Mustererkennung SS17 (Kropatsch)
Fragen

Formatierung:
Units: Formatvorlage "Überschrift 2"
Tasks: Formatvorlage "Überschrift 3"
Fragen: Text normal als schwarze Bullet Points

- Ist das eine Frage?

Antworten: Richtige Antworten **bold**,
bei Multiple Choice bitte diese Unicode Symbole verwenden (copy+paste)

☐ Die Falschheit
☒ **Die Wahrheit**

Bilder bitte in-line ausrichten, mit einem Zeilenumbruch (shift+enter) nach dem Bullet Point, damit nicht alles durcheinander kommt! :)

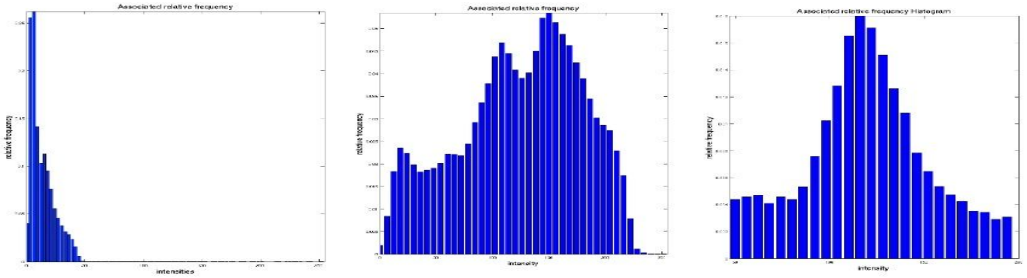
Unit 1: Feature Extraction

Task 1.1: Histograms of Oriented Gradients

- Is HOG per se scale invariant?
  ☐ Yes
  ☒ **No**

- What is the goal of HoG?
  ☒ **Robust visual object recognition**
  ☐ Finding differences in images
  ☐ Texture Classification

- Which histogram corresponds to the provided image?
  Answer: (a)



(a) Answer1 (b) Answer2 (c) Answer3

- Why does the HOG algorithm normalize the feature vector?
  Answer: For better invariance to illumination and contrast.

- What is calculated for a cell in the context of HOG features?
  ☒ **histogram of gradient angles weighted by the gradient magnitude**
  ☐ histogram of gradient angles weighted by the pixel luminance
  ☐ histogram of pixel luminance
  ☐ histogram of gradient magnitudes

- What are some of the main differences to approaches like SIFT?
  ☒ **HOGs calculated one just one scale space**
  ☐ No rotation invariance
  ☒ **No keypoint detection**
  ☒ **Dense calculation of histograms**

- What does a block size of 2x2 indicate?
  ☐ 4 gradients are computed per block
  ☐ The number of orientational bins in the histogram is 4
  ☒ **4 histograms are contrast normalized by average intensity**
  ☐ 4 unique image regions are exclusively used in each separate block

- Hogs Histogram is calculated by ..?
  ☐ Block
  ☒ **Cell**

- Why is normalization applied to the histograms?
  ☐ to reduce noise artifacts in histograms
  ☒ **to achieve lighting independency**
  ☐ to simplify the resulting feature vector

- What preprocessing steps might be useful?
  ☐ desaturation
  ☒ **resolution reduction**
  ☐ white correction

- How are the histograms normalized?
  Answer: The algorithm combines several cells to larger spatial regions ("blocks") and normalizes the combined histograms for each cell in the block.
- What is the difference in the histogram for unsigned gradients and signed gradients?
  The histogram is a vector of 9 bins corresponding to angles 0, 20, 40, ...,160. The angles are between 0 and 180 degrees instead of 0 to 360 degrees. These are called unsigned gradients because a gradient and it's negative are represented by the same number.
- What does HOG stand for?
  ☐ Harmonic Oriented Gradients
  ☒ **Histograms of Oriented Gradients**
  ☐ Histograms over Gradients
  ☐ Histograms of Gaussian
- How does HOG normalize each cell?
  ☐ By the energy of the cell
  ☒ **By the energy of the surrounding block**
  ☐ By the energy of the blocks which surround the block of the cell
  ☐ By the energy of the whole image

- Why does it make sense to use a local contrast normalization step in the feature extraction process?
  To increase the photometric invariance of the descriptors.

- Which two values do form the HOG?
  ☒ **The gradient magnitude and the gradient's orientation**
  ☐ The gradient's orientation and the colours of the image
  ☐ The gradient's orientation and the intensities of the image

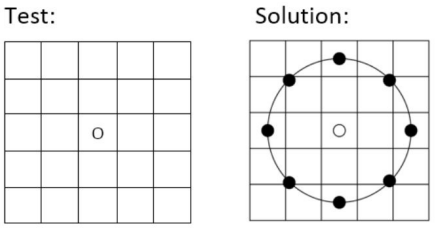
- The HOG as feature descriptor is able to
  ☒ **identify all sorts of objects with sharp edges.**
  ☐ only identify humans.

- What are the histograms of the HOG displaying?
  ☒ **The amount of gradients, which point in a specific angle**
  ☐ The amount of different colors in the image
  ☐ The amount of cells in the image

- Which steps actually happen in the HOG algorithm?
  ☒ **Dividing the image into several cells**
  ☒ **Weighting the various gradients through its magnitude**
  ☒ **Normalizing the local histograms with the histograms of the neighbouring blocks**

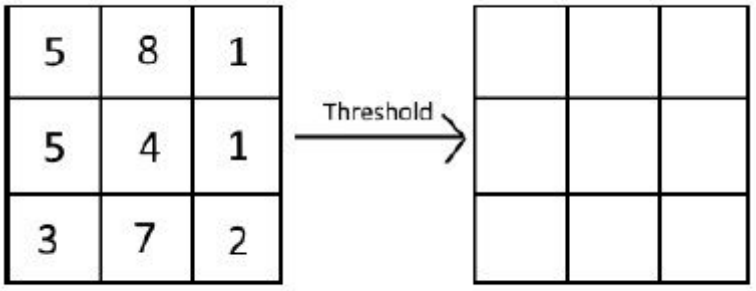
Task 1.2: Local Binary Pattern

- What would a circular (8,2) neighbourhood of the center pixel (marked with circle) with interpolated pixel values look like?



- How many unique bit patterns are possible on a (8,2) Neighbourhood-LBP, without any bitshifting?
  ☐ 2
  ☐ 8
  ☐ 16
  ☒ **256**

- Threshold the following 3x3 neighborhood according to the Local Binary Pattern (LBP) operator:



Solution: **1 1 0**
1 0
0 1 0

- Which properties are true for the original LBP operator?
  ☒ **Robust to illumination variation**
  ☒ **Computational simple**
  ☐ Scale invariant

- Which value is taken as the threshold in Local Binary Patterns?
  Answer: The value in the middle of the filter kernel

- Why are Local Binary Patterns insensitive to global illumination changes?

Answer: Because only the difference between the central value and its neighbours is relevant

- How are Local Binary Patterns used to describe an image?
  Answer: By a histogram of local descriptors consisting of the local descriptor calculated with the LBP operator for each pixel.

- What does the LBP operator assign to each pixel of an image?
  Answer: a label calculated with the values assigned to the neighboring pixels of a center pixel (= pixel of interest)

- What statistical tool is mainly used in the analysis of Local Binary Patterns?
  ☐ Pie charts
  ☒ **Histograms**
  ☐ Line charts
  ☐ Venn diagrams

- Why is LBP rotation invariant?
  ☐ Invert Bits
  ☒ **Shift Bits**

- Why do LBPs have the qualifier *binary*?
  The description is essentially in the form of a binary number.

- What makes an LBP descriptor *uniform*?
  When only at most two bitwise transitions occur in it.

- How many patterns can be described with a neighbourhood of 3x3?
  ☒ **2^8**
  ☐ 2^9
  ☐ 2^10

- Which of the following patterns is a uniform pattern?
  ☒ **00000000**
  ☐ 11001001
  ☐ 01010011

- Which of these answers is correct?
  ☐ LBP only works with a 3x3 neighbourhood
  ☐ The calculation of the neighbours works only clockwise
  ☒ **The direction of the calculation of the neighbours is irrelevant, it only has to be consistent**

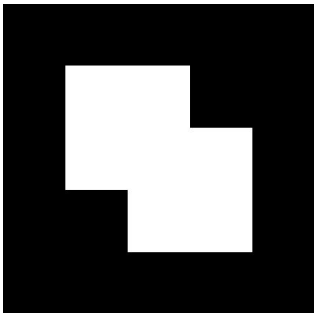
- Which two parameters are needed to describe a scalable neighbourhood?
  ☒ **Amount of sample points**
  ☒ **Radius**

☐ Angle

- What two Parameters have to be considered when Local Binary Patterns are used?
  Neighbors P (=sampling points) and Radius R of the circle around the center pixel

Task 1.3: Features from binary regions

- What is the Euler number of this figure? (black = background)



☐ -1
☐ 0
☒ **1**
☐ 8

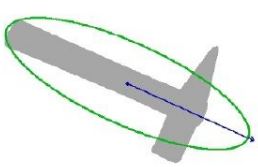
- Do these regions have the same centroid?

☐ Yes
☒ **No**

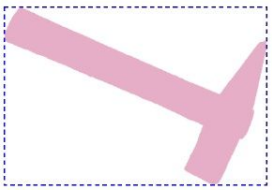
- Which perimeter of the same binary region will be larger: the one calculated with a 4-neighborhood or the one calculated with a 8-neighborhood?
  Answer: 4-neighborhood

- Describe the convex hull of a binary region!
  Answer: convex hull is the smallest convex polygon that contains all points of the region. A physical analogy is a board in which nails stick out in correspondence to each of the points in the region. If you were to place an elastic band around all the nails, then, when you release it, it will contract into a convex hull around the nails

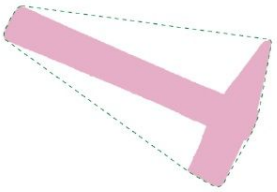
1. What image describes the Bounding Box feature?



(a) img1

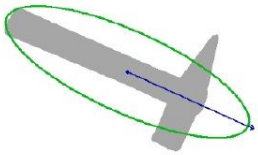


(b) img2

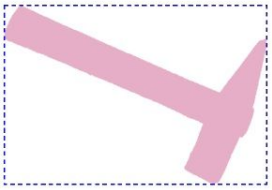


(c) img3

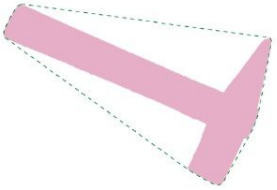
2. What image describes the Convex Hull feature?



(a) img1



(b) img2

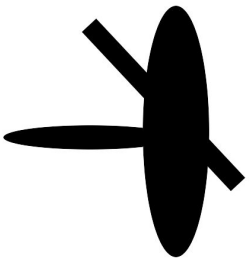


(c) img3

- **Answers:**
  1. (b)
  2. (c)

- Is the perimeter of a circle or a square bigger if both have same area?
  Answer: Square
  because circle has always the smallest perimeter per area, think of waterdrops for example, they form spheres (3D-circles) because they want a minimal surface (in 2D: perimeter)

- Draw the bounding box and the convex hull for the following binary region:



The convex area of the left image is very roughly estimated at around 5000. Give a rough estimate of the convex area of the right image. Note that both images have the same resolution.

(a) Convex Area $\approx 5000 \pm 500$ (b) Convex Area $\approx$??

- **Answer: Same**

- Which of the following characters have the same Euler number as ”R”
  ☒ **O**
  ☐ 8
  ☐ W
  ☒ **A**

- ...

- What makes the bounding box special in contrast with the convex hull?
  The bounding box is axis-parallel.

- What physical property describes the centroid?
  The center of gravity.

Task 1.4: Feature extraction in MATLAB

- What are uniform patterns and why are they commonly used in a LBP histogram?
  Answer: Local Binary Patterns with at most two transitions from 0 to 1 or 1 to 0. They are used to reduce the feature vector length and achieve rotation invariance via bit shifting

- The Viola Jones algorithm for face detection uses:
  ☒ **Haar feature selection**
  ☐ Sift feature selection
  ☐ HOG feature selection

- For the Viola Jones algorithm a higher thresholds may suppress false detections, but what could happen if it is set too high (more false positive or **more false negative**)?

- Draw a 3x3 array with random values and calculate the center points LBP label value as a decimal. Is it a uniform pattern?
  See question" Threshold the following 3x3 neighborhood according to the Local Binary Pattern (LBP) operator" as an example
  Solution: 1 1 0
  1 0
  0 1 0
  this pattern is not uniform.

- What faces tend to be not detected by Viola Jones? Rotated faces or smiling faces?
  Answer: Rotated faces because basic Haar features have fixed alignment:
  goo.gl/qkLf5U goo.gl/S68swt
  https://en.wikipedia.org/wiki/Viola%E2%80%93Jones_object_detection_framework

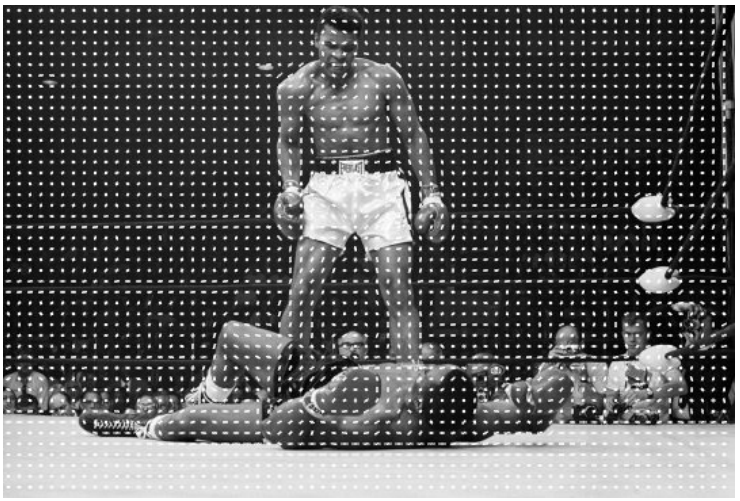
- Does the Viola Jones face detector detect faces which are only visible sideways (e.g. the ear points to the camera)?
  Answer: No, because the basic features (see the question above), which for example represent the eye region, are not visible in the same way.

- What is the main usage of the Viola Jones detector?
  ☐ Edge detection
  ☐ Corner detection
  ☒ **Face detection**
  ☐ Coin detection

- For which applications can Harris Features be used?
  Answer: Corner Detection

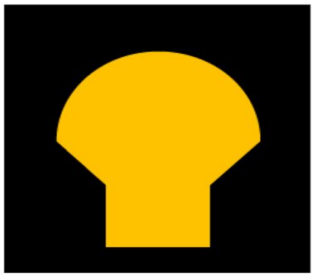
- Select which feature extraction method yields the most matching results?
  ☐ Harris
  ☒ **SIFT**
  ☐ SURF

- What was most likely applied to this image? **HOG features**

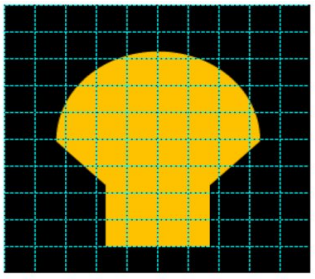


- Which of the following feature descriptors are scale and rotation invariant?
  ☒ **SIFT [YES]**
  ☐ Harris [NO]
  ☒ **SURF [YES]**

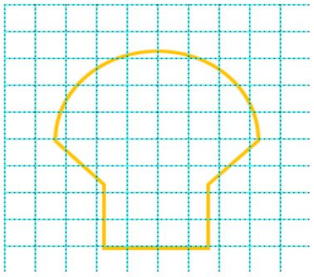
- Draw a visualization of the HOG features that can be extracted from the image 16a to the grid 16c. The cells of the grid should be used as the HOG algorithms cell-size. The yellow shape should help you to find the right cells.



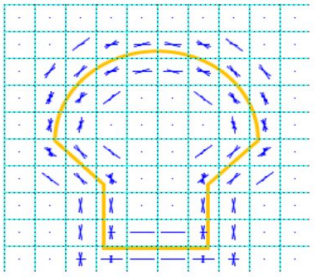
(a) source image



(b) source image with grid



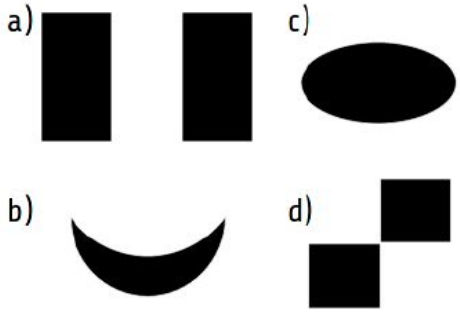
(c) target image



(d) sample solution

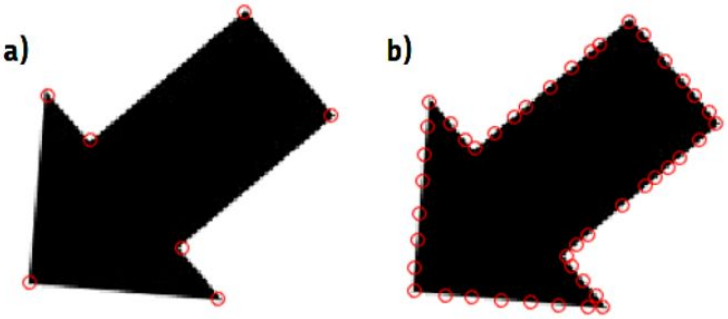
- Answer: **(b) and (c)**

Which of the following patterns do NOT belong to the Haar features used by the Viola-Jones method?



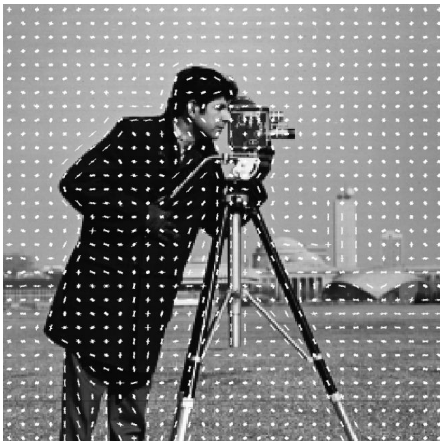
- Answer: **(a)**

A gaussian filter is often applied before using a corner detector, such as the Harris-Stephens method, to detect corners on different scales. Which of the following uses the filter kernel of greater size?



- What are the two drawbacks of the MinimumEigenvalue corner detector when applying it with a high maximum number of corners?
  Answer: When increasing the maximum number of corners from 200 to 500, the Harris detector found only 300 corners and the MinimumEigenvalue detector found 500. The MinimumEigenvalue detector detects edges that are not horizontal or vertical as corners. Furthermore, it reacts strongly on noisy image parts for example the grass in the Figure.

- What influence has the cell size on the detected HOG features?
  The HOG Features describe clearly the contour of the man in the image:



  The higher the cell size the less features are detected:

- How is SIFT used in panorama stitching?
  Answer: Find Keypoints at the side regions of the picture. Align (rotate, move, scale, i.e. find homography mapping) the next picture so that its keypoints overlap the first ones.

- What is the length of the SIFT feature vector?
  Answer: The SIFT vector has a length of 128

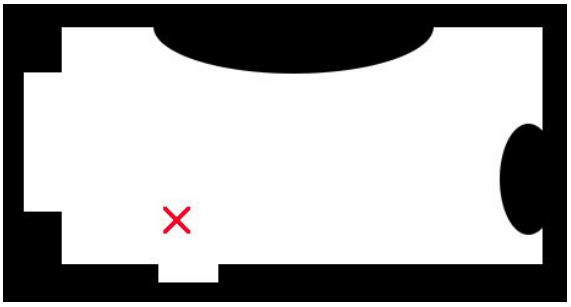
Unit 2: Structural Pattern Recognition

Task 2.1: Eccentricity Transform in MATLAB

- What could be an application of the Eccentricity Transform?
  Answer: Finding connected regions/Shape Matching

- The naive implementation of the Eccentricity Transform is very complex because of two nested loops. What is being calculated in these loops?
  Answer: The inner loop searches for the longest geodesic distance for a point of interest (needs to compare this point to every other pixel). The outer loop executes the inner loop for every pixel.

- The eccentricity transform value of the marked pixel in the image below equals the distance of the marked pixel to a certain other pixel. Highlight that other pixel! (The background is black and the relevant digital shape is white.)
  Answer: Always the pixel with the highest distance to the pixel is relevant. This case: top right



- Imagine a square. Where are the points with the lowest distance transform values and where are the points with the lowest eccentricity values?
  Answer: Lowest distance transform: border. Lowest eccentricity: center (the edges are equally far away)

- On what is the eccentricity transform dependent on?
  ☒ **distance transform**
  ☐ time transform
  ☐ value transform

- Is the eccentricity transform robust to salt and pepper noise? **Answer: Yes**

- What is the relationship between the radius of G and central vertices and the diameter and peripheral vertices?
  Answer: The radius of G is the minimum eccentricity and diameter is the maximum eccentricity.
  For a central vertex the eccentricity is equal to the radius and for a peripheral vertex the

eccentricity is equal to the diameter of G.

- What is the reason the eccentricity transform is more robust to salt and pepper noise than the distance transform?
  Answer: As the eccentricity transform finds the shortest path to the farthest pixel, it simply finds a path around a single noise pixel.

- What is the maximum error, if the eccentricity transformation is applied on a mask with a hole of size Epsilon? **Answer: Epsilon**
- For what can the Eccentricity Transformation be used?
  ☒ **Shape matching**
  ☐ To transform the shape
  ☐ Edge detection
  ☐ To find holes

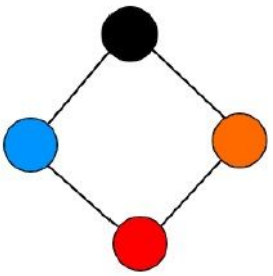
Task 2.2: Graph-based Pattern Recognition

- What are the main rules for determining the weight of the operations for the calculation of the edit distance?
  Answer: Insertions and Deletions have fixed weights and are usually more costly than substitutions which weight depends on the difference of the properties between the original and the substitute.

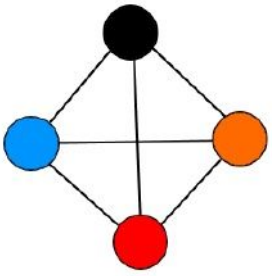
- What are the requirements of a similarity measure to be a metric? Can the edit distance fulfill them?
  Answer:
  positive definiteness, symmetry, triangle inequality
  Yes: "Edit distance with non-negative cost satisfies the axioms of a metric"

- Answer: **(a)**

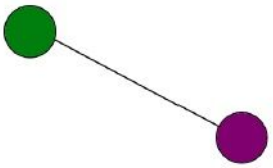
Which of these graphs (a, b, c) has the edit distance of 2 to the given graph? All operations have the cost of one.



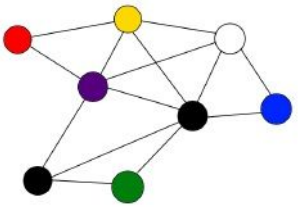
(a) Provided graph



(b) Answer a)



(c) Answer b)



(d) Answer c)

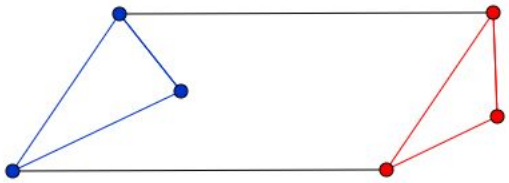
- Which data structure can be used for structural pattern recognition?
  ☒ **Trees, Graphs**
  ☐ Feature vectors
  ☐ Histograms

- What do you need for "exact graph matching"?
  ☒ **Graph Isomorphism**
  ☐ Subgraph Isomorphism

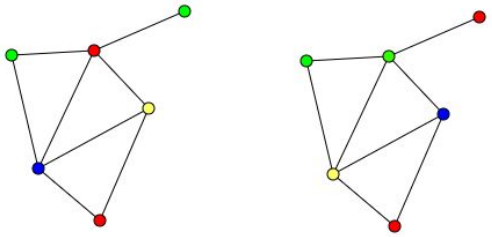
- Why are graphs more expressive than feature vectors?
  Answer: Because graphs can contain more information/attributes/characteristics of a feature and do not have a fixed size. Can formulate relations.

- How can graphs describe a certain object/pattern - (what are the nodes/edges)?
  Answer: Nodes are mostly regions (e.g. in an image) and edges indicated if two regions are adjacent.

- What type of graph matching can be seen here? **Answer: Exact graph matching**

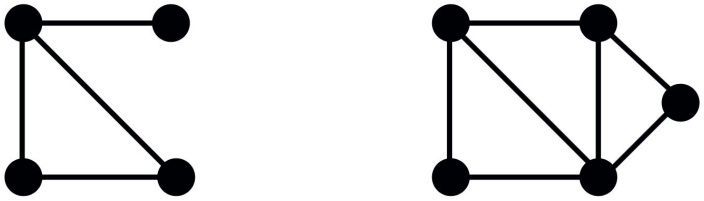


- What is the graph edit distance between the two graphs (using insertion, deletion, relabeling of one node or edge, all have cost 1)? **Answer: 4 (4x relabeling)**

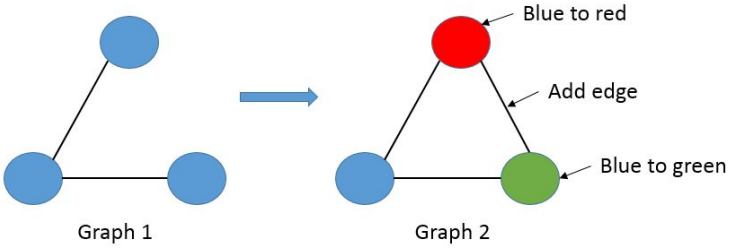


- What are the advantages of structural pattern recognition in comparison to statistical pattern recognition?
  ☐ easier calculations [NO]
  ☒ **objects which differ in their complexity can be represented by different complex data structures [YES]**
  ☒ **relations between information extracted from different regions are preserved [YES]**
  ☐ training can be automated better [NO]

- What is the graph edit distance between the following two graphs?
  Answer: 4 (1 node insertion + 3edge insertions)



- Design a small graph, where the graph edit distance is 3 (all operations have a cost of 1).



Answer:

- What are the advantages and disadvantages of the statistical and structural pattern recognition?
  Answer: Statistical pattern recognition is characterized by describing objects by means of feature vectors. The advantage of the statistical approach is that it is theoretically well founded and many powerful algorithms are available like for example the SVM. The

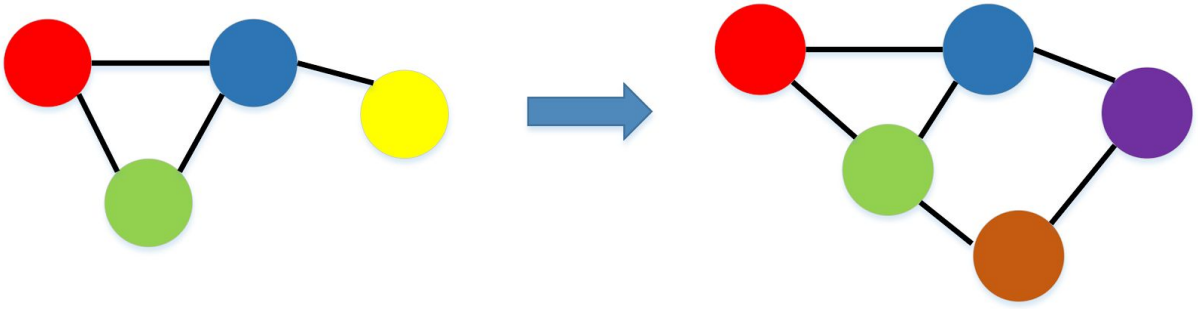
disadvantage is that the dimension of the feature vector is fixed and one can represent only unary features, but no relations between them. In the structural approach we use symbolic data structures like trees or graphs to describe the objects. The advantage of this approach is that the representation size is variable and the representational power is higher because you can show relations as well. But the disadvantage is the lack of mathematical structure in the graph domain.

This is easier to remember	**Advantage**	**Disadvantage**
Statistical	• **theoretically well founded** • **many powerful algorithms are available (like SVM)**	• **feature vector size is fixed** • **can only represent unary features and no relations between**
structural	• **representation size is variable** • **representational power is higher because you can show relations**	• **lack of mathematical structure**

- Which approach to pattern recognition (statistical or structural) utilizes hierarchical pattern descriptions?
  Answer: Structural.

- Between graph G1 and G2 there exists an edit path P1 with associated cost 7. Between graph G1 and G3 there exists an edit path P2 with associated cost 2. Which graph (G2 or G3) is more similar to G1?
  Answer: G3 is more similar to G1 (edit distance cost denotes dissimilarity).

- What is the result of the graph edit distance of the graph, when all operations have a cost of 1?



☒ **4**
☐ 5
☐ 6

- What are the two main fields in which Pattern Recognition is divided?

☐ Color Pattern Recognition
☒ **Structural Pattern Recognition**
☐ Number Pattern Recognition
☒ **Statistical Pattern Recognition**

- What is true about structural pattern recognition?
  ☒ **There is a lack of operations in the graph domain.**
  ☒ **Structural pattern recognition can be used to characterize different patterns and model relations between objects.**
  ☐ All graphs must have the same dimensionality.

- Which statements to "Graph Edit Distance" are correct?
  ☒ **This method is error-tolerant**
  ☒ **Every transformation has a cost**
  ☐ A distribution model is used
  ☐ The higher the accumulated cost the more similar are the graphs

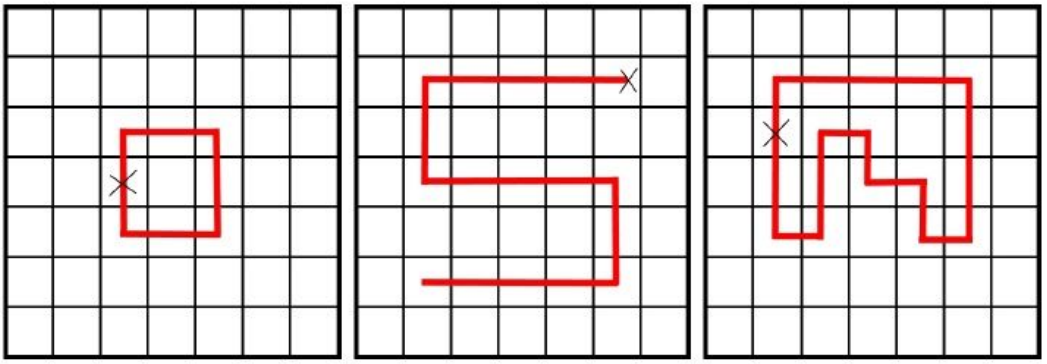
Task 2.3: RULI Chain Code

- What has to be done to make the RULI Chain Code NRIILUIE rotation invariant?
  assuming N indicates initial direction in Freeman code: just remove the N

- Has a curve defined by a RULI chain same/larger/smaller spatial resolution than the pixel grid of the image is is drawn in? **QUESTION UNCLEAR**
- **NOTE: $(IR)^4$ not $(IR)^3$**
  Answer: (a) Graph1

What image belongs to the provided RULI chain code?
The curve is painted in red and the starting point is marked by a black X.

Code = $(IR)^3$



(a) Graph1 (b) Graph2 (c) Graph3

- Where does the name “RULI” chain code come from? **Right, U-turn, Left, I: straight line**

- Draw the following RULI chain: $I^2\,RLR^2I(LR)^2$

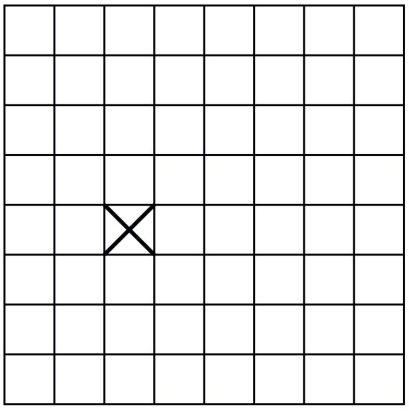


Figure 1: Start at the X

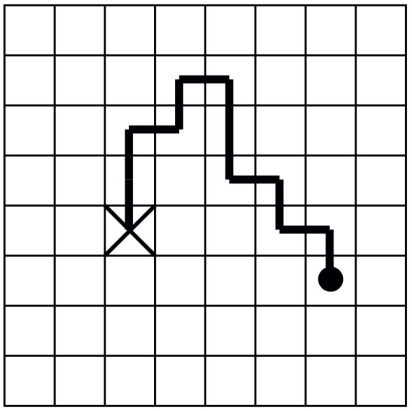


Figure 2: Solution

- Is the RULI chain I^2R^2IR closed?
  Answer: No (because the direction at the start and at the end do not match)

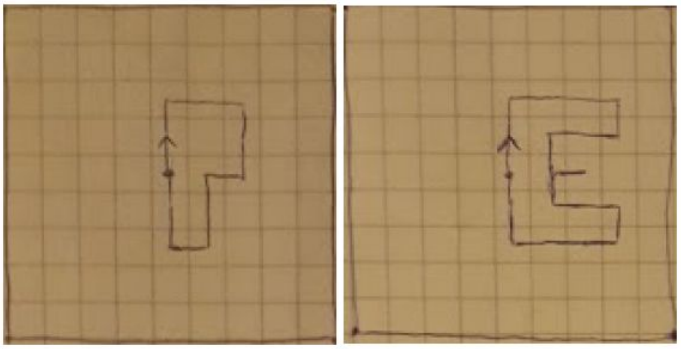
- Draw the following closed RULI chain: $[RL(LI)^2L^2]$

- Rewrite the following Freeman chain code (giving absolute directions N,E,S,W) as RULI chain code. Regarding the RULI code between start and end, note that the chain is closed.
  Freeman code: $[E^2(NW)^2SWSE]$
  Answer: RULI code: $[I^2L^2(RL^2)^2]$

- Draw the following RULI chain code: $I^2LRI^2R2I^5UR^2I^2$
- Insert the missing RULI chain code elements to make it cyclic: I^2R^2____$(LR^4)UI^2$
  Answer: I^2R

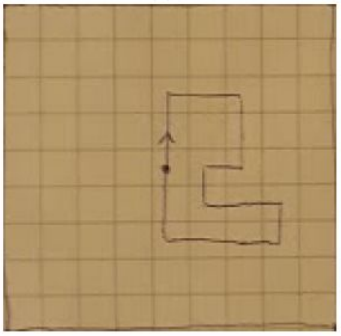
- Which of the following paths represents the following RULI code: **answer: a**

$$P = [I(IR)^3LIR^2I]$$

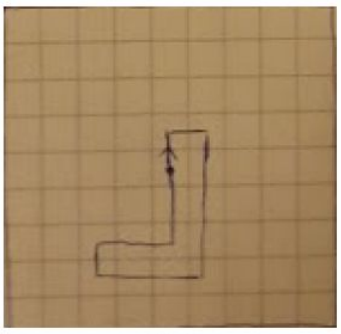


(a) a

(b) b



(c) c



(d) d

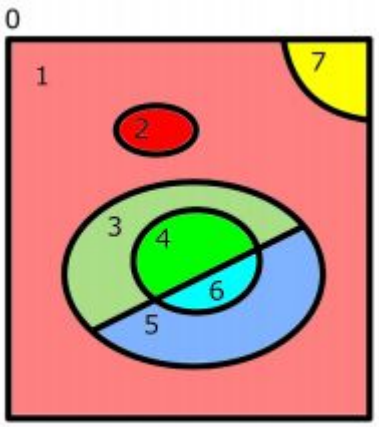
- How many rules do exist in RULI chain codes and what are those?
  Answer: 4 Rules: R: Right, U: U-Turn, L: Left, I: Intersect the current cell

- For a curve to be closed is it enough to reach the starting position at the end again?
  Answer: No, the direction must also be the same

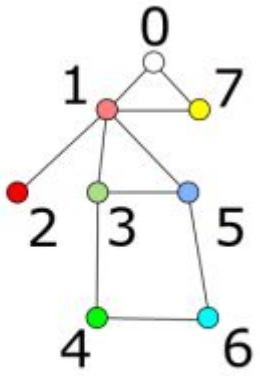
Task 2.4: Region Adjacency Graph and Dual Graph

- Under which condition can a Region Adjacency Graph be cut?
  Answer: If a region encloses other regions, then the part of the graph corresponding with the areas inside can be separated by a cut in the graph.

- Draw the Region Adjacency Graph of the graph displayed here:



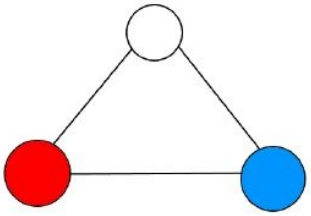
(a) Question



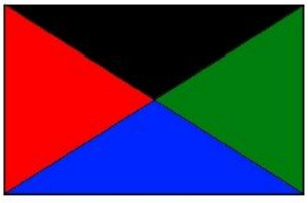
(b) Answer

- **Answer: Graph 2**

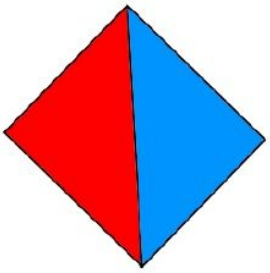
Which image belongs to the given region adjacency graph? White vertices describe the virtual region surrounding the image.



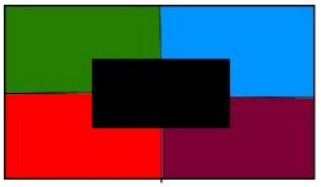
(a) Provided image



(b) Graph1



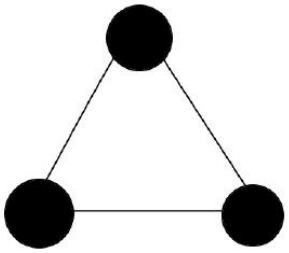
(c) Graph2



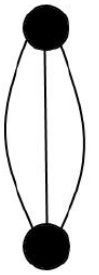
(d) Graph3

- Answer. **Graph 1**

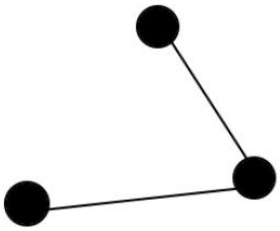
Which dual graph corresponds to the given graph?



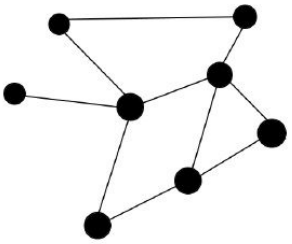
(a) Provided graph



(b) Graph1

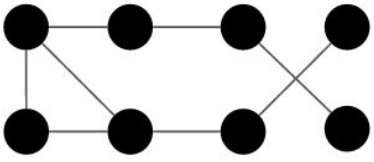


(c) Graph2

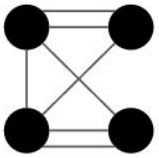


(d) Graph3

- Is there a homomorphism between the following two graphs?



and

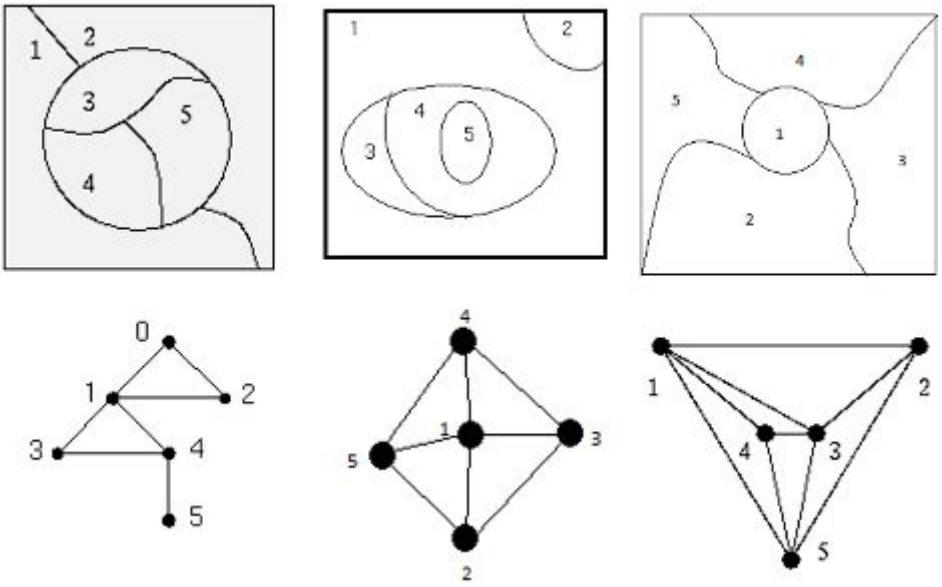


Answer: No, not possible, the number of edges is not equal ;)

- If a graph has 4 vertices, 6 edges and 4 faces (including the surrounding face), how many vertices does the dual graph have?
  Answer: A dual graph has a vertex for every face of the original graph ⟶ 4 vertices

- Does a planar graph always have a dual graph? **YES**

- Map the given images to the corresponding region adjacency graphs



Answer: (the first number indicates the image in the first row, the second in the second row)

(1) - (3)

(2) - (1)

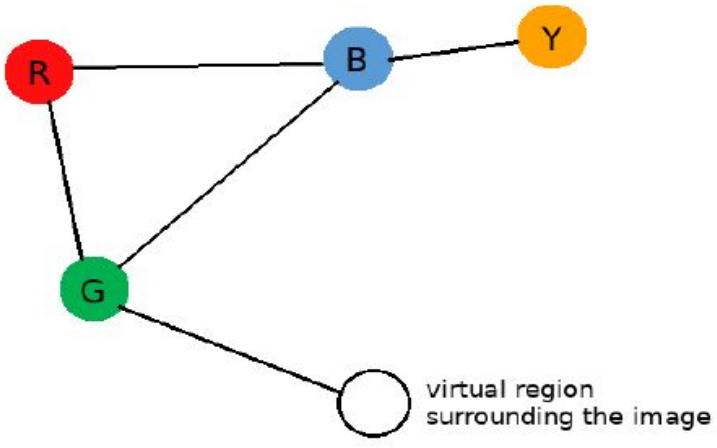
(3) - (2)

- Given a planar graph, what are the vertices of the corresponding dual graph?

  Answer: A dual graph has a vertex for every face of the original graph

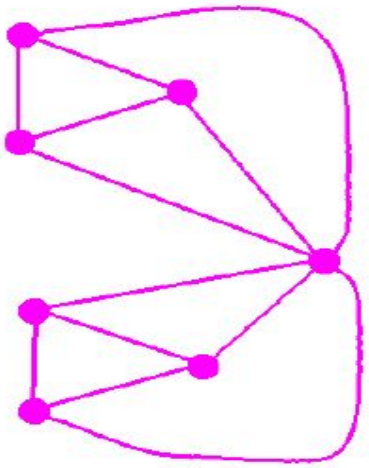
- Draw region adjacency map for given adjacency graph

  Draw the regions (region adjacency map) corresponding to the following region adjacency graph. Draw the regions in color or label them with the respective letter.

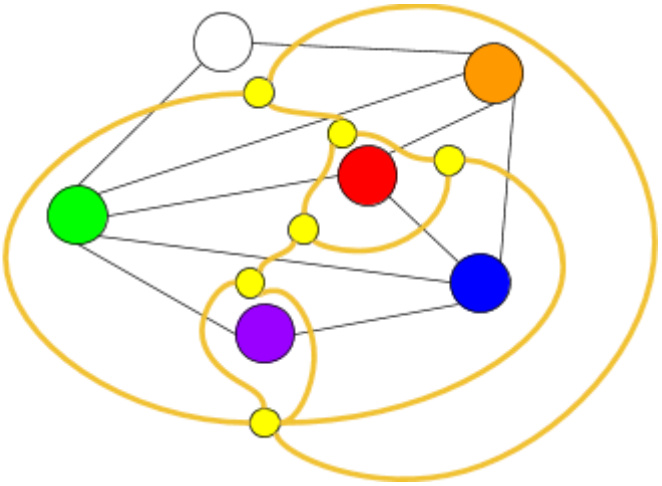


- **Answer:** https://commons.wikimedia.org/wiki/File:Self-dual_graph.svg

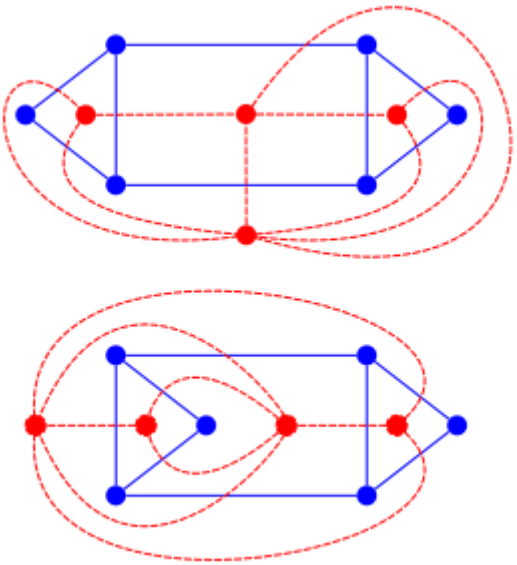
Show that the following graph is isomorphic to its dual graph, and in a few words explain why. Hint: Think symmetrically. Also don't forget to include a vertex corresponding to the outer face surrounding the graph.



- Create the region adjacency graph and its dual graph of the following image:



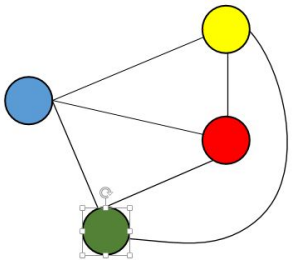
- Are two dual graphs of isomorphic primal graphs always isomorphic to themselves? Either give a proof or a counter example.
  Answer: No, the dual graph depends on the embedding.



- The dual graph G' of the planar graph G has a vertex for every …
  ☐ vertex
  ☐ edge
  ☒ **face**
  ... of G.

- In a region adjacency graph, what type of edge represents a region that is on both sides of a single edge?
  Answer: A loop.

- How many edges does the dual graph of the primal graph shown below have?
  ☐ 4
  ☐ 5
  ☒ **6**



- What is true about Region Adjacency Graphs (RAG)?
  ☒ **Edges represent neighbouring relations.**
  ☒ **When one region encloses another region, this relationship can easily be visualized with RAGs.**

☒ **Vertices describe image regions.**
☐ RAGs do always have multiple edges and self-loops.

Task 2.5: Image Segmentation with Graphs

- How is the internal contrast defined?
  ☐ the largest dissimilarity inside one component.
  ☐ the smallest dissimilarity between two different components.
  ☒ **the largest similarity inside one component.**

- What is true about Minimal Spanning Trees (MST) and their calculation:
  ☐ connect pixels with the highest feature distance.
  ☒ **They split pixels into segments according to the distance of their features.**
  ☒ **The euclidean distance can be used to calculate the feature distance in MSTs.**
  ☒ **Distance Features can be for example colour values or intensity values.**

- What are possible ways to evaluate a segmentation?
  ☒ **Let the segmentation result get evaluated by a human.**
  ☒ **Evaluate parameters of the algorithm, like efficiency.**
  ☐ Use a Gaussian filter for evaluation.

- How are contrast zones defined?
  ☐ Internal contrast is the largest dissimilarity inside one component.
  ☒ **Internal contrast is the lowest dissimilarity inside one component**
  ☐ External contrast is the smallest dissimilarities between two components

Unit 3: Classifiers

Task 3.1: Classifiers: Terms & Definitions

- Give an example of a descriptor
  Answer: Area, Perimeter.
  A descriptor describes the visual features and elementary characteristics of an image like shape, color, texture.

- Name two types of classifiers

- What is the lower accuracy limit for a weak classifier to still be considered a weak classifier?
  Answer: It has to be better than random guessing. So its accuracy must be better than 1/number_of_classes

- What is the difference between supervised, semi-supervised and unsupervised learning processes?
  Answer: In supervised learning the classifier expects a manually given set of descriptors with assigned classes as an example for each class. Often this requires expert knowledge or preprocessed data.
  In unsupervised learning there are no examples or prototypes for expected classes. The aim there is to find interesting structure in the given unlabeled data or more generally speaking to guess the underlying distribution function.
  A combination of supervised and unsupervised learning is semi-supervised learning and an example could be self-training which starts off as a supervised learning process but readjusts the class definitions with the most confident label guesses as new data is incoming.

- What is the difference between a descriptor and a feature?
  Answer: Feature = Descriptor + Keypoint (2D Position in Image)

- How can weak classifiers be turned into strong ones?
  Answer: Combine multiple weak ones

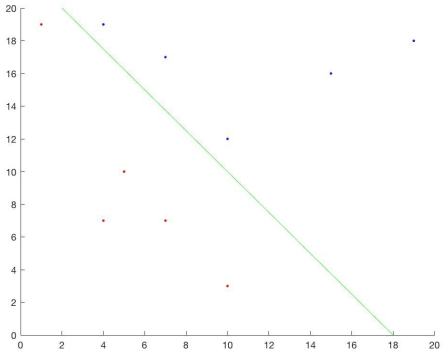
- Which properties describe a Classifier?
  ☐ Classifiers are elements of an image like color, edges etc.
  ☐ A classifier is also often called a ‘feature’
  ☒ **A classifier assigns class numbers to descriptors**
  ☒ **Classifiers learn rules to label new examples from a given set**

- Which properties are true for a weak classifier?
  ☐ A learner is called weak if the classification accuracy if over 90%
  ☒ **Simple thresholding is an example for a weak classifier**
  ☒ **Weak classifiers can iteratively be combined to a strong classifier**

- Which of the following classifier algorithms supports Unsupervised Learning?
  ☒ **k-means [YES]**
  ☐ support vector machines (SVM) [NO]
  ☐ **neural networks [NO]**

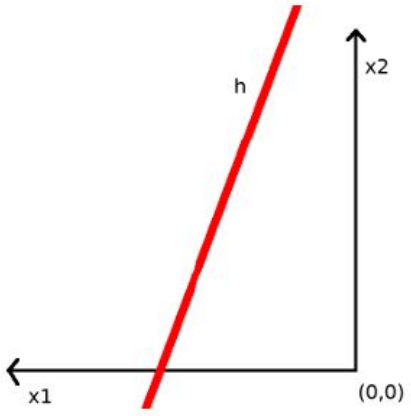
- What does a classifier do with a descriptor?
  Answer: Classifier assigns a class label to each descriptor.

- Show in the image: What is a descriptor.

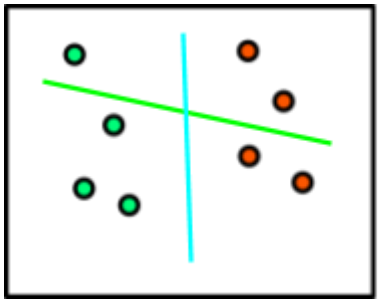


- What is the difference between supervised and unsupervised learning?
  Answer:
  Supervised: with correct label Y.
  Unsupervised: without correct label. Looking for interesting pattern in data.

- Which of the following statements are true?
  ☒ **A feature is a descriptor and a keypoint**
  ☐ Binary classifiers are by definition weak classifiers
  ☒ **A linear binary classifier can be described by a hyperplane**
  ☒ **The feature space of a SIFT feature is 130-dimensional**
- A binary linear classifier my be defined as a hyperplane $h : \mathbf{w}^T\mathbf{x} - b = 0$, to classify a descriptor $\mathbf{x}$. In the following example of a 2D descriptor space, draw $\mathbf{w}$ and b and explain their function.



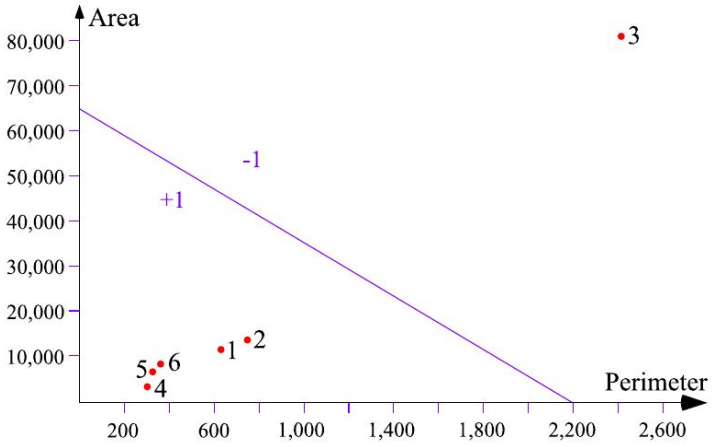
- Draw a weak and a strong classifier for the following image:

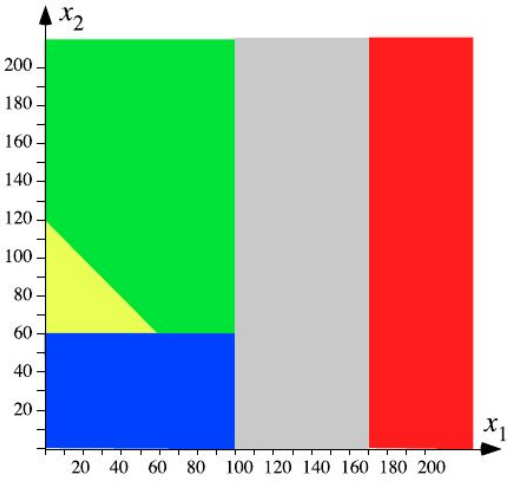


weak classifier: better than a coin flip

strong classifier: better than 90% correct classification

- For the following two image state if the shown classifier is a binary or general classifier.





Answer: The first is a binary classifier the second is a general classifier

- State a simple example where multiple weak classifiers can be combined to form a strong classifier.

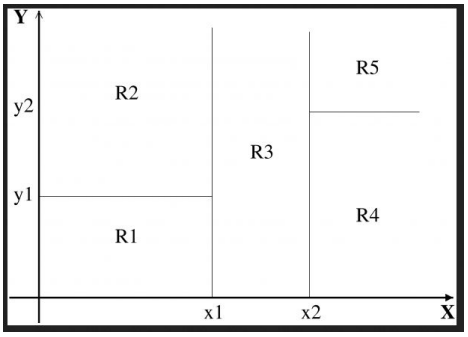
Answer: Haar features used for face detection work by combination of several weak classifiers

- **In the terminology of image analysis a feature is ...**
  ☐ a keypoint in the image.
  ☐ a descriptor of the image.
  ☒ **a combination of the two above.**

- What could be one possible method to turn weak classifiers into strong classifiers?
  Answer: Multiple weak classifiers can be mapped into a strong classifier, to achieve at a satisfactory solution of a classification problem.

- What is true about features?
  ☒ **They are specific properties extracted or calculated from images, like size or shape.**
  ☐ They can also be called descriptors.
  ☒ **Similar objects should have similar feature vectors.**
  ☐ Features can only be quantitative.

- Is a linear binary tree able to calculate a separation as shown below?
  ☐ No
  ☒ **Yes**



Task 3.2: Support Vector Machine (SVM)

- Which of the following is true?
  ☒ **SVM is a supervised learning model**
  ☒ **SVM can be used for binary classification**
  ☐ The margin of the perceptron algorithm is in general bigger

- What does an SVM do if a problem is not linearly separable?
  Answer: Soft margin approach: Allow misclassification but punish them by adding the weighted slack variables (how far the misclassified element is beyond the separating hyperplane)

- What would be a possible way to find candidates for the support vectors in a binary task?

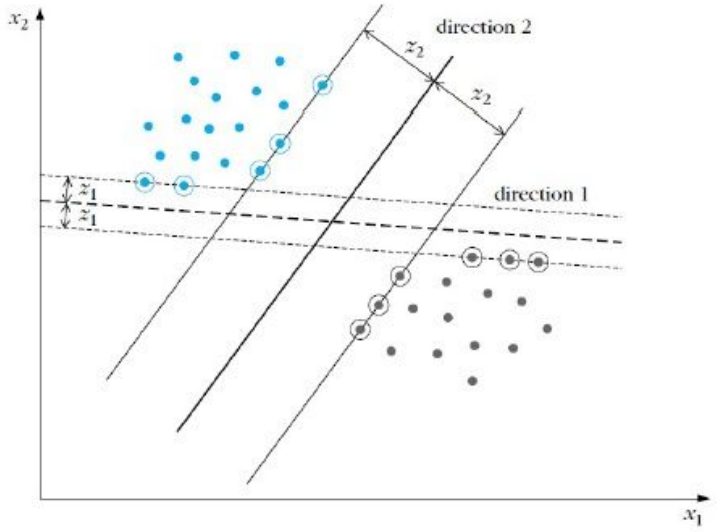
Answer: Finding the centroid of each class. Imagine a line between the centroids. Project all elements onto this line. Starting from one side of the line, take the last few elements which are still in the first class before an element of the second class occurs. And also take the first elements of the second class after there are no more elements of the first class.

- How does SVM separate two classes?
  Answer: By finding a linear boundary between them where the nearest elements of both classes are equally distant (maximum margin).

- What does it mean that SVM is a maximum margin classifier?
  Answer: Finds separating plane that maximizes distance between samples of different classes

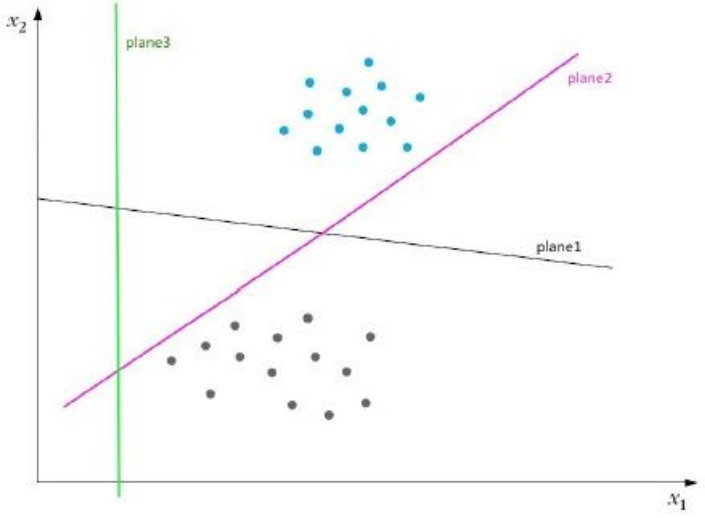
- **Answer: Direction 2**

Which of these two directions describes a maximum margin hyperplane?



- **Answer: Plane1**

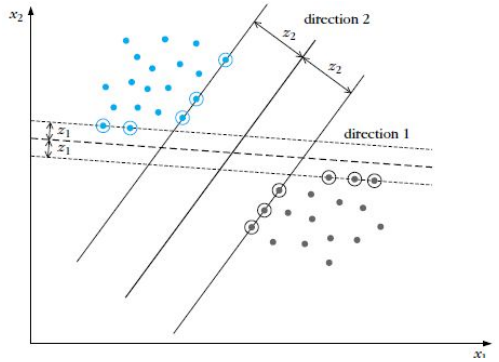
Which of these hyperplanes would be used for a SVM classification?



- Which of the following statements are true?
  ☒ **SVM is a maximum-margin classifier**
  ☒ **SVM is a binary classifier**
  ☐ SVM only works for linearly separable data
  ☐ There is no way for a SVM to perform non-linear classification

- What is the advantage of the SVM compared to the Perceptron Algorithm?
  Answer: SVM finds a decision hyperplane that maximizes the margin between the classes.

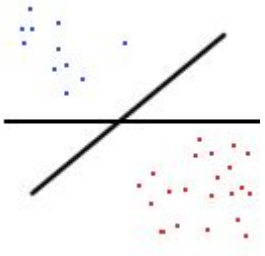
- What are Support vectors, support lines and the margin?
  Answer: A two lines that go through the support vectors are called support lines. The region between the two lines is called the margin.



-

- Can a Support Vector Machine (SVM) be used for data that is not separable? Why or Why not?
  Answer: Yes by using the Soft margin approach: Allow misclassification but punish them by adding the weighted slack variables (how far the misclassified element is beyond the separating hyperplane)

- In the following picture which of those hyperplanes would a SVM have calculated and why? Can this also be said about the Perceptron Algorithm?



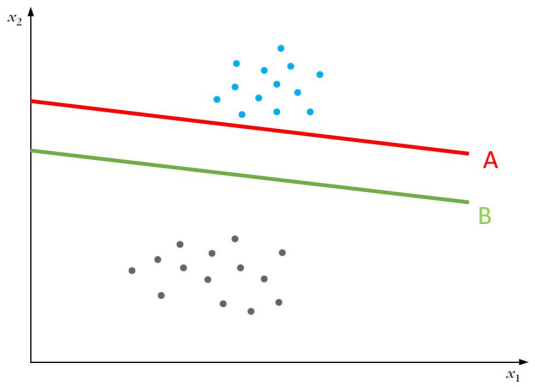
Answer: The line that is around 45° as it calculates the maximum margin to both classes

- Which statements about the Support Vector Machine are true?
  ☒ **It is a linear classifier.**
  ☒ **It consists of a weight vector and a threshold.**
  ☐ It can differentiate between more than two classes.

- What is the difference between the SVM and the Perceptron Algorithm?
  ☒ **The SVM calculates a hyperplane furthest away from both classes.**
  ☐ The SVM calculates the nearest hyperplane to one of the two classes.
  ☐ The Perceptron Algorithm calculates the perfect hyperplane

- What is true about the Support Vector Machine?
  ☒ ***The SVM is a linear classifier.***
  ☒ ***The SVM separates feature vectors by spanning a hyperplane.***
  ☐ The SVM belongs to the type of classifiers of unsupervised learning.
  ☐ Since there is an infinite number of possible hyperplanes, the SVM converges to any of these possible solutions.

- Figure 1 shows hyperplane A, which is red, and hyperplane B, which is coloured green. Which classifiers could have computed hyperplane A and hyperplane B?
  ☐ A has been computed by the Support Vector Machine and B has been computed by the Perceptron Algorithm.
  ☒ ***A has been computed by the Perceptron Algorithm and B has been computed by the Support Vector Machine.***
  ☐ A and B have been computed by the Support Vector Machine (SVM).
  ☒ **A and B have been computed by the Perceptron Algorithm (unlikely but possible)**



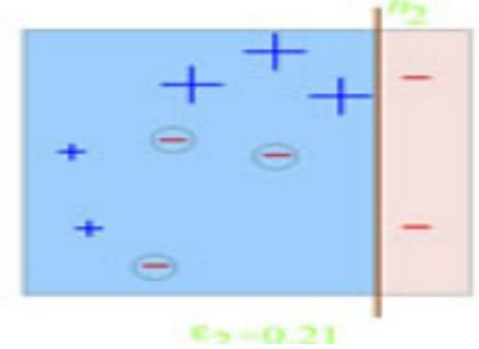
Task 3.3: Boosting (AdaBoost)

- What is the main advantage of AdaBoost over other Boosting algorithms to the time it was invented?
  Answer: Weak classifier's error rate does not need to be known in advance

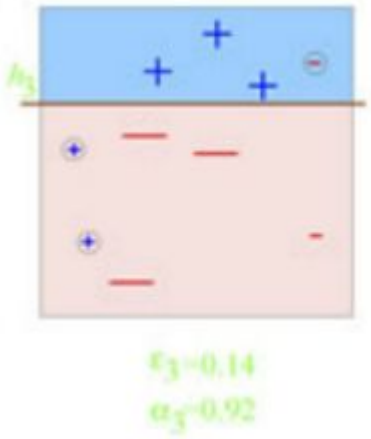
- Which weak learner is usually used for a binary classification task using AdaBoost?
  Answer: classification tree

- Which of these following statements are true for Boosting?
  ☒ **Boosting generates a classification rule out of rough rules of thumb (weak learners)**
  ☐ A learner classifies as weak if the classification accuracy is over 90%
  ☒ **Weak learners get "boosted" into strong learner**

- Which of the following statements are true for AdaBoost?
  ☒ **AdaBoost is a machine learning algorithm**
  ☒ **Wrong classifications get weighted more heavily in the next learning cycle**
  ☐ Initially, some weak learner are weighted more heavily than others
- Draw the next distribution for each data sample by adjusting the size of the cross. The current distribution and low learner are shown in the following image:



Answer:



- Is it possible to use Boosting when you have more than two classes?
  Answer: Yes it is possible if you use it on pairs of the classes

Task 3.4: Evaluation of Classifiers

- Which of the six evaluation possibilities can't handle a case where no positives (true or false) are present?
  Answer: Precision and False-positives per image because of division by 0

- If a classifier evaluation only reports the positives (true or false) which evaluation measures can still be calculated?
  Answer: Precision and False-positives per image

- A classifier is used for classifying exactly one item and it assigns this item the class Positive, but the ground truth of the item is Negative. Give the values of TP, FP, TN, FN!
  Answer: TP:0, FP:1, TN:0, FN:0

- Create a small dataset (<5 items), and assign every item a ground truth and the result of a classifier in such a way that the accuracy of the classifier is 0.75 and TN is 2.
  Answer: For example

Result	Negative	Negative	Positive	Positive
Ground Truth	Negative	Negative	Positive	Negative

- There are 15 elements, where 8 elements are relevant. Only 4 relevant elements were selected. Compute the recall.
  Answer: recall = TP / Relevant = 4 / 8 = 0.5

- There are 5 elements selected, whereas 10 elements are not selected. 4 elements of the selected ones are relevant. Compute the precision.
  Answer: precision = TP / Selected = 4 / 5 = 0.8

- Which one of the paired performance metrics takes into account all four elements of the confusion matrix?
  ☐ *precision*: the ratio of true-positives compared to all detected as positive TP/(TP+FP)
  ☐ *recall*: the ratio of true-positives compared to all actual positives TP/P = TP/(TP+FN)
  ☒ ***accuracy*: the ratio of correct decisions compared to all conditions (TP+TN)/(P+N)**

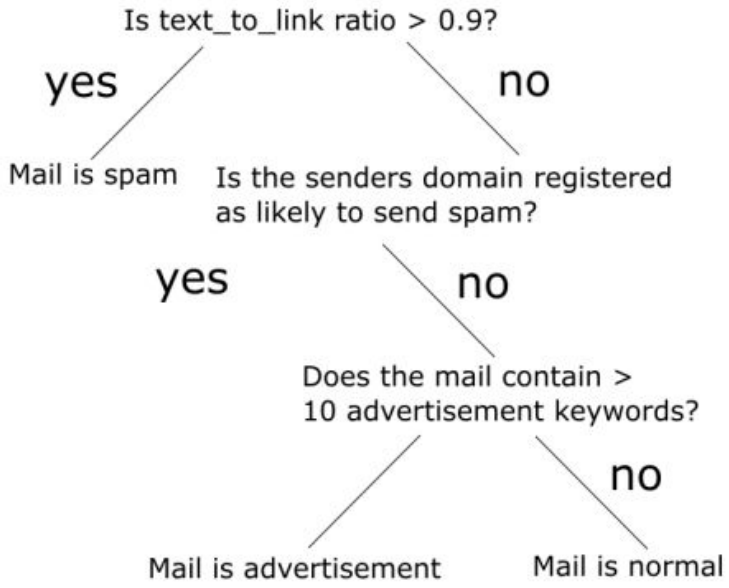
- You want to detect cars in a parking lot. For this you scan each parking space to check if there is a car there or not. In this scenario, how would you describe a *true-negative* decision?
  Answer: Empty parking spaces detected as such.

Unit 4: Decision Trees and Clustering

Task 4.1: Decision Trees – Introduction

- Name a popular method for building a decision tree.
  Answer: C4.5 or C5.0 or ID3 or QUEST

- Design an abstract algorithm how an e-mail service provider could use decision trees



- What do the nodes represent in a decision tree?
  Answer: individual decisions

- What is the difference between internal nodes and leaf nodes?
  Answer: Internal nodes are decisions, leaf nodes form the output/class labels

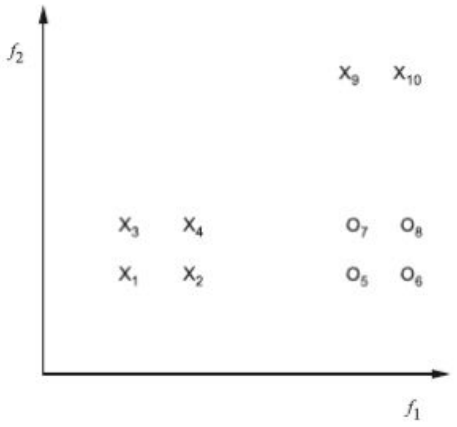
- Which problems can occur, when you train a decision tree?
  ☒ **overfitting**
  ☒ **underfitting**

- In which kind of nodes do class labels appear in decision trees?
  ☐ root-nodes [NO]
  ☐ inner-nodes [NO]
  ☒ **leaf-nodes [YES]**

- What are the strengths of decision trees?
  ☒ **The decision rules of a decision tree are easy to interpret [YES]**
  ☐ Decision trees are easy to create. [NO]
  ☒ **Decision trees, once constructed, are very fast since they require very little computation. [YES]**

- Which of the following statements about decision trees are true?
  ☐ Decision trees can only be used for categorical data
  ☐ To avoid overfitting to training data we must eliminate all impurity in leaf nodes
  ☒ **The measure of entropy is commonly used to measure the impurity of data clusters at the nodes**
  ☐ Decision trees are expensive during classification but their construction is fast

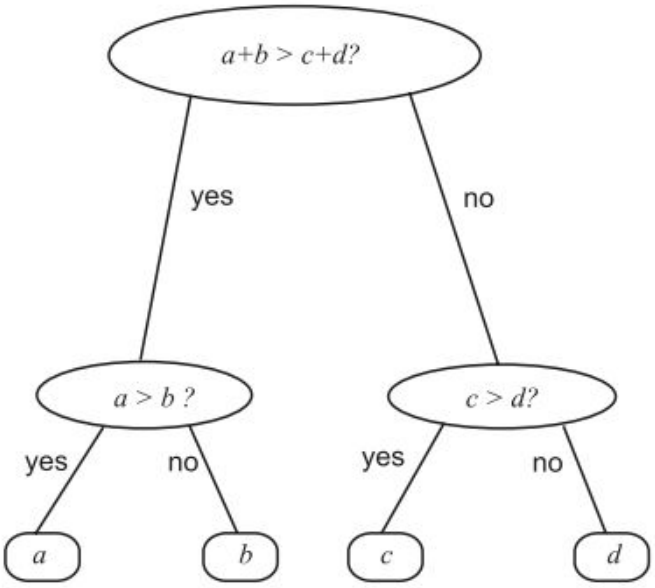
- Draw a binary decision tree that can correctly classify the given data points into the two classes x and o, by splitting the 2D space parallel to the coordinate axes. Provide labels for all nodes and edges and write the data points next to the leaf nodes where they are classified.



- What kind of data can be handled by a decision tree?
  ☒ **data with numerical values**
  ☒ **data with categorical values**
  ☒ **data with ordinal values**
  ☒ **data with missing values**

- Draw a valid binary decision tree for the following situation (Write the decisions and results on the tree): There are four coins a,b,c,d out of which three coins are of equal weight and one coin is heavier.

Answer:



- Are there only binary decision trees? Why?/why not?
  Answer: No, but a non binary decision tree can be reformulated as a binary tree.

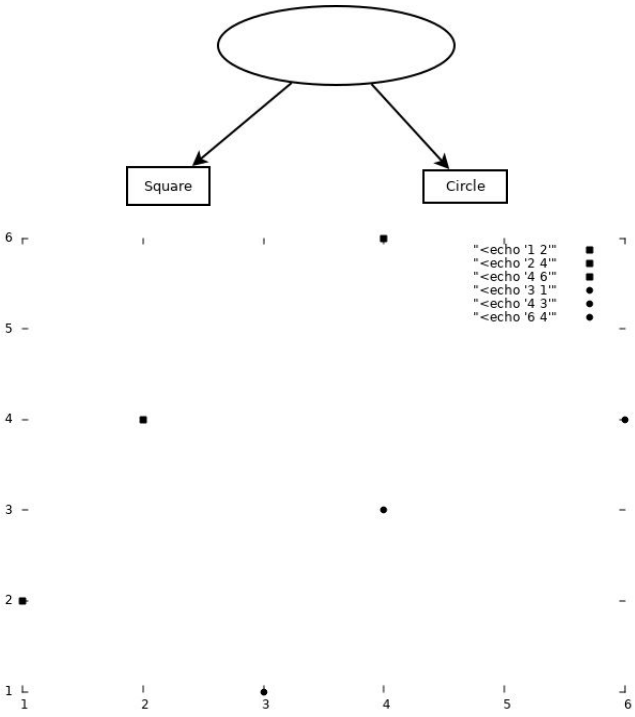
- Create a possible Decision Tree to classify this data:
  1. {v1 : 1, v2 : 2, v3 : 1, v4 : 2, v5 : 1, class : 1}
  2. {v1 : 8, v2 : 2, v3 : 1, v4 : 3, v5 : 1, class : 1}
  3. {v1 : 3, v2 : 2, v3 : 1, v4 : 2, v5 : 1, class : 1}
  4. {v1 : 6, v2 : 2, v3 : 1, v4 : 3, v5 : 2, class : 2}
  5. {v1 : 5, v2 : 1, v3 : 1, v4 : 2, v5 : 2, class : 2}
  6. {v1 : 4, v2 : 1, v3 : 1, v4 : 3, v5 : 2, class : 3}
  7. {v1 : 7, v2 : 1, v3 : 1, v4 : 2, v5 : 2, class : 3}
  8. {v1 : 2, v2 : 1, v3 : 0, v4 : 3, v5 : 2, class : 3}

  Answer (using ternary operators): v2==1 & v1!=5 ? class:3 : v5==1 ? class1 : class2

- Can multiple features be tested in a single node in a decision tree
  Answer: Yes

- A decision tree can be built with:
  ☒ **numerical data.**
  ☒ **nominal data.**
  ☒ **ordinal data.**
  ☒ **categorical data.**

- Consider the following dataset and decision tree. How could the decision look like to separate the circle and square data?



Answer: A simple solution could be "y > x " (note: axes should be labeled)

Task 4.2: Random Forests

- Why is the out-of-bag error important and what is it used for?
  Answer: With it it is possible to assess the generalization error of the individual and combined classifier and also calculate the importance of features.

- What is the difference of using bagging in classification and regression?
  Answer: in classification the result is the class label which got the majority vote of the guesses and in regression it is an average over all guesses.

- What is **not** a benefit of Bagging, compared to only taking one single classifier of the same kind?
  A. Reduced calculation time
  B. Higher accuracy
  C. Solve problems of higher complexity
  Answer: A, because you have to calculate multiple classifiers and then combine them instead of only calculating a single classifier, the calculation time is increased.

- Bagging, two classifiers A,B: A: trained using P1, returns class 'a' on item P1 and 'b' on item P2. B: trained using P2, returns class 'b' on both items. When calculating the out-of-bag error, which training items (P1,P2?) are considered in classifier A?
  Answer: Classifier A considers only P2 for the Out Of Bag error because it was trained on P1.

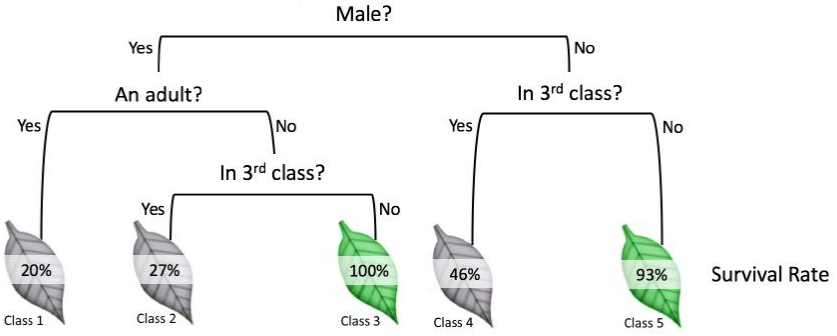
- How does the random forest make a decision?
  Answer: By vote of majority

- Does the random forest use all the features in the learning process?
  ☐ Yes
  ☒ **No**

- Which of the following statements is true for bagging?
  ☐ Bagging sets must all have the same size
  ☒ **Bagging generated new training sets out of a given one**
  ☒ **Bagging stands for Bootstrap aggregating**
  ☐ There is no restriction on how much of the new training set can be build by Replacement

- Dataset:
  - 28y/o male, 1st class passenger -> **Class 1**
  - 7 y/o female, 3rd class passenger -> **Class 4**
  - 65 y/o female, 3rd class passenger -> **Class 4**
  - 21 y/o male, 3rd class passenger -> **Class 1**

The following picture shows a decision tree (survival rates of Titanic passangers). Which of the following Datasets gets sorted into which class based on this decision tree?



Task 4.3: Comparison: Clustering Algorithms

- If you want to cluster cars and this results into 4 clusters, which can be labeled as American cars, European cars, old cars (<2005), new cars (>2005), which of the following types of clustering were probably applied?
  ☒ **Overlapping Clustering**
  ☐ Exclusive Clustering
  ☐ Complete Clustering
  ☒ **Partial Clustering**
  Overlapping Clustering, because cars could be for example European AND old.
  NOT Exclusive Clustering, because that's the opposite of Overlapping Clustering
  NOT Complete Clustering, because there can be items without cluster (2005 and Asian)
  Partial Clustering, because that is the opposite of Complete Clustering

- Which of the following algorithms is **not** a clustering algorithm?
  ☐ k-means
  ☐ Gaussian Mixture Models
  ☒ **kNN**
  ☐ Fuzzy c-means
  kNN is k-nearest-neighbours and performs classification, not clustering (therefore not mentioned in this chapter)

- What does the k in k-means clustering stand for?
  Answer: k is the expected number of clusters

- What is the difference of k-means to fuzzy c-means clustering?

- Each of the following images shows a visualization of one clustering technique. Order the following techniques to the correct images:
  - K-Means Clustering **(Image 2)**
  - Gaussian Mixture Models, GMM **(Image 3)**
  - Hierarchical Clustering **(Image 1)**

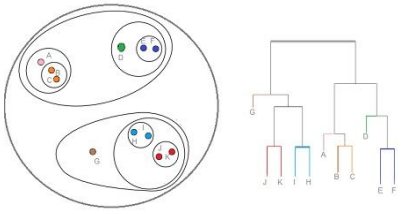


Figure 1: Image1

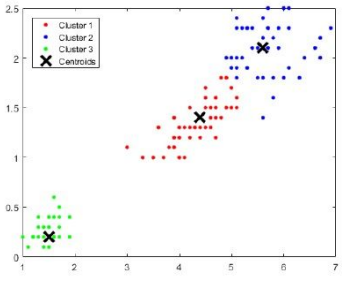


Figure 2: Image2

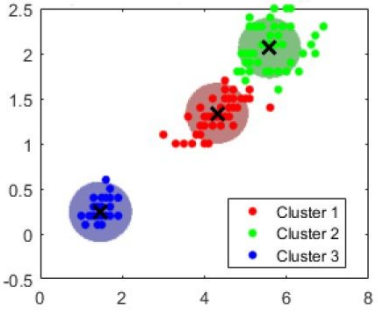


Figure 3: Image3

- Explain what the difference between complete and partial clustering is?
  A complete clustering assigns every data point to a cluster. Partial clustering on the other hand allows some data objects to left alone.

- At which clustering type(s) one object can belong to more than one cluster (class)?
  ☐ Exclusive clustering [NO]
  ☒ **Overlapping clustering [YES]**
  ☒ **Fuzzy clustering [YES]**

- Draw the visualization for one possible result of a k-Means clustering with k=3.
  Note: in actual solution cells should be formed according to Voronoi tessellation

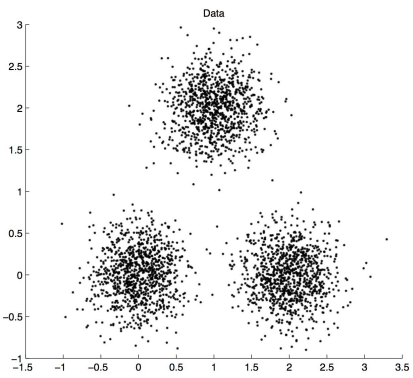


Figure 1: Data

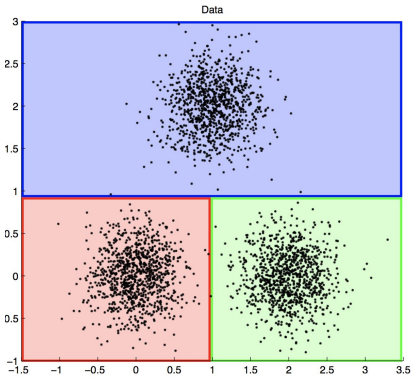
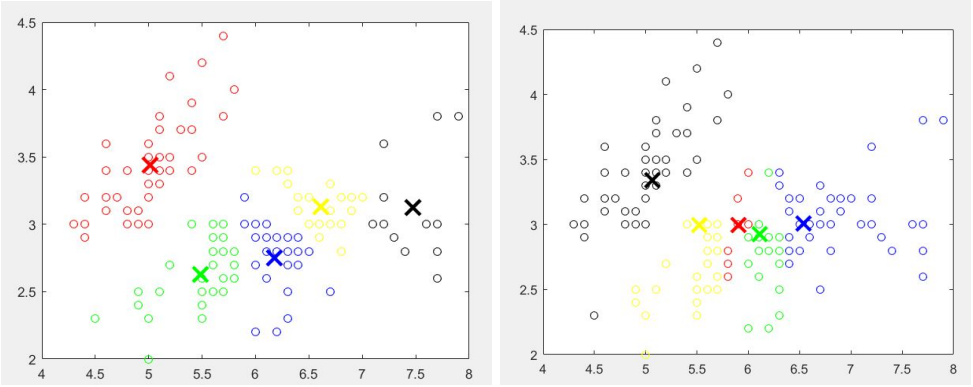


Figure 2: Example solution

- Which clustering algorithm depends on the start position of the centroids?
  Answer: kMeans

- What influence has a high exponent for the fuzzy partition matrix?
  Answer: When you increase the exponent the cluster centers are closer together.
  Left: Exponent is 1.1:, Right: Exponent is 8:



- Describe in a few short points how a basic K-Means algorithm works.
  1. **Select k random centroid points as cluster representatives**
  2. **Calculate the clusters for each point belonging to the closest centroid**
  3. **Calculate new centroids as means of all data points in cluster**
  4. **Repeat from Step 2**
  5. **Stop when changes are under a threshold or certain number of iterations reached**

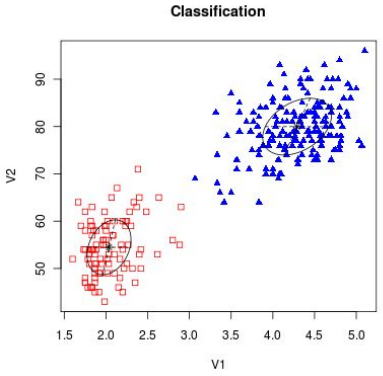
- Describe in a few short points how a basic hierarchical clustering algorithm works.

  1. **Look for the two closest elements and merge them to a single cluster**
  2. **Repeat until desired number of clusters or until only a single cluster remains**

Task 4.4: Parameter Estimation in Gaussian Mixture Models

- Describe Expectation Maximization Algorithm in a few short points.
    1. **Initialize the parameters and evaluate the initial value of the log likelihood**
    2. **E step Evaluate the responsibilities using the current parameter values**
    3. **M step Re-estimate parameters using the current responsibilities**
    4. **Evaluate the log likelihood and check for convergence otherwise return to step 2**

- Plot a simple two dimensional point set where a Gaussian Mixture Model is better to describe the data than a simple Gauss distribution.



- k-means performs overlapping clustering?
  ☐ yes
  ☒ **no**

- Fuzzy clusterings can not be used for classification because they do not provide an unambiguous statement regarding cluster/class membership?
  ☐ they can not be used
  ☒ **they can be used (apply thresholding)**

- Which statements for the Expectation-Maximization Algorithm are <u>false</u>?
  ☐ The Algorithm consists out of 2 main steps.
  ☐ The Algorithm terminates when the likelihood value converges.
  ☒ **The Algorithm is sometimes not defined when the function goes to infinity for some points.**

- Which parameters have to be estimated for Gaussian Mixture Models, when they are defined as $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$?

  $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k)$

  ☐ x_n

☒ π_k

☒ Σ_k

☒ μ_k

- Which two steps are essential for the EM-Algorithm, when a maximum likelihood solutions should be found?
  ☐ Entropy Step
  ☒ **Expectation Step**
  ☒ **Maximization Step**
  ☐ Minimization Step

Task 4.5: Spectral Clustering

- The first three eigenvalues of the graph Laplacian are zero. What does this mean for the similarity graph?
  Answer: There are at least 3 connected components in the similarity graph

- From which two matrices is the graph laplacian calculated?
  Answer: The similarity and the degree matrix

- What is represented by a Similarity Graph?
  A Similarity Graph uses all objects of a data set as nodes and connects all similar objects / nodes with edges.
  OR: **A similarity graph models the local neighborhood relationships between objects (features).**

- Why can Spectral Clustering find clusters with such a complex shape which k-Means alone would never find a good solution?
  Answer: Because Spectral Clustering changes the representation of the objects from the input data to a representation that enhances the cluster-properties in the data so that clusters can be trivially detected by k-Means.

- Given a set of datapoints and their similarities. How is the similarity graph constructed?
  Answer: Each vertex in the graph corresponds to a data point. Two vertices are connected if the similarity between the corresponding data points is positive or larger than a certain threshold, and the edge is weighted based on the similarity.

- Given the matrix U of the eigenvectors of the Laplacian matrix. How are the original datapoints clustered?
  Answer: The i-th datapoint is mapped to the i-th row of U. The rows in U are clustered using k-means. That means that if the i-th row is classified as being part of cluster 1, the i-th datapoint is also classified as being part of cluster 1.

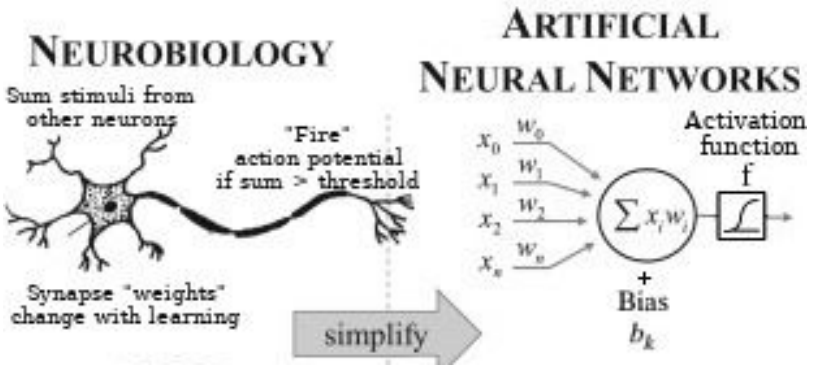
Unit 5: Deep Learning

Task 5.1: Recall Neural Networks

- What property must a pattern fulfill to be learnable by a single layer perceptron?
  Answer: It must be linearly separable

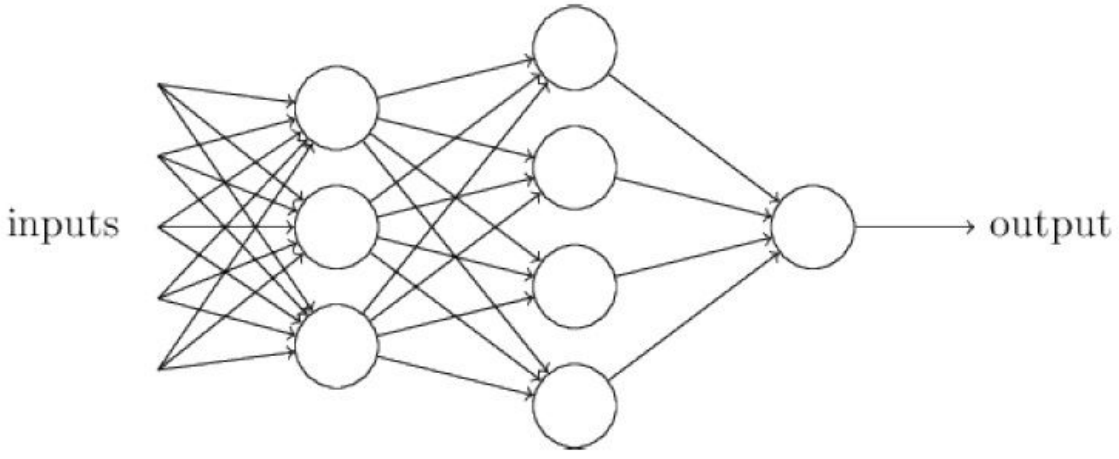
- What property must a typical activation function have?
  Answer: The range of values should be bounded, i.e. should not be open at either side (typical range of values is between -1 and 1 or 0 and 1) and should be monotonically increasing.

- What are the components of a neuron?
  Answer: weighted input, one bias, sum and activation function, output (activation value)
  Note: Bias is often represented via a variable weight w0 for the fixed input x0 = 1.



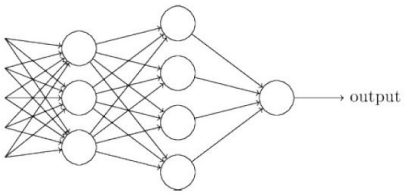
- What is the output of a neural network?
  Answer: mostly an output vector where the dimension correspond to the number of classes in classification problem and the values of the vector elements are between 0 and 1 (highest element value = assigned class). But can also be only a single value (0 or 1)

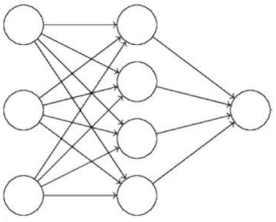
- How many hidden layers does the network shown below contain? **Answer: 2**

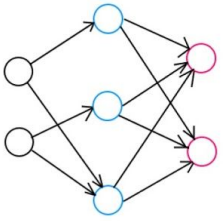


- Which of the shown networks fulfills these requirements:
    - Fully connected
    - One hidden layer
    - Single output

  Answer: (b)







- What is represented by the output values of the output units of a neural network?
  Answer: Each output unit yields the probability of the input data belonging to a certain class.

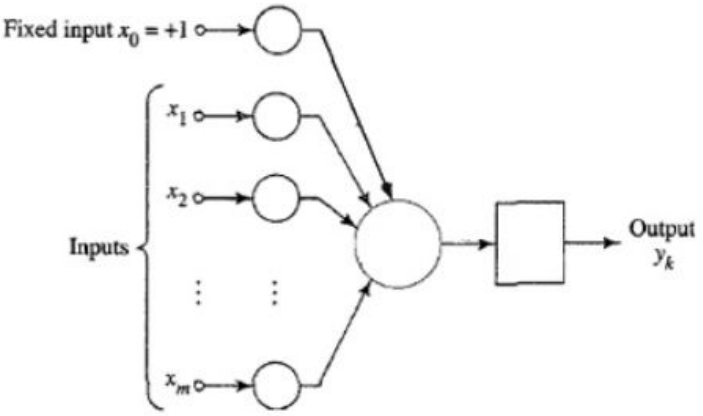
- Between how many classes can a neural network with 5 output units classify at max.?
  Answer: 5

- Is the backpropagation algorithm a supervised algorithm or a unsupervised algorithm?
  Answer: supervised

- Which of the following statements about artificial neural networks are true?
  ☐ The output of an artificial neuron is either 1 or 0
  ☒ **Backpropagation is a supervised training algorithm to update the weights in a neural network such that the error is minimized**
  ☒ **Output layer neurons have weights just like hidden layer neurons**
  ☒ **Without a hidden layer, only linearly separable classifications in input vector space can be made**

- **Answer: Bias is represented via a variable weight w0 for the fixed input x0 = 1.**

In the following illustration of neuron j, put the following elements into the corresponding places: (1) weights $w_{0j} \dots w_{mj}$, (2) $\sum_{i=0}^{m} w_{ij} x_i$, (3) activation function σ. Indicate how the bias is represented and draw and name an example of a an activation function.



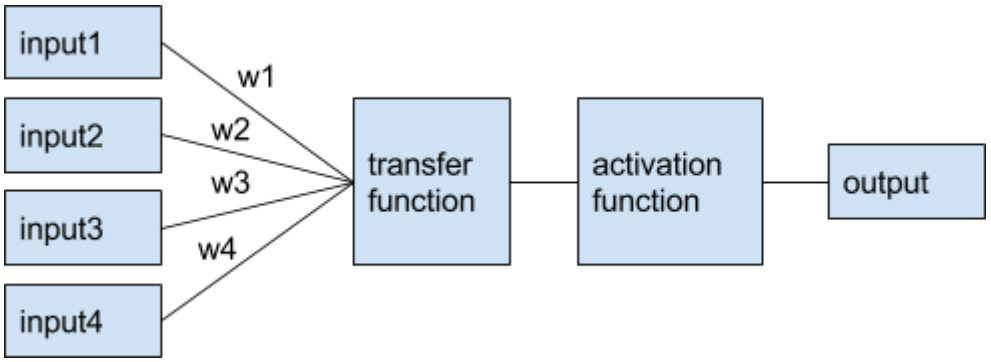
- Name one activation function

Name	Plot	Equation
Identity		$f(x) = x$
Binary step		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}$
Logistic (a.k.a. Soft step)		$f(x) = \frac{1}{1 + e^{-x}}$
TanH		$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$
Softsign [7][8]		$f(x) = \frac{x}{1 + \lvert x \rvert}$
Rectified linear unit (ReLU)[9]		$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
Parameteric rectified linear unit (PReLU)[11]		$f(\alpha, x) = \begin{cases} \alpha x & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases}$
SoftPlus[16]		$f(x) = \ln(1 + e^x)$

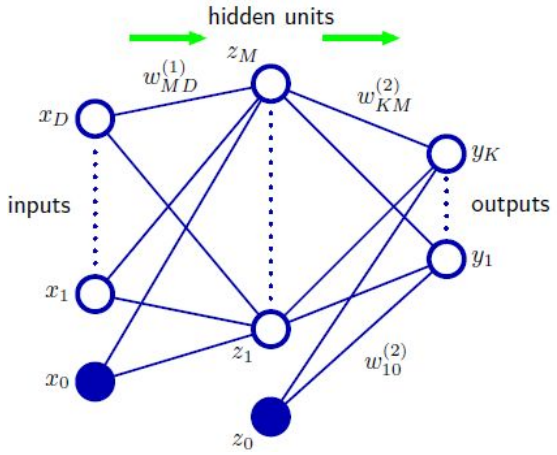
- If there are 5 input nodes, 3 hidden nodes, 1 output node, how many parameters are trainable (no direct connections between input and output nodes)?
  ☒ **> 18 (Answer: 22, 5*3 hidden layer weights + 3 bias, 3*1 output layer weights + 1 bias)**
  ☐ 18
  ☐ < 18

- Draw a simple schema consisting the following elements of an artificial neuron: 4 inputs + weights, transfer function element, activation function element and output (each element is represented as a box).

Answer:



- Draw an example of a single hidden layer neural network with the weights and biases!



- Imagine you designed the following neural network: input 20x20 grayvalue image, 5 hidden units and 2 output units. How many parameters do you need to train?
  Answer: A neural network with input 20x20 grayvalue images, 5 hidden units and 2 output units consists of one input layer, one hidden layer and one output layer. We therefore need one bias for computing the hidden layer (vector with 5 values) and one bias for the output layer (vector with 2 values). Furthermore, we need a 400x5 weight-matrix for the hidden layer and a 5x2 weight matrix for the output layer. We have therefore to compute 400 * 5 + 5 * 2 + 5 + 2 = 2017 parameters.

- What is the difference between a hidden layer and the output layer?
  Answer: Output layer labels data, hidden layer not.

- What do you train in a neural network?
  Answer: Weight matrix and biases (note: biases can be represented as weights, too)

- **What type of activation function is best suited for multi-class tasks?**
  ☐ linear
  ☐ sigmoid
  ☒ **softmax**

- What different type of layers are there in an artificial neural network?
  Answer: Input layer (1) -> hidden layer (1 - *) -> output layer (1)

- Which statements about the Backpropagation Algorithm are true?
  ☒ **It uses the sum squares error.**
  ☐ You don't need a ground truth for it.
  ☒ **The algorithm tries to minimize the cost function.**

- How many parameters have to be trained when the neural network consists of 5 input units, 2 hidden units and 3 output units.
  ☒ **21**
  ☐ 40
  ☐ 10
  ☐ 33

- What is correct about neural networks:
  ☒ **For training a neural network the weights and bias values are changed.**
  ☐ Every input unit has a bias.
  ☐ A neural network is only allowed to have one hidden layer.

Task 5.2: Deep Convolutional Neural Networks (DCNN)

- What is the purpose of (max) pooling layers in CNNs?
  ☐ Feature Extraction.
  ☒ **Dimensionality reduction.**
  ☐ Classification.
  Dimensionality reduction, because they only take one value out of a region (ex, 2x2). Classification is done by the fully connected layer in the end (or alternatively something like a SVM in the last layer). Feature Extraction is done by Convolutional Layers.

- Order the following convolutional neural network layers (input to output):
  - Fully Connected Layer
  - Convolutional Layer
  - Max Pooling Layer

  Answer: Convolutional Layer → Max-Pooling-Layer → Fully Connected. Reason: Max-Pooling needs input to downsample from feature maps generated by convolutional layers. Fully connected layers are usually the last layer in a CNN.

- What is the operation performed in the convolutional layer?
  Answer: convolution operation
  Feature Extraction via Filtering from last layer, i.e. linear combination of pixels with different weights. Same as in normal Neural Network but much faster for pixels.

- What is the output of the ReLU activation function for negative input values? **Answer: 0. Reason: ReLU f(x) = max(0,x)**
  https://en.wikipedia.org/wiki/Rectifier_(neural_networks)

- Which of the following statements are correct?
  ☐ A Convolutional Neural Network needs more training samples than a neural network.
  ☒ **The higher possible depths of a Convolutional Neural Network is an advantage over a neural network.**
  ☒ **Convolutional Neural Networks learn the values of filter kernels.**
  ☒ **Convolutional Neural Networks perform a feature extraction.**
  ☒ **Convolutional Neural Networks perform a classification.**
  ☐ The learned parameters of a neural network can be easily understood by human.

- What does the pooling layer of a Convolutional Neural Network do?
  Answer: It performs a downsampling. It makes a classification more robust against small object movements. Often performed by taking the maximum of some (e.g. 4) signals.

- For a convolutional neural network with an input rgb image of dimensions 256x256x3 and a convolution filter of dimensions 5x5x3, what is the size of the output of this layer?
  Answer: For a single filter: 252x252x1

- For the following matrix write down the result of a maximum pooling layer with a size of 2x2 and a stride of 2.
  Answer: Due to 2x2 block size and stride 2 we have no overlap and we have 4 blocks, where the max values are: 6, 9, 4, 9 (in 'Z' order)

3	6	1	5
1	5	6	9
4	3	7	9
1	2	1	3

6 | 9
----
4 | 9

- Explain the Dropout layer.
  With a user-specified dropout probability this layers can turn neurons of their corresponding functional layer on or off. A turned of neuron will retain its value from the previous iteration and use that for the next time it is turned on again.

- What does the Batch Normalization layer do?

Transforms set of feature responses into a standard normal distribution and is usually used before a ReLU. This transforms an entirely positive or negative range of values to a zero-mean normal distribution so the ReLU won't just activate or deactivate all neurons.

Task 5.3: Neural Network Training for Recognition

- Name some typical classes which occur often in object classification datasets (like CIFAR10).
  Answer: For example car, cat, chair, plane, boat, dog,...

- What is the difference between hidden neurons and other ones?
  Answer: Not in the input and not in the output layer

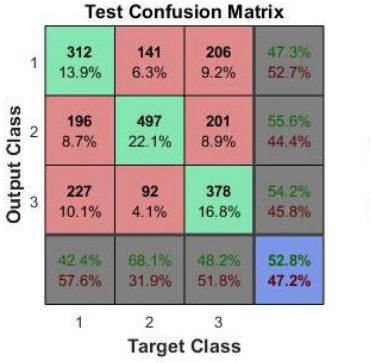
- How can you see that a network is overfitted?
  Answer: Results are too good for the trainingset but to poor for the testset.

- How can you test a neural network?
  Answer: The created NN gets a testset as input and checks if the result and the correct labels match.

- When applying a neural network to a dataset with classes easy to classify for humans, why can it be easier for the network as well?
  Answer: Images are not very similar

- Does the neural network always perform better when increasing the number of hidden layers and hidden units?
  A**nswer: No**

- How many output nodes does the neural network have?



Answer: 3

- A given structure of a neural network always performs consistent, no matter what the input data is?
  ☐ yes

☒ **no: the number of neurons and hidden layers depends on input dimensions**

- For a classification problem the output dimension of a neural network equals the input image resolution?
  ☐ yes
  ☒ **no: the output dimension after softmax typically equals the number of classes**

- 3-input neuron is trained to output a zero when the input is 110 and a one when the input is 111. After generalisation, the output will be zero when and only when the input is:
  - 1. (a) 000 or 110 or 011 or 101
  - 2. (b) 010 or 100 or 110 or 101
  - **3. (c) 000 or 010 or 110 or 100**

Task 5.4: Opinions about Deep Learning

- Which of the following statements are true?
  ☐ Hype in the AI community has been very helpful
  ☒ **Convolutional neural networks are enjoying great popularity**
  ☐ The mammal brain is a helpful analogy to deep learning
  ☐ Unsupervised machine learning is currently dominating the field

- Which of the following statements are true?
  ☐ Ethics are of no concern in regard of deep learning
  ☐ Private corporations and secret algorithms processing private data are a good basis for accountability and rule of law
  ☐ Automated decision making e.g. by machine learning is not regulated in the EU
  ☒ **In the EU, everyone about whom data is being processed for automated decision making has a right to an explanation of that decision**