# Examples of control questions and problems for Pattern Recognition

Problems that have been given at previous examinations are marked by ***. A typical examination would consist of 4-5 such problems or a larger number of simpler problems. Use of books, print-outs and similar will not be allowed at the examination. If a (difficult) formula or a table is needed to solve a given problem, it will be provided. Questions on some of the student presentations may be asked too. For calculations, you may bring a simple electronic calculator to the exam and use it.

**Patterns and pattern recognition**

1. What is a pattern? Which are the main steps in pattern recognition? Give examples of patterns and pattern recognition problems.
2. What do you understand by a feature and a feature vector? Give a few examples.
3. Describe the Universal Product Code (UPC).
4. *** What are the main differences between the Universal Product Code (UPC) and feature vectors extracted from natural objects?
5. What does 'statistical' refer to in 'statistical pattern recognition'? How does it manifest itself?
6. What do you understand by a histogram of the observed values of a feature?
7. What do you understand by the probability density function of a feature?
8. What is the relation between a histogram and the probability density function of a feature?
9. Various terms are used in pattern recognition, depending on the task at hand: classification, detection, identification, authentication. What do they mean?
10. How is the effectiveness of a pattern recognition system measured?
11. What do you understand by a decision boundary? What is the difference between a decision criterion and a decision boundary?
12. What is a training set? What is a test set?
13. What is supervised learning?

**Statistical decision theory (Hypothesis testing). Authentication by iris pattern**

1. *** Assume that you are given a set of 100 000 binary feature vectors, each of which is a binary code of the iris pattern of a person. The set contains 100 iris codes of each of 1000 persons. Describe how you would use this data to design an authentication system based on statistical decision theory.
2. Why can the distribution of the Hamming distance between the iris codes of different persons be described well by a Gaussian function?
3. When we obtain experimental data from the comparisons of iris codes, we can build a histogram of the Hamming distances of pairs of iris codes. Why should we fit such a histogram with a mathematical function (e.g. a Gaussian function)? What do we need this function for?
4. Describe the main components of hypothesis testing in statistical decision theory.
5. What are type I and type II errors? What is 'false acceptance' and 'false rejection rate'?

6. How can we deal with the problem of missing bits (binary features) in an iris code? What are the consequences of missing bits for the choice of an appropriate decision criterion?
7. What do you understand under the term "degrees of freedom" in statistical decision theory?
8. Let us have a vector of n bits. Let each bit take the values 0 and 1 with equal probability. The sum of all bits is a stochastic variable (across different vectors) that is normally distributed and has a certain mean and variance. What is the relation between that mean and variance and n?
9. The comparison of some biometric characteristic (iris, fingerprint, palm, ear shape, voice) of different persons produces a dissimilarity that is normally distributed with some mean and standard deviation. What is the number of independent binary degrees of freedom that can be associated with the observed ratio of the mean and the standard deviation? Given two biometric methods, the one with 30 and the other with 200 independent binary degrees of freedom, which one would you prefer and why?
10. Given are the parameters (mean and standard deviation) of a normal distribution of the dissimilarity of two different persons, using some biometric. Which decision criterion will you use to decide that two persons are identical or different, with a significance level of 0.025?
11. Assume that an iris recognition system extracts iris code of 100 bits that are statistically independent. Assume that each individual bit can take the values 0 and 1 with equal probability (0.5) across the iris codes of different people. Estimate the probability that two iris codes of two different persons agree in more than 60% of the bits. Estimate the same probability for an iris code of 16 bits.

**Bayesian Decision Theory**

1. What do you understand under the following terms: class conditional probability, likelihood, prior probability, evidence, and posterior probability?
2. Explain Bayes rule for the posterior probability and Bayes decision rule.
3. Give a definition of conditional risk, explaining the quantities that participate in that definition.
4. Formulate Bayes decision rule in terms of conditional risk.
5. What is minimum error rate classification and which loss function does lead to it?
6. *** **Naïve Bayes rule.** Give and explain naïve Bayes rule. How would you use this rule to design a spam filter? Give an example.
7. Describe a situation of a two-dimensional two-class problem in which the naïve Bayes approach would fail.
8. *** **Bayesian classification.** A medical test of a disease presents 5% false positives. The disease strikes 1 on 1000 of the population. People are tested at random, regardless of whether they are suspected of having the disease. A patient's test is positive. What is the probability of the patient having the disease?
9. Let x be a feature vector that is extracted from a dark skin spot (mole) that can be a benign nevus or a melanoma mole. Let the posterior probabilities associated with this specific feature vector be P(melanoma | x) = 0.1 and P(benign nevus | x) = 0.9. Let now consider the following two possible actions: $A_1$ – remove the

mole, $A_2$ – do not remove the mole. Let the costs (in Euro) of these actions be: $a_{1,mel}$ = 50 (cost of removing a melanoma mole), $a_{1,nev}$=50 (cost of removing a nevus), $a_{2,mel}$ = 100000 (current and future costs incurred by not removing a melanoma mole), $a_{2,nev}$=0 (current and costs of not removing a nevus).

a. Compute the average expected costs (weighted average over many cases with the same feature vector x) of the two possible actions: $C_1$ of removing a mole and $C_2$ of not removing a mole.
b. Which action will you recommend to take, based on the average costs of the two possible actions?
c. Which class would you assign to this mole, using Bayesian classification?

**Eigenvalues and eigenvectors**

1. Give a definition of an eigenvalue and an eigenvector of a matrix.
2. Compute the eigenvalues and eigenvectors of the following 2x2 matrix …
3. Give an example of a situation in which you can use the concepts of eigenvalue and eigenvector in a pattern classification problem.
4. What is an eigenface and how is this concept used in face recognition?
5. For a 2D normal distribution, the points of equal probability density form ellipses around the mean of the distribution. What is the relation of the eigenvectors and eigenvectors of the co-variance matrix of the distribution to these ellipses?
6. How are eigenvectors used in principal component analysis?
7. Consider the following co-variance matrix of a normal distribution:

   $$\begin{matrix} 3 & -1 \\ -1 & 1 \end{matrix}$$

   Determine the eigenvalues and eigenvectors of this matrix. Assume that the mean of the distribution is in the point (0,0) and draw an ellipse of points around the mean for which the probability density is the same. Draw the eigenvectors of the co-variance matrix to illustrate their relation to this ellipse. For this distribution, define the Mahalanobis distance of a point $(x_1,x_2)$ to the mean (0,0). What is the relation of the Mahalanobis distance to the co-variance matrix of a normal distribution?

**Bayesian Decision Theory – Discriminant functions. Normal distributions**

1. What is a discriminant function and how is it used?
2. What is a dichotomiser and how is it used?
3. Define the whitening transform. What is the result of performing the whitening transform?
4. Consider a bivariate normal distribution with a mean … and a covariance matrix …. Determine the whitening transform for this distribution.
5. Consider a bivariate normal distribution with a mean … and a covariance matrix …. Consider the following linear transform $y = 2x_1 + 3x_2$. Show that y is normally distributed and compute the parameters of its distribution.
6. Consider the following two bivariate normal distributions … with priors … and … Propose discriminant functions that lead to minimum error rate classification. Give an analytical expression for the decision boundary.
7. Given the following normal distributions:

$$p_1(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} \frac{(x-4)^2}{\sigma^2} \right) \qquad p_2(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} \frac{(x-8)^2}{\sigma^2} \right)$$

and equal prior probabilities for the classes:

$P(\omega_1) = P(\omega_2)$

find the decision boundary between the two classes in analytical form.

8. The following two sets of feature vectors:
   S1= {(1,5),(2,4),(3,5),(2,6)}
   S2={(1,7),(3,6),(5,7),(3,8))}
   originate from two bivariate normal distributions.
   a) Estimate the corresponding covariance matrices using ML estimation.
   b) Find the analytical form of the optimal decision boundary between the two classes, assuming equal prior probabilities for the two classes.

9. ***The following two sets of feature vectors originate from two bivariate normal distributions:
   $S_1$= {(0,0), (2,0), (0,2), (2,2), (1,1)}
   $S_2$= {(2,6), (3,3), (5,5), (6,2), (4,4)}

Problems:

a) Estimate the corresponding means and covariance matrices using unbiased maximum likelihood estimation (of the covariance matrices).
b) Find the analytical form of the optimal decision boundary between the two classes, assuming the following relation between the prior probabilities $P_2 = 2P_1$. (Analytical form means an equation.) Draw a sketch of the boundary, together with the data sets and the estimated means.

c) How can you estimate the classification error of this classifier? (If possible, write a mathematical expression.)

*Reminders:*

The pdf of a normal distribution is defined as follows (d is the dimensionality):

$$p(\mathrm{x}) = \frac{1}{(2\pi)^{d/2} \mid \sum \mid^{1/2}} \exp\left[ -\frac{1}{2} (\mathrm{x} - \mu)^t \sum^{-1} (\mathrm{x} - \mu) \right]$$

Relation of a matrix A, its determinant |A|, and its inverse $A^{-1}$:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad A^{-1} = \frac{1}{\mid A \mid} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

|A| = ad – bc

9. ***The following two sets of feature vectors originate from two bivariate normal distributions that represent two different classes:
   Class 1: S1 = {(3,3), (4,1), (4,3), (5,3), (4,5)}
   Class 2: S2 = {(2,7), (4,6), (6,7), (4,8), (4,7)}.

a) Using these feature vectors and the maximum likelihood method, estimate the parameters of the distributions.

b) Find the analytical form of the optimal Bayesian decision boundary between the two classes, assuming equal prior probabilities. Give an approximate drawing of this boundary together with the positions of the means of the distributions and class labels of the regions in which the 2D feature space is divided by the decision boundary.

c) Using the Bayesian decision rule and the parameters of the distributions, classify the following feature vectors: (9,3), (11,6), (5,14), (7,3). Classify the same vectors using nearest neighbor classification and Euclidian distance and compare the results of the two classification methods.

10. *** Let us consider a two-category classification problem, with categories A and B with prior probabilities $P_A$ and $P_B$. The class-conditional probability densities $p_{x|A}$ and $p_{x|B}$ are one-dimensional normal distributions:

$$p_{x|A} \sim N(\mu_A, \sigma^2_A), \qquad p_{x|B} \sim N(\mu_B, \sigma^2_B)$$

a) Express analytically the position(s) of the optimal Bayesian decision boundary or boundaries in terms of $P_A$, $\mu_A$, $\sigma_A$, $P_B$, $\mu_B$, $\sigma_B$.

b) Find the analytical conditions for having 0, 1, 2, or 3 decision boundaries. For each possible case, draw qualitative graphs of the posterior probability functions $P_A p_{x|A}$ and $P_B p_{x|B}$, which illustrate why the number of decision boundaries depends on the parameters $P_A$, $\mu_A$, $\sigma_A$, $P_B$, $\mu_B$, $\sigma_B$.

c) Let us consider the sets of observations {-2, -1, 0, 1, 2} for category A and {3.2, 4.1, 5, 5.9, 6.8} for category B.

c1) Compute *unbiased* maximum likelihood estimations of $\mu_A$, $\sigma_A$, $\mu_B$, $\sigma_B$.

c2) Show that the condition $P_A = P_B$ gives rise to two decision boundaries and compute their positions $x_1$ and $x_2$.

c3) Are both decision boundaries in c2 equally relevant? Justify your answer.

**Bayesian Decision Theory – Binary features**

1. Let p be the probability that a given keyword, such as viagra, appears in a spam email, and let q be the probability that this word appears in a normal email. Let f = 1 if the keyword appears in an email and let f = 0 if it does not appear. Let the prior probabilities of spam and normal email are P(s) and P(n), respectively. Give an expression for the posterior probability P(spam | f). Give such an expression P(spam | $f_1$, $f_2$) for the case of two statistically independent keywords with class conditional probabilities of appearing in spam $p_1$ and $p_2$, respectively, and of appearing in a normal mail $q_1$ and $q_2$, respectively.

**Bayesian classification. Missing features**

1. *** Consider a two-dimensional, three-category pattern classification problem, with equal priors $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$. We define the 'disk distribution' $D(\boldsymbol{\mu},r)$ to be uniform inside a circular disk centered on $\boldsymbol{\mu}$ and having radius r, and elsewhere 0. The class-conditional probabilities for the three categories are such disk distributions $D(\boldsymbol{\mu}_i,r_i)$, i = 1, 2, 3, with the following parameters:
$\omega_1$: $\boldsymbol{\mu}_1 = (3, 2)$, $r_1 = 2$; $\omega_2$: $\boldsymbol{\mu}_2 = (4, 1)$, $r_2 = 1$; $\omega_3$: $\boldsymbol{\mu}_3 = (5, 4)$, $r_3 = 3$.

   a) Classify the points (6, 2) and (3, 3) with minimum probability of error.
   b) Classify the point (*, 0.5), where * denotes a missing feature.

2. *** Consider a two-dimensional, three-category pattern classification problem, with priors $P(\omega_1) = 0.5$, $P(\omega_2) = 0.25$, $P(\omega_3) = 0.25$. We define the 'square distribution' $S(\boldsymbol{\mu},a)$ to be uniform inside a square of size a × a centered on $\boldsymbol{\mu}$, and elsewhere 0. The sides of the square are parallel to the coordinate axes. The class-conditional probabilities for the three categories $\omega_1$, $\omega_2$, and $\omega_3$ are such square distributions $S(\boldsymbol{\mu}_i,a_i)$, i = 1, 2, 3, with the following parameters:
$\omega_1$: $\boldsymbol{\mu}_1 = (0, 0)$, $a_1 = 3$; $\omega_2$: $\boldsymbol{\mu}_2 = (-1, 1)$, $a_2 = 1$; $\omega_3$: $\boldsymbol{\mu}_3 = (1, -1)$, $a_3 = 2$.

   a) Classify the points (1,1), (0.5,-0.5) and (-0.7,1) with minimum probability of error.
   b) Classify the patterns (*,1) and (1,*), where * denotes a missing feature.

**3.** *** Consider a two-dimensional, three-category pattern classification problem with equal priors $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$. We define the 'disk distribution' $D(\boldsymbol{\mu},r)$ to be uniform inside a circular disk centered on $\boldsymbol{\mu}$ and having radius r, and elsewhere 0. The class-conditional probabilities for the three categories are such disk distributions $D(\boldsymbol{\mu}_i,r_i)$, i = 1, 2, 3, with the following parameters:

   $\omega_1$: $\boldsymbol{\mu}_1 = (3, 3)$, $r_1 = 2$; $\omega_2$: $\boldsymbol{\mu}_2 = (4, 2)$, $r_2 = 1$; $\omega_3$: $\boldsymbol{\mu}_3 = (5, 5)$, $r_3 = 3$.

   a) Classify the points (6, 3) and (3, 4) with minimum probability of error.
   b) Classify the point (*, 1.5), where * denotes a missing feature.

   *Hint*: Draw the three disks and the points to be classified in a 2D feature space.

**Classification error. Receiver operating characteristics**

1. Define the classification error probability for a two-class problem.
2. What is a dichotomizer?
3. What is a reducible error? How large is this error for a two-class problem, if the optimal Bayes decision boundary is used?
4. Define the probabilities of hit, false alarm, miss and correct rejection in a two-category classification problem.
5. What is discriminability? How can the discriminability of a two-category classification system be experimentally determined?

6. *** What do you understand by a receiver operating characteristics (ROC)? To which class of problems does it apply? What is the common property of points situated on the same ROC curve?
7. Given the following two normal distributions:

$$p_1(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} \frac{(x-4)^2}{\sigma^2} \right) \qquad p_2(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} \frac{(x-8)^2}{\sigma^2} \right)$$

and equal prior probabilities for the two classes:

$$P(\omega_1) = P(\omega_2)$$

determine the classification error of an optimal Bayesian classifier. (You first need to find the decision boundary between the two classes.)

**ML and Bayesian Parameter Estimation**

1. What do you understand by parameter estimation? When does it apply?
2. Define Maximum Likelihood (ML) estimation. What are the values of the mean and standard deviation estimated using ML in case of a univariate normal distribution? How are the corresponding formulae derived?
3. What is a biased estimator? Give an example.
4. *** Consider a one-dimensional two-category classification problem with equal priors, $P(\omega_1) = P(\omega_2) = 0.5$, where the class conditional probability densities have the form

   $p(x|\omega_i) = 0$ for $x < 0$, $\quad p(x|\omega_i) = \theta_i \exp(-\theta_i x)$ for $x \geq 0$,

   where $\theta_1$ and $\theta_2$ are positive but unknown parameters.

a) Show that the distributions are normalized.
b) The following data are collected: $D_1 = \{1, 5\}$ and $D_2 = \{3, 9\}$ for $\omega_1$ and $\omega_2$, respectively. Find the maximum-likelihood values of $\theta_1$ and $\theta_2$.
c) Given your answer to part b), determine the decision boundary $x^*$ for minimum classification error. Indicate which category is to the right and which to the left of $x^*$.
d) What is the expected error of your classifier in part c)?

5.*** The task is to use Bayesian methods to estimate a one-dimensional probability density. The fundamental density function is a normalized triangle distribution $T(\mu,1)$ with center at $\mu$ and half-width equal 1, defined by

$$p(x|\mu) \sim T(\mu,1) = 1 - |x - \mu| \text{ for } |x - \mu| \leq 1,$$
$$= 0 \text{ otherwise}.$$

The prior information on the parameter $\mu$ is that it is equally likely to take any of the three discrete values -1, 0 or 1.

a) Plot the 'estimated density' before any data are collected (which we denote by $D_0 = \{\}$). That is, plot $p(x|D_0)$. Here and below, be sure to label and mark your axes and ensure normalization of your final estimated density.
b) The single point $x = 0.25$ was sampled, and thus $D_1 = \{0.25\}$. Plot the estimated density $p(x|D_1)$.

c) Next the point $x = 0.75$ was sampled, and thus the data set is $D_2 = \{0.25, 0.75\}$. Plot the estimated density $p(x|D_2)$.

**6.** *** Let us consider a two-category classification problem, with categories A and B with prior probabilities $P_A = 1/3$ and $P_B = 2/3$. The class-conditional probability densities $p_{x|A}$ and $p_{x|B}$ are one-dimensional normal distributions:

$$p_{x|A} \sim N(\mu_A, \sigma^2_A), \qquad p_{x|B} \sim N(\mu_B, \sigma^2_B)$$

Let us consider the sets of observations {-1, 0.5, 1, 1.5, 3} for category A and {2, 3.5, 4, 4.5, 6} for category B.

a) Compute *unbiased* maximum likelihood estimations of $\mu_A$, $\sigma_A$, $\mu_B$, $\sigma_B$.
b) Plot sketches of the two probability density functions.
c) Classify the following points: -2, 0, 2, 3, 5, 7.

**Nonparametric classification techniques**

1. What do you understand by nonparametric classification techniques? When can these techniques be applied?
2. *** Explain, using a simple example, the density estimation with Parzen windows. Under which conditions does this method give reliable results?
3. What is k-nearest neighbor pdf estimation? When can it be applied? Under which conditions does this method converge?
4. What is the relation between k-nn classification and Bayesian classification?
5. You have a data set of feature vectors that belong to different classes. Each feature vector has a class label. How can you use this data set to determine the value of the parameter k which should be used in order to achieve optimal classification results?
6. What is a Voronoi tesselation? How can it be used in classification?
8. What metrics can one use in nearest neighbor classification?
9. Consider the following set of labeled patterns from … different classes: … Using … distance and ?-nearest-neighbor classification, classify the following patterns: …
10. Problem 9 page 202 text book of Duda et al.
11. *** Consider the following two sets of training patterns from two different classes:
    Class 1: $S_1$ = {(7,31), (8,32), (10,32), (6,31), (5,32), (4,28), (5,30)}
    Class 2: $S_2$ = {(10,31), (8,29), (9,33), (10,32), (14,33), (12,31), (11,30), (9,30)}
    Using city block distance and 3-nearest-neighbour classification, classify the following test patterns: (8,31), (5,29), (9,30). If for a given test pattern more than three training patterns fall in its neighborhood defined by the first three nearest neighbors from the training sets, use all training patterns falling in this neighborhood to determine the class.
12. *** Present the LVQ algorithm.
13. *** Describe the relevance LVQ algorithm.
14. Give a list of advantages and disadvantages of non-parametric classifiers versus parametric classifiers.

**Cross validation.**

**1.** *** Consider a data set that includes 1000 128-dimensional feature vectors that come from four different categories. Using this data set, you want to construct a k-nearest neighbor classifier.

a) Describe how you can select the value of the parameter k using cross validation.

b) What is over-fitting and how can you detect and prevent it using cross validation?

c) Which types of cross validation can you use and what are their advantages and disadvantages?

**Unsupervised learning and clustering.**

1. What do you understand by unsupervised learning and clustering?
2. In which situations is clustering useful? Give a short example of one such situation.
3. Present the k-means algorithm (Lloyd's algorithm). Which quantity is referred to as the quantization error? Demonstrate that the algorithm converges.
4. What method can you use to automatically determine an appropriate value of the parameter k for the k-means clustering algorithm?
5. Describe the Fuzzy k-means clustering algorithm. Which cost function is minimized in this algorithm?
6. Name and illustrate problems that can be encountered by Lloyd's algorithm for k-means clustering.
7. Describe the vector quantization algorithm for clustering. What are the similarities and differences between this algorithm and Lloyd's algorithm for k-means clustering?
8. *** Present shortly the fuzzy k-means algorithm. What are the differences between k-means and fuzzy k-means?
9. How can you use the quantization error of the k-means algorithm to determine the number of clusters in a data set?
10. *** Consider the application of the *k*-means clustering algorithm to the one-dimensional data set $D = \{0, 1, 5, 8, 14, 16\}$ for $k = 3$ clusters.
    a) Start with the following three cluster means: $m_1(0) = 2$, $m_2(0) = 6$ and $m_3(0) = 9$. What are the values of the means at the next iteration?
    b) What are the final cluster means after convergence of the algorithm?
    c) For your final cluster means, to which cluster does the point $x = 3$ belong? To which cluster does $x = 11$ belong?
11. *** Consider the following set of data points: S = {(1, 9.5), (3, 8.5), (4, 8), (7, 6.5), (9, 5.5), (14, 3), (16, 2), (17, 1.5)}. Explain the vector quantization algorithm for clustering using three prototypes (cluster centroids) initialized as follows: (0, 6), (8, 8), (15, -5) and learning rate $\eta = 0.3$. Compute the new positions of the prototypes for one epoch and the assignment of points to clusters after that epoch. *Hint*: Plot the data.
12. Describe the neural gas clustering algorithm.

13. Describe a clustering algorithm that deploys distance-based connected components.
14. ***K-means clustering. Vector quantization.**

a) Present Lloyd's algorithm for k-means clustering.
b) Which function is minimized in the k-means problem?
c) Show that this algorithm converges.
d) Name problems encountered by Lloyd's algorithm for k-means clustering.
e) Describe the gap statistics method to determine the appropriate number of clusters in a data set.
f) Consider the application of the *k*-means clustering algorithm to the one-dimensional data set $D = \{-3, -2, 2, 5, 11, 13\}$ for $k = 3$ clusters. Start with the following three cluster means: $m_1(0) = -1$, $m_2(0) = 3$ and $m_3(0) = 6$. What are the values of the means at the next iteration? What are the final cluster means and clusters after convergence of the algorithm?
g) Sketch the vector quantization algorithm for clustering. What are the similarities and dissimilarities between k-means clustering and vector quantization clustering? Comment on their advantages and disadvantages.

**Hierarchical clustering.**

15. What is a dendrogram? In which situations is appropriate to use it for characterizing data?
16. Write in pseudocode the algorithm for agglomerative hierarchical clustering.
17. What is the difference between single-linkage agglomerative algorithm and the complete-linkage algorithm?
18. Consider the following matrix of dissimilarities between four objects. Using hierarchical clustering and the single-linkage algorithm, built a dendrogram.
19. *** Construct a cluster dendrogram for the one-dimensional data D = {2, 3, 5, 10, 13}. As a distance between two clusters $D_i$ and $D_j$ use the maximum distance between a point from $D_i$ and a point from $D_j$ , for all possible pairs of such points.
20. *** Consider the following set of points S = {(1, 9.5), (3, 8.5), (4, 8), (7, 6.5), (9, 5.5), (14, 3), (16, 2), (17, 1.5)}. The dissimilarity between two points is defined as the Euclidean distance between them. The dissimilarity between two clusters of points is defined by the dissimilarity of their least dissimilar elements.
    b) Build a dendrogram for this set.
    c) Using the dendrogram, cluster the points in two clusters.
    d) Using the dendrogram, group the points in three clusters.
    e) *Hint*: Using the dissimilarity matrix approach will take you a lot of time. You can build the dendrogram faster if you plot the data and decide visually how to cluster data.
21. *** Construct two cluster dendrograms for the one-dimensional data S = {1, 3, 6, 10, 16}
    a) using the distance measure $d_{max}(S_i,S_j) = \max_{x \in S_i,\, x' \in S_j} |x-x'|$,
    b) using the distance measure $d_{min}(S_i,S_j) = \min_{x \in S_i,\, x' \in S_j} |x-x'|$.

**Independent component analysis.**

1. What is meant by ‘independent components’?
2. Formulate the concept of statistical independence.
3. Given a set of two-dimensional feature vectors, how can you conclude if the two features are statistically independent or not?
4. What is meant by ‘source signals’ and by ‘mixed signals’?
5. Consider the scatter plot of 2D feature vectors shown in Fig.X. The covariance matrix of this data set is diagonal. Are the two features $x_1$ and $x_2$ independent? Can you infer from the figure which the independent components are? Give arguments in support of your answers.
6. Name measures of deviation from the normal distribution (non-Gaussianity).
7. What is the statistic-theoretical basis of the ICA method?
8. Describe an ICA application.

**Support vector machines.**

1. Consider two linearly separable data sets of 2D feature vectors. Draw a linear decision boundary and define the margin in the sense of SVM. Give an expression for the margin as a function of the parameters of the linear decision boundary. Formulate the SVM method as an optimization problem.
2. Give an example of two data sets that are not linearly separable but can become linearly separable after mapping them to a higher-dimensional space.
3. Explain the concept of a kernel function in the context of SVM.