"Machine Learning and Computational Statistics"

8th Homework

(A) k-NN classifier

Exercise 1 (**mult. choice question):** Which of the following statements regarding the k-nearest neighbor (k-NN) classification rule is/are true?

1. The k-NN rule is a parametric classifier.

2. The parameter k in the k-NN rule results after the solution of an appropriately defined optimization problem.

3. The k-NN rule requires knowledge of the probability density functions of the classes.

4. There is no training phase for the k-NN classification rule.

Exercise 2 (**mult. choice question):** Consider an M-class classification task and let $Y = \{(y_i, \boldsymbol{x}_i), \boldsymbol{x}_i \in R^l, y_i \in \{1,2,\dots,M\},\ i = 1,\dots,N\}$ be the available data set associated with it. Let $\boldsymbol{x}$ be a vector that is to be classified to one of the M classes, using the k-nearest neighbor (k-NN) classification rule. Which of the following statements is/are true?

1. The k-NN rule finds the set $S_j, j = 1,\dots,M$, of the k-nearest neighbors of $\boldsymbol{x}$ from each class, it determines the mean distance, d_j, of $\boldsymbol{x}$ with the points in each $S_j, j = 1,\dots,M$, and assigns to the class that corresponds to the minimum d_j.

2. The k-NN rule determines the set S of the k nearest neighbors of $\boldsymbol{x}$ (among all the vectors in Y from all classes), it counts how many of the vectors in S stem from each category and assigns $\boldsymbol{x}$ to the class to which the majority of the vectors in S belong.

3. For $k = 1$, the k-NN rule assigns $\boldsymbol{x}$ to the class where its closest vector in Y belongs.

4. For $k = 3$, if the set S of the nearest neighbor of $\boldsymbol{x}$ consists of three training points from different classes, then $\boldsymbol{x}$ remains unclassified.

Exercise 3 (**mult. choice question):** Consider four classification tasks for which only classification error probability values less than 0.1 are acceptable. Consider the nearest neighbor (NN) classification rule ($k = 1$) and the Bayes classifier. Let P_{NN} and P_B denote the respective classification error probabilities. The classification error probabilities P_B of the Bayes classifier, for each of the four classification tasks, are shown below. Taking into account the asymptotic bounds of P_{NN}, that is, $P_B \leq P_{NN} \leq 2 \cdot P_B$, in which of these cases, the NN rule is expected to fulfill this requirement?

1. $P_B = 0.3$

2. $P_B = 0.03$

3. $P_B = 0.09$

4. $P_B = 0.002$

Exercise 4 (**mult. choice question):** Consider a specific M-class classification task and let N be the number of the available training points that stem from the M classes. Also, let P_{NN}, P_{kNN} and P_B be the classification error probabilities for the nearest neighbor (NN), the k-nearest neighbor (k-NN) and the Bayes classifier, respectively. Which of the following statements is/are true?

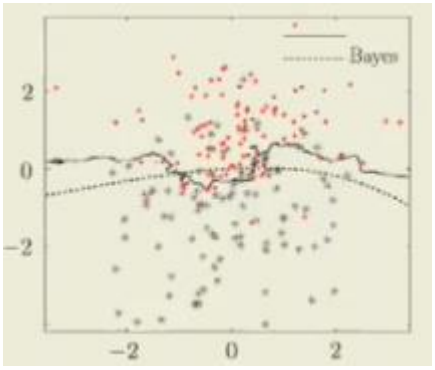
1. As $N \to \infty$, $k \to \infty$ and $k/N \to 0$, P_{NN} tends to P_B.

2. As $N \to \infty$, P_{kNN} tends to P_B.

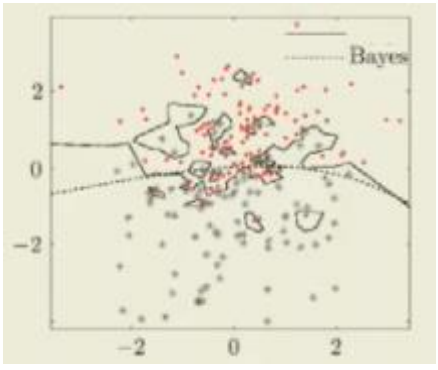
3. As $N \to \infty$, P_{NN} tends to $3 \cdot P_B$.

4. As $N \to \infty$, $k \to \infty$ and $k/N \to 0$, P_{kNN} tends to P_B.

Exercise 5 (**mult. choice question):** Consider a two-class classification task, where the data points from the two classes are denoted via red and black colors (see figure below). The dashed line is the decision curve obtained by the Bayes classifier and the solid lines are the ones obtained by the nearest neighbor (NN) and the 13-nearest neighbor (13-NN) classifiers. Which of the following statements is/are true?



(a)



(b)

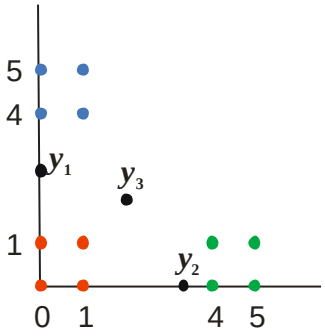
1. The solid line in Fig. (a) corresponds to the NN classifier.

2. The solid line in Fig. (a) corresponds to the 13-NN classifier.

3. The solid line in Fig. (b) corresponds to the NN classifier.

4. The solid line in Fig. (b) corresponds to the 13-NN classifier.

Exercise 6 (**mult. choice question):** Consider a three-class classification problem for which four points from each class are available, denoted by different colors in the following figure (red, green, blue correspond to classes ω_1, ω_2 and ω_3, respectively). Specifically, the data points $[0,0]^T$, $[1,0]^T$, $[0,1]^T$ and $[1,1]^T$ belong to class ω_1, the data points $[4,0]^T$, $[4,1]^T$, $[5,0]^T$ and $[5,1]^T$ belong to class ω_2 and the data points $[0,4]^T$, $[0,5]^T$, $[1,4]^T$ and $[1,5]^T$ belong to class ω_3. The points $\boldsymbol{y}_1 = [0, 2.8]^T$, $\boldsymbol{y}_2 = [3.2, 0]^T$ and $\boldsymbol{y}_3 = [2,2]^T$, are to be classified via the 5-nearest neighbor rule, utilizing the squared Euclidean as distance between two data points. Which of the following statements is/are true?



1. The points $\boldsymbol{y}_1, \boldsymbol{y}_2$ and $\boldsymbol{y}_3$ are classified to the classes ω_3, ω_2 and ω_1, respectively.

2. The points $\boldsymbol{y}_1, \boldsymbol{y}_2$ and $\boldsymbol{y}_3$ are classified to the classes ω_1, ω_2 and ω_3, respectively.

3. The points $\boldsymbol{y}_1, \boldsymbol{y}_2$ and $\boldsymbol{y}_3$ are classified to the classes ω_3, ω_1 and ω_2, respectively.

4. The points $\boldsymbol{y}_1, \boldsymbol{y}_2$ and $\boldsymbol{y}_3$ are classified to the classes ω_2, ω_3 and ω_1, respectively.

Exercise 7 (**mult. choice question):** Which of the following statements is/are true for the k-nearest neighbor (k-NN) classification rule, when applied to a specific classification task?

1. The decision surfaces that separate the class regions remain unaltered, independently of the choice of the specific distance measure between two data vectors.

2. The classification of a single data point requires the consideration of all the data vectors in the training set.

3. The time required for the training of the k-NN classifier is usually very large.

4. In general, the decision surfaces that separate the class regions, produced by the k-NN classifier, vary as k increases.

(B) Logistic regression

Exercise 8 (**mult. choice question):** Which of the following statements concerning logistic regression (LR) is/are true?

1. LR, as its name implies, is a regression method.

2. LR is a generative modeling approach.

3. In LR, the probability distribution of the data is not taken into account.

4. LR models the class prior probabilities.

Exercise 9 (**mult. choice question):** Consider a two-class classification task, where the classes are denoted as ω_1 and ω_2. Which of the following expressions, which relate the posterior probabilities for a data vector $\boldsymbol{x}$ in the framework of logistic regression (LR), is/are valid?

1. $\ln \frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = \boldsymbol{\theta}^T \boldsymbol{x}$

2. $\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = \boldsymbol{\theta}^T \boldsymbol{x}$

3. $\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = -\exp(\boldsymbol{\theta}^T \boldsymbol{x})$

4. $\frac{P(\omega_2|\boldsymbol{x})}{P(\omega_1|\boldsymbol{x})} = \exp(-\boldsymbol{\theta}^T \boldsymbol{x})$

Exercise 10 (**mult. choice question):** Consider a two-class classification task, where the classes are denoted as ω_1 and ω_2. Let (H) be the hyperplane defined by the equation $\boldsymbol{\theta}^T \boldsymbol{x} = 0$ (the offset θ_0 is absorbed in $\boldsymbol{\theta}$ and $\boldsymbol{x}$ is redefined accordingly). Let $\boldsymbol{x}'$ be a specific data point and let $\ln \frac{P(\omega_1|\boldsymbol{x}')}{P(\omega_2|\boldsymbol{x}')} = \boldsymbol{\theta}^T \boldsymbol{x}'$, where $P(\omega_j|\boldsymbol{x}')$, $j = 1,2$, are the posterior class probabilities given $\boldsymbol{x}'$. Which of the following statements is/are true?

1. If $P(\omega_1|\boldsymbol{x}') > P(\omega_2|\boldsymbol{x}')$, then $\boldsymbol{x}'$ lies on the positive side of (H).

2. If $P(\omega_1|\boldsymbol{x}') < P(\omega_2|\boldsymbol{x}')$, then $\boldsymbol{x}'$ lies on the nonnegative side of (H).

3. If $\boldsymbol{\theta}^T\boldsymbol{x}' > 0$, then $P(\omega_2|\boldsymbol{x}') > P(\omega_1|\boldsymbol{x}')$.

4. The logistic regression is a linear classifier.

Exercise 11 (**mult. choice question):** Consider a two-class classification task, where the classes are denoted as ω_1 and ω_2. Consider a specific data point $\boldsymbol{x}$. In the framework of logistic regression, the class posterior probabilities with respect to $\boldsymbol{x}$ are related via the expression $\ln\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = \boldsymbol{\theta}^T\boldsymbol{x}$. Which of the following expressions, related to the class posterior probabilities, is/are valid?

1. $P(\omega_1|\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x})}$

2. $P(\omega_1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$

3. $P(\omega_2|\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x})}$

4. $P(\omega_2|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x})}$

Exercise 12 (**mult. choice question):** Which of the following statements related to the logistic function $\sigma(t)$ is/are true?

1. Its domain is the interval $(0,1)$.

2. It holds: $lim_{t\to\infty}\,\sigma(t) = 1$ and $lim_{t\to-\infty}\,\sigma(t) = 0$

3. It holds: $\sigma(0) = 0$.

4. The value set of $\sigma(t)$ is the set of the real numbers.

Exercise 13 (**mult. choice question):** The derivative of the sigmoid function, $\sigma(t) = \frac{1}{1+exp(-t)}$, is equal to:

1. $\sigma(t)\big(1-\sigma(t)\big)$

2. $\sigma(t)\big(1+\sigma(t)\big)$

3. $\frac{\sigma(t)}{1-\sigma(t)}$

4. $\sigma^2(t)\big(1-\sigma(t)\big)$

Exercise 14 (**mult. choice question):** Consider a two-class classification task, where the classes are denoted as ω_1 and ω_2. Assume that we have at our disposal a data set $X=\{(y_n,\boldsymbol{x}_n)\colon \boldsymbol{x}_n \in R^l,\ y_n \in \{0,1\}, n=1,\dots,N\}$, where $y_n=1$ (resp. 0) if $\boldsymbol{x}_n \in \omega_1$(resp. ω_2). Consider the logistic regression framework, which is based on the equation $\ln\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}=\boldsymbol{\theta}^T\boldsymbol{x}$, where $P\big(\omega_j|\boldsymbol{x}\big)$, $j=1,2$, are the class posterior probabilities, with respect to $\boldsymbol{x}$**,** and $\boldsymbol{\theta}$ the parameter vector that defines the classifier. Which of the following statements is/are true?

1. $P(\omega_1|\boldsymbol{x}_n)^{y_n}\cdot P(\omega_2|\boldsymbol{x}_n)^{1-y_n}=P(\omega_1|\boldsymbol{x}_n)$, if $\boldsymbol{x}_n \in \omega_1$.

2. $P(\omega_1|\boldsymbol{x}_n)^{y_n}\cdot P(\omega_2|\boldsymbol{x}_n)^{1-y_n}=P(\omega_2|\boldsymbol{x}_n)$, if $\boldsymbol{x}_n \in \omega_2$.

3. $y_n \ln P(\omega_1|\boldsymbol{x}_n)+(1-y_n)\ln P(\omega_2|\boldsymbol{x}_n)=P(\omega_1|\boldsymbol{x}_n)$, if $\boldsymbol{x}_n \in \omega_1$.

4. $y_n \ln P(\omega_1|\boldsymbol{x}_n)+(1-y_n)\ln P(\omega_2|\boldsymbol{x}_n)=P(\omega_2|\boldsymbol{x}_n)$, if $\boldsymbol{x}_n \in \omega_2$.

Exercise 15 (**mult. choice question):** Consider a two-class classification task, where the classes are denoted as ω_1 and ω_2. Assume that we have at our disposal a data set $X=\{(y_n,\boldsymbol{x}_n)\colon \boldsymbol{x}_n \in R^l,\ y_n \in \{0,1\}, n=1,\dots,N\}$, where $y_n=1$ (resp. 0) if $\boldsymbol{x}_n \in \omega_1$(resp. ω_2) (all $\boldsymbol{x}_n$'s are independent from each other). Recall that in the logistic regression framework, the class posterior probabilities of ω_1 and ω_2 with respect to a specific data vector $\boldsymbol{x}$**,** are $P(\omega_1|\boldsymbol{x})=\sigma(\boldsymbol{\theta}^T\boldsymbol{x})$ and $P(\omega_2|\boldsymbol{x})=1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x})$, where $\sigma(\cdot)$ is the logistic function and $\boldsymbol{\theta}$ the parameter vector that defines the classifier. The associated likelihood function is the following

$$P(\boldsymbol{y};\boldsymbol{\theta})=P(y_1,\dots,y_N;\boldsymbol{\theta})=\prod\prod_{n=1}^{N}\big(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_n)\big)^{y_n}\big(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_n)\big)^{1-y_n}$$

Where $\boldsymbol{y}$ comprises all y_n's. Which of the following statements is/are true?

1. The vector $\widehat{\boldsymbol{\theta}}$, where $P(\boldsymbol{y};\boldsymbol{\theta})$ is maximized defines the hyperplane (H): $\widehat{\boldsymbol{\theta}}^T\boldsymbol{x}=0$, that leaves as many as possible points from ω_1(resp. ω_2) on its negative (resp. positive) side.

2. If the groups of data points from the two classes are linearly separable and distant from each other, it is expected that the position $\widehat{\boldsymbol{\theta}}$ where $P(\boldsymbol{y};\boldsymbol{\theta})$ is maximized defines the hyperplane (H): $\widehat{\boldsymbol{\theta}}^T\boldsymbol{x}=0$, that leaves all points from ω_1(resp. ω_2) on its positive (resp. negative) side.

3. For each $\boldsymbol{x}_n$**,** both factors $\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_n)$ and $\big(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}_n)\big)$ contribute in a nontrivial way to $P(\boldsymbol{y};\boldsymbol{\theta})$.

4. It holds: $P(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{\boldsymbol{x}_n \in \omega_1} \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n) \prod_{\boldsymbol{x}_n \in \omega_2} \big(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n)\big)$.

Exercise 16 (**mult. choice question):** Consider a two-class classification task, where the classes are denoted as ω_1 and ω_2. Assume that we have at our disposal a data set $X = \{(y_n, \boldsymbol{x}_n): \boldsymbol{x}_n \in R^l,\ y_n \in \{0,1\}, n = 1, \dots, N\}$, where $y_n = 1$ (resp. 0) if $\boldsymbol{x}_n \in \omega_1$(resp. ω_2) (all $\boldsymbol{x}_n$'s are independent from each other). Recall that in the logistic regression framework, the class posterior probabilities of ω_1 and ω_2 with respect to a specific data vector $\boldsymbol{x}$**,** are $P(\omega_1|\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ and $P(\omega_2|\boldsymbol{x}) = 1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$, where $\sigma(\cdot)$ is the logistic function and $\boldsymbol{\theta}$ the parameter vector that defines the classifier. The associated negative log-likelihood function is the following

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \big(y_n \ln\big(\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n)\big) + (1 - y_n) \ln\big(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n)\big)\big)$$

Which of the following statements is/are true?

1. There exists a closed form solution for computing the $\widehat{\boldsymbol{\theta}}$ value of $\boldsymbol{\theta}$ that minimizes $L(\boldsymbol{\theta})$.

2. $L(\boldsymbol{\theta})$ exhibits several local minima.

3. The gradient descent algorithm for the above problem starts from an initial estimate $\boldsymbol{\theta}_0$ and iteratively updates it, by moving at each iteration towards the opposite direction of the gradient of $L(\boldsymbol{\theta})$ (computed on the previous estimate of $\boldsymbol{\theta}$), until a maximum of $L(\boldsymbol{\theta})$ is reached.

4. Once the optimum value for $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, is reached, a new data vector $\boldsymbol{x}$' is assigned to class ω_1(resp. ω_2) if $\widehat{\boldsymbol{\theta}}^T \boldsymbol{x}' >$ (resp. $<$) 0.

(C) Scatter Matrices

Exercise 17 (**mult. choice question**)**:** Which of the following statements, regarding a specific classification task, is/are correct?

1. Feature selection is always performed independently of the design of the associated classifier.

2. Feature generation may be performed in a combined way with the classifier design phase.

3. Although a set of "bad" features may have been selected, a smart classifier can always guarantee an acceptable performance of the overall classification system.

4. An information-rich feature is one whose values lie in the same interval for all the entities from all classes.

Exercise 18 (mult. choice question): Which of the following statements, regarding a specific classification task, is/are true?

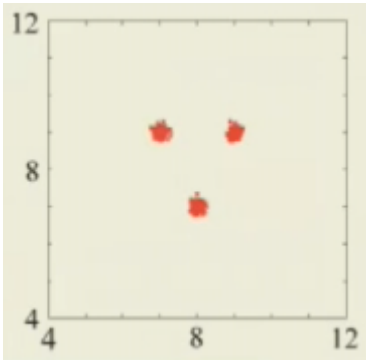
1. The dimension of the feature space is independent of the number of the selected features.

2. The patterns from different classes are represented in different feature spaces.

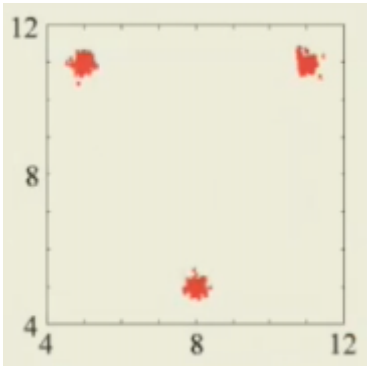
3. A desired characteristic of the selected features is to have large between-class distance.

4. A desired characteristic of the selected features is to have large within-class variance.

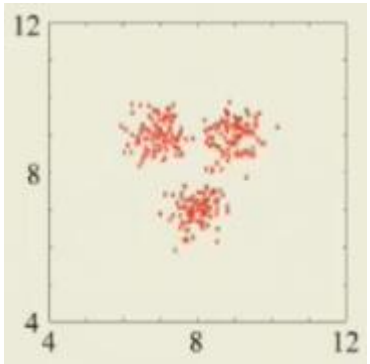
Exercise 19 (mult. choice question): Consider a specific three-class classification task where the involved patterns are represented by using three different combinations of pairs of features, as shown in the following figures.



(1)



(2)



(3)

Which of the following statements is/are true?

1. In figure 1, the classes exhibit “small” within-class variance and “small” between-class distances.

2. In figure 1, the classes exhibit “small” within-class variance and “large” between-class distances.

3. In figure 2, the classes exhibit “small” within-class variance and “large” between-class distances.

4. In figure 2, the classes exhibit “large” within-class variance and “small” between-class distances.

5. In figure 3, the classes exhibit “large” within-class variance and “large” between-class distances.

6. In figure 3, the classes exhibit “small” within-class variance and “large” between-class distances.

Exercise 20 (mult. choice question): Which of the following statements regarding a specific classification task is/are true?

1. The probability distribution of the points corresponding to the involved patterns is independent from the features that will be selected to represent these patterns.

2. The scatter matrices model the way the data points, from all classes, are distributed in the feature space.

3. The within-class scatter matrix quantifies the distance between the different classes.

4. The between-class scatter matrix quantifies the distance of the mean vectors of classes from the mean vector of the data set as a whole.

Exercise 21 (mult. choice question): Consider a specific M-class classification task and assume that l features have been selected for representing the involved patterns. Assume, also, that the data points from each class are aggregated around a single point and let $\boldsymbol{\mu}_m$, Σ_m and $P(\omega_m)$, be the mean vector, the covariance matrix and the a priori probability of the m-th class, $m = 1, \dots, M$. Also, Σ_w, Σ_b and Σ_{mixt} denote the within-class, the between-classes and the mixture scatter matrices, respectively and $\boldsymbol{\mu}_o$ the mean vector of the whole associated data set. Which of the following statements is/are true?

1. The scalar quantity Σ_w is defined as $\Sigma_w = \sum_{m=1}^{M} P(\omega_m)\Sigma_m$.

2. The quantity Σ_b is defined as $\Sigma_b = \sum_{m=1}^{M} P(\omega_m)(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)$.

3. It holds $\Sigma_{mixt} = \Sigma_w - \Sigma_b$.

4. For equiprobable classes (that is, with equal a priori probabilities), It holds $\boldsymbol{\mu}_o = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\mu}_m$.

Exercise 22 (mult. choice question): Consider a specific M-class classification task and assume that l features have been selected for representing the involved patterns. Assume, also, that the data points from each class are aggregated around a single point and let $\boldsymbol{\mu}_m$, Σ_m and $P(\omega_m)$, be the mean vector, the covariance matrix and the a priori probability of the m-th class, $m = 1, \dots, M$. Also, Σ_w and Σ_b denote the within-class and the between-classes scatter matrices, respectively, and $\boldsymbol{\mu}_o$ the mean vector of the whole associated data set. Which of the following statements is/are true?

1. The scatter matrices Σ_w and Σ_b are independent from the choice of the selected l-dimensional feature space.

2. The trace of the matrix $(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o) \cdot (\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)^T$ equals to $\sqrt{(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)}$.

3. The trace of Σ_b equals to the mean squared Euclidean distance of the mean vectors $\boldsymbol{\mu}_m$'s of the classes from $\boldsymbol{\mu}_o$.

4. If $P(\omega_1) \gg P(\omega_m)$, $m = 2, \dots, M$, the trace of Σ_b is approximately equal to $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_o)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_o)$

Exercise 23 (mult. choice question): Consider a specific M-class classification task and assume that l features have been selected for representing the involved patterns. Assume, also, that the data points from each class are aggregated around a single point and let $\boldsymbol{\mu}_m$, Σ_m and $P(\omega_m)$ be the mean vector, the covariance matrix and the a priori probability of the m-th class, $m = 1, \dots, M$. Also, Σ_w, Σ_b and Σ_{mixt} denote the within-class, the between-classes and the mixture scatter matrices, respectively and $\boldsymbol{\mu}_o$ the mean vector of the whole associated data set. Which of the following statements is/are true?

1. The trace of the covariance matrix Σ_m of the m-th class equals to the sum of the variances of all features for (the data points of) this class.

2. The trace of Σ_w equals to the average sum of the variances of all features over all classes.

3. It is $\frac{trace\{\Sigma_{mixt}\}}{trace\{\Sigma_w\}} = 1 - \frac{trace\{\Sigma_b\}}{trace\{\Sigma_w\}}$

4. It is $\frac{|\Sigma_{mixt}|}{|\Sigma_w|} \leq 1$

Hints: (a) For any two matrices A and B it is $trace\{A + B\} = trace\{A\} + trace\{B\}$.

(b) The covariance matrices are positive semidefinite.

(c) For two positive semidefinite matrices A and B, it is $|A + B| \geq |A| + |B|$.

Exercise 24 (mult. choice question): Consider a specific M-class classification task and assume that l features have been selected for representing the involved patterns. Assume, also, that the data points from each class are aggregated around a single point and let $\boldsymbol{\mu}_m$, Σ_m and $P(\omega_m)$ be the mean vector, the covariance matrix and the a priori probability of the m-th class, $m = 1, \dots, M$. Also, Σ_w, Σ_b and Σ_{mixt} denote the within-class, the between-classes and the mixture scatter matrices, respectively and $\boldsymbol{\mu}_o$ the mean vector of the whole associated data set. Two criteria that measure the "goodness" of the specific feature space are the $J_1 = \frac{trace\{\Sigma_{mixt}\}}{trace\{\Sigma_w\}}$

and the $J_2 = \frac{|\Sigma_{mixt}|}{|\Sigma_w|}$. Assuming that the data points from each class form a single cluster (aggregated around a specific point in the feature space), which of the following statements is/are true?

1. Large values of J_1 imply classes whose mean vectors $\boldsymbol{\mu}_m$ are distant from $\boldsymbol{\mu}_o$ and/or have small within-class variance.

2. Large distances $(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)^T(\boldsymbol{\mu}_m - \boldsymbol{\mu}_o)$, for $m = 1, \dots, M$, imply necessarily large distances $\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_j\right)^T\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_j\right)$, for $m, j = 1, \dots, M$ and $m \neq j$.

3. Small values of J_2 imply classes whose mean vectors $\boldsymbol{\mu}_m$ are distant from $\boldsymbol{\mu}_o$ and/or have small within-class variance.

4. If J_1 has a large value and it is known that the within-class variance of the classes is not very small, then their mean vectors $\boldsymbol{\mu}_m$ are very distant from $\boldsymbol{\mu}_o$.

Exercise 25 (mult. choice question): Consider a two-class classification task, where the two classes ω_1 and ω_2 are equiprobable. Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be the mean vectors of the two classes. Then, the associated between-class scatter matrix Σ_b can be expressed as

1. $\Sigma_b = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

2. $\Sigma_b = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$

3. $\Sigma_b = \frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$

4. $\Sigma_b = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$

Exercise 26 (mult. choice question): Consider a two-dimensional three-class classification task, where the involved classes ω_1, ω_2 and ω_3 are equiprobable. In the specific selected feature space, the data points from each class are aggregated around a specific point. The associated mean vectors are $\boldsymbol{\mu}_1 = [2,\ 2]^T, \boldsymbol{\mu}_2 = [-2,\ 0]^T$ and $\boldsymbol{\mu}_3 = [0, -2]^T$, while the associated covariance matrices are $\Sigma_1 = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.2 \end{bmatrix}$ and $\Sigma_3 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.4 \end{bmatrix}$. Then, the value of the criterion $J_1 = \frac{trace\{\Sigma_m\}}{trace\{\Sigma_w\}}$ is

1. 10.4

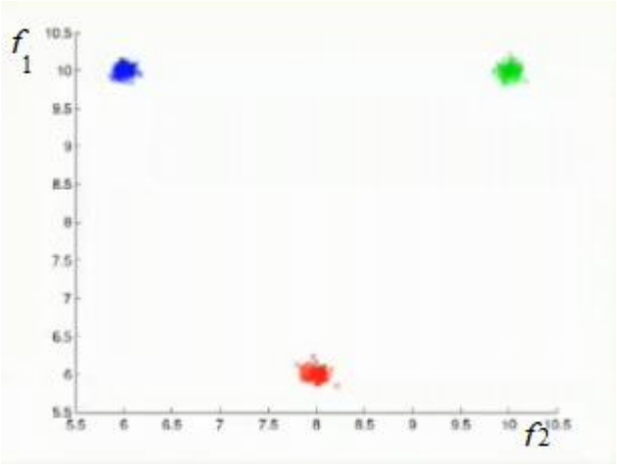
2. 104

3. 1.04

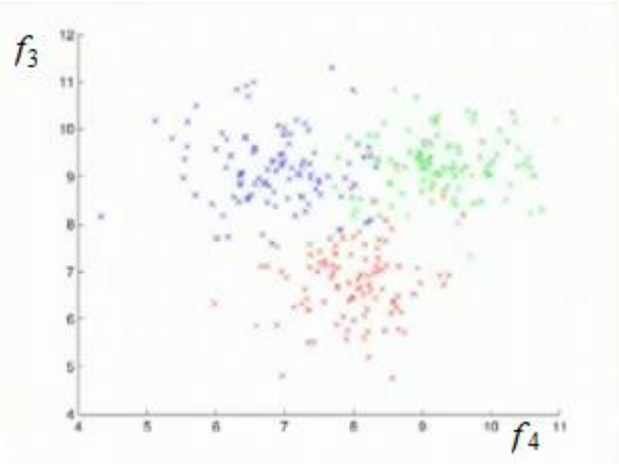
4. 5.7

Hint: Use the approximation $\frac{8}{3} \approx 2.6$.

Exercise 27 (mult. choice question): Consider a three-class classification task, where in figure (a) the available patterns are represented by the combination of the features f_1 and f_2, while in figure (b), the same available patterns are represented by the combination of the features f_3 and f_4. The values of the criterion $J_3 = trace\{{\Sigma_w}^{-1}\Sigma_b\}$ for (a) and (b) are 1350 and 4.5, respectively. Consider now the case where the patterns are represented by the combination of features f_2 and f_3. In this case the value of J_3 cannot be



(a)



(b)

1. 2.5

2. 1350

3. 1400

4. 500

(D) FDR criterion

Exercise 28 (mult. choice question): Which of the following statements is/are true?

1. All linear classifiers are defined irrespective of the data distribution in each class.

2. Two different linear classifiers may be optimal with respect to two different criteria (e.g., the Least Squares Error criterion and the negative log-likelihood in the framework of the Logistic criterion).

3. A linear classifier corresponds to a decision hyperplane in the feature space.

4. Consider the hyperplane (H): $g(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x} + \theta_0 = 0$. The relation $g(\boldsymbol{x}_1) > 0$ implies that the point $\boldsymbol{x}_1$ lies on the non-negative side of (H).

Exercise 29 (mult. choice question): Consider the vectors $\boldsymbol{x} = [1, 2]^T$ and $\boldsymbol{\theta} = [-1, 2]^T$. Then, the projection of $\boldsymbol{x}$ along the direction of $\boldsymbol{\theta}$ is

1. $[-0.6, 1.2]^T$

2. $\frac{3}{\sqrt{5}}$

3. $\left[-\frac{3}{\sqrt{5}}, \frac{6}{\sqrt{5}}\right]^T$

4. $[1.2, -0.6]^T$

Exercise 30 (mult. choice question): Consider a two-class classification task, where the class labels are $+1$ and -1 and let X be a specific data set. Also, let (H): $g(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x} + \theta_0 = 0$ be a hyperplane that defines a linear classifier. In the Fisher's Linear Discriminant rationale, one seeks for that $\boldsymbol{\theta}$ so that:

1. (H) leaves all points of X from class $+1$ (resp. -1) on its positive (resp. negative) side, even for nonlinearly separable classes.

2. (H) leaves the maximum possible margin from each class.

3. the projections of the points of X from class $+1$ along the direction of $\boldsymbol{\theta}$ are as far away as possible from the projections of the points of X from class -1.

4. the projections of the points of X from classes $+1$ and -1 along the direction of $\boldsymbol{\theta}$ exhibit as much as high variance around the respective class means.

Exercise 31 (mult. choice question): Consider a two-class classification task, where the class labels are 1 and 2 and let $X = \{(y_n, \boldsymbol{x}_n),\ \boldsymbol{x}_n \in R^l,\ y_n \in \{1, 2\}, n = 1, \dots, N\}$ be a specific data set. Let $\boldsymbol{\mu}_i$ be the mean of the vectors of the i-th class and Σ_i the covariance matrix of the i-th class, $i = 1,2$. The projection of $\boldsymbol{\mu}_i$ along the direction defined by a vector $\boldsymbol{\theta}$, is $\mu_i = \boldsymbol{\theta}^T\boldsymbol{\mu}_i$, $i = 1,2$, while the variance of the projections of the points of the i-th class along the direction of $\boldsymbol{\theta}$ around μ_i is ${\sigma_i}^2 = \boldsymbol{\theta}^T\Sigma_i\boldsymbol{\theta}$. Which of the following expressions are valid for the Fisher's linear discriminant ratio (FDR) criterion?

1. $FDR = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2}$

2. $FDR = \frac{\mu_1 - \mu_2}{{\sigma_1}^2 + {\sigma_2}^2}$

3. $FDR = \frac{(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T (\boldsymbol{\theta}\boldsymbol{\theta}^T)(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})}{\boldsymbol{\theta}^T (\Sigma_1 + \Sigma_2)\boldsymbol{\theta}}$

4. $FDR = \frac{(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T (\boldsymbol{\theta}^T\boldsymbol{\theta})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})}{\boldsymbol{\theta}^T (\Sigma_1 + \Sigma_2)\boldsymbol{\theta}}$

Exercise 32 (mult. choice question): Consider a two-class classification task, where the class labels are 1 and 2 and let $X = \{(y_n, \boldsymbol{x}_n),\ \ \boldsymbol{x}_n \in R^l,\ \ y_n \in \{1,2\}, n = 1, \dots, N\}$ be a specific data set. Let $\boldsymbol{\mu}_i$ be the mean of the vectors of the i-th class and Σ_i the covariance matrix of the i-th class, $i = 1,2$. Also, Σ_w and Σ_b stand for the within-class and the between-classes scatter matrices. Which of the following choices of $\boldsymbol{\theta}$ minimize the Fisher's linear discriminant ratio (FDR) criterion?

1. ${\Sigma_w}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

2. ${\Sigma_b}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

3. ${\Sigma_w}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$

4. ${\Sigma_b}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$

Exercise 33 (mult. choice question): Consider a two-class classification task, where the involved classes are equiprobable and their class labels are 1 and 2. Let $X = \{(y_n, \boldsymbol{x}_n),\ \ \boldsymbol{x}_n \in R^l,\ \ y_n \in \{1,2\}, n = 1, \dots, N\}$ be a specific data set and $\boldsymbol{\mu}_i$ be the mean of the vectors of the i-th class and Σ_i the covariance matrix of the i-th class, $i = 1,2$. Also, Σ_w and Σ_b stand for the within-class and the between-classes scatter matrices. Which of the following statements regarding the Fisher's Linear Discriminant (FDR) ratio criterion is/are true?

1. If both Σ_1 and Σ_2 are diagonal, the direction $\boldsymbol{\theta}$ that maximizes FDR is $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

2. If $\Sigma_w = \Sigma_b$, then $FDR = 1$.

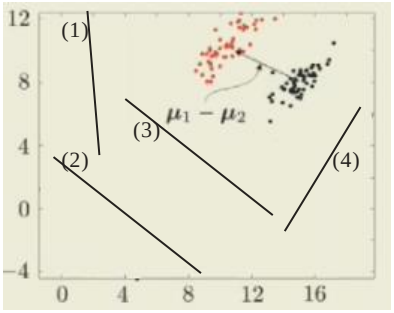
3. If $\Sigma_w = a \cdot I$ and $\Sigma_b = b \cdot I$, then $FDR = \frac{b}{a}$

4. If $\Sigma_b = b \cdot I$, the direction $\boldsymbol{\theta}$ that maximizes FDR is always equal to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Exercise 34 (mult. choice question): Consider a two-class classification problem, where the involved classes are equiprobable and their class labels are 1 and 2. Let $X = \{(y_n, \boldsymbol{x}_n),\ \boldsymbol{x}_n \in R^l,\ y_n \in \{1,2\}, n = 1, \dots, N\}$ be a specific data set. Let $\boldsymbol{\mu}_1 = [1,1]^T$ and $\boldsymbol{\mu}_2 = [2,2]^T$ be the mean of the vectors of classes 1 and 2, respectively, and and $\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$ be the covariance matrices of the two classes, respectively. Then, the direction $\boldsymbol{\theta}$ that maximizes the Fisher's Linear Discriminant (FDR) ratio criterion is:

1. $\left[-\frac{1}{3}, -\frac{1}{3}\right]^T$
2. $\left[\frac{1}{3}, \frac{1}{3}\right]^T$
3. $\left[-\frac{1}{3}, \frac{1}{3}\right]^T$
4. $[-1,-1]^T$

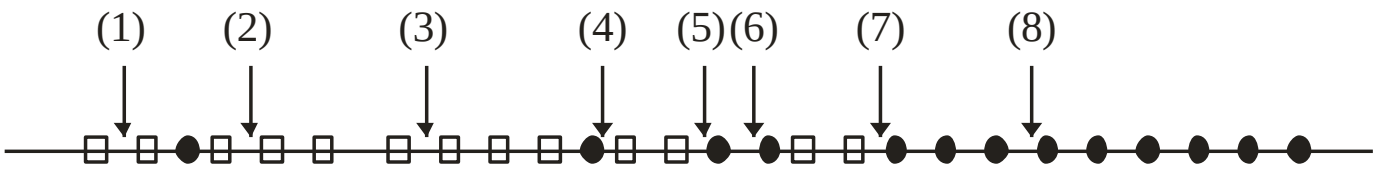
Exercise 35 (mult. choice question): Consider the two-class classification problem shown in the figure below, where the two classes (the red one and the black one) are equiprobable. Let also, FDR_1, FDR_2, FDR_3 and FDR_4 are the values of the Fisher's Linear Discriminant (FDR) ratio associated with the respective directions shown in the figure (the directions (2) and (3) are parallel to each other). Which of the following statements is/are true?



1. $FDR_1 < FDR_3$
2. $FDR_2 = FDR_3$
3. $FDR_4 > FDR_2$
4. $FDR_4 > FDR_3$

Exercise 36 (mult. choice question): Consider a two-class classification problem where the two equiprobable classes are denoted by 1 and 2. In order to utilize the FDR as a classifier, we need to determine θ_0 in the decision rule $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T {\Sigma_w}^{-1} \boldsymbol{x} + \theta_0 \begin{cases} > 0, & class\ 1 \\ < 0, & class\ 2 \end{cases}$. In the following figure, the projections of the points of the data set at hand from the two classes along

the optimal (according to FDR) direction $\boldsymbol{\theta}$ are shown (the coordinate of the projection of $\boldsymbol{x}$ along the optimal direction $\boldsymbol{\theta}$, is $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T {\Sigma_w}^{-1} \boldsymbol{x}$).



The value(s) of θ_0 that minimize the misclassification error on the projections of the data points is/are the one/those that correspond to the position(s)

1. (1)
2. (2)
3. (3)
4. (4)
5. (5)
6. (6)
7. (7)
8. (8)

(E) Decision trees

Exercise 37 (mult. choice question): Which of the following statements regarding classification trees is/are true?

1. Classification trees are multistage systems.

2. The classification of a pattern to a class is performed in one step.

3. The class where a specific pattern will be assigned is determined by sequentially rejecting classes where the pattern will not be assigned.

4. A class is rejected after the application of a ternary “yes”, “no”, “undefined” test.

Exercise 38 (mult. choice question): Which of the following statements regarding ordinary binary classification trees (OBCT) is/are true?

1. Classification trees are non-linear classifiers.

2. As is the case with the Bayes classifier, for classes that are modeled by normal distributions, classification trees partition the space via second degree curves.

3. Ordinary binary classification trees can also be used for regression tasks.

4. Ordinary classification trees partition the feature space via hyperplanes so that each one of them to be perpendicular to one of the axes that define the feature space.

Exercise 39 (mult. choice question): Which of the following statements regarding ordinary binary classification trees (OBCT) is/are true?

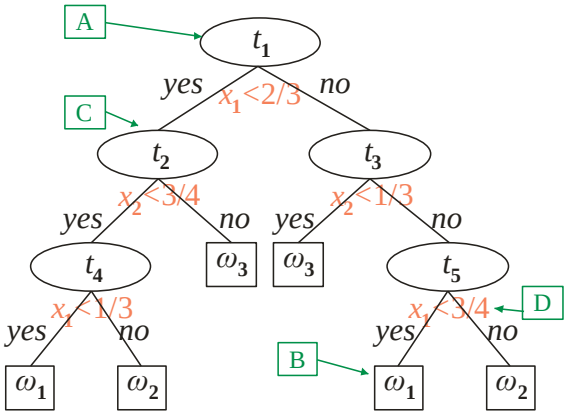
1. The equation of a hyperplane defined by an OBCT is of the form $x_i{}^2 - a = 0$, where x_i is the i-th feature of the associated feature space.

2. The decision region for a class, as defined by a classification tree, is a union of hyper-rectangles.

3. The knowledge of the probability density functions of each class is of vital importance in OBCTs.

4. In an OBCT with several non-leaf nodes, the class decision region to which a specific pattern belongs is always determined exclusively by the application of just a single test of the type $x_i < a$, where x_i is the i-th feature of the associated feature space.

Exercise 40 (mult. choice question): Consider the ordinary binary classification tree (OBCT) shown in the figure on the right. Which of the following statements is/are true?



1. Element A is a leaf node.

2. Element A is a root node.

3. Element B is an intermediate node.

4. Element B is a leaf node.

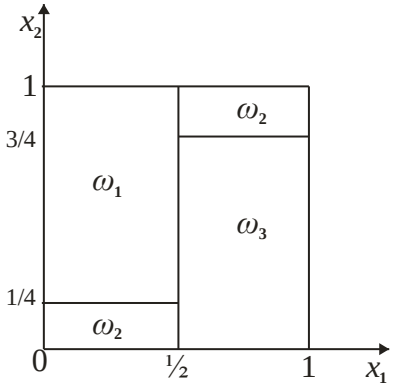
5. Element C is a splitting criterion

6. Element C is an intermediate node.

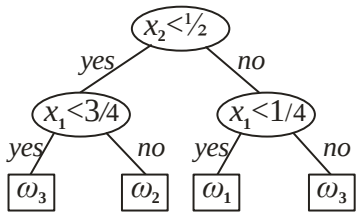
7. Element D is a splitting criterion

8. Element D is a leaf node.

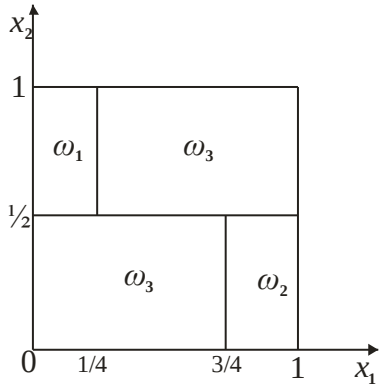
Exercise 41 (mult. choice question): Consider a three-class two-dimensional classification task, where the classes are denoted as ω_1, ω_2 and ω_3 and the data points are bounded in the $[0,1] \times [0,1]$ unit square. In the left column below, three different partitions of feature space are shown, as performed by three ordinary binary classification trees (OBCTs) (right column).



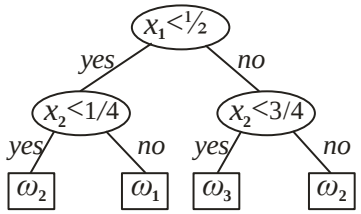
(1)



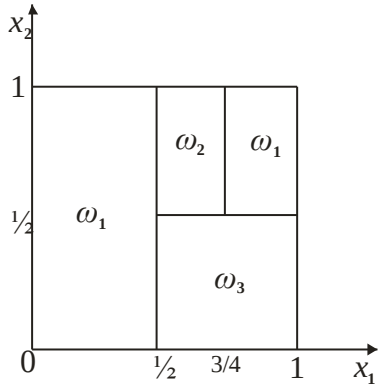
(a)

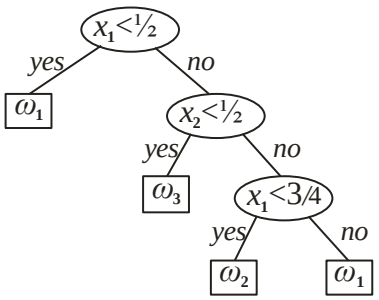


(2)



(b)





(3) (c)

Which of the following pairs "partition ⟷ OBCT" are compatible (in the sense that the OBCT produces the partition) to each other?

1. Partition (1) ⟷ OBCT (b)

2. Partition (1) ⟷ OBCT (a)

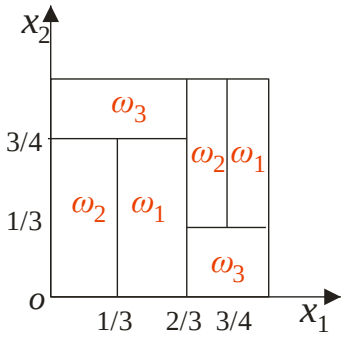
3. Partition (2) ⟷ OBCT (a)

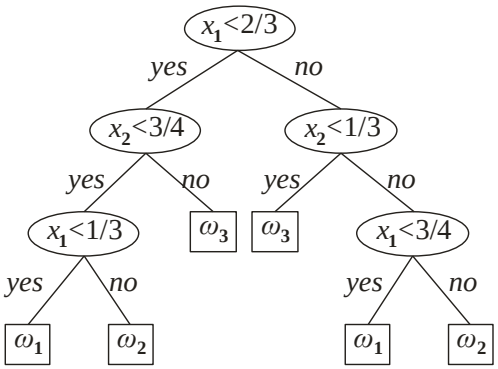
4. Partition (2) ⟷ OBCT (c)

5. Partition (3) ⟷ OBCT (c)

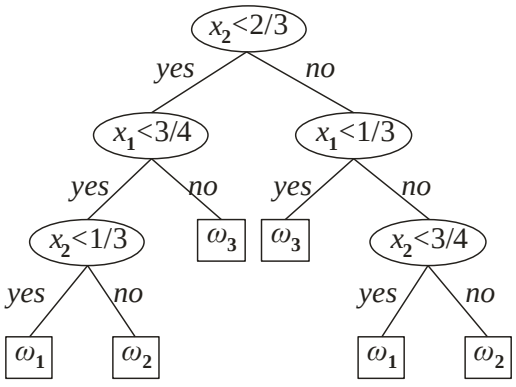
6. Partition (3) ⟷ OBCT (b)

Exercise 42 (mult. choice question): Consider a three-class two-dimensional classification task, where the classes are denoted as ω_1, ω_2 and ω_3 and the data points are bounded in the $[0,1] \times [0,1]$ unit square. The figure on the right shows a partition of the feature space performed by an ordinary binary classification tree (OBCT). Which of the following is the OBCT that performs this partition?

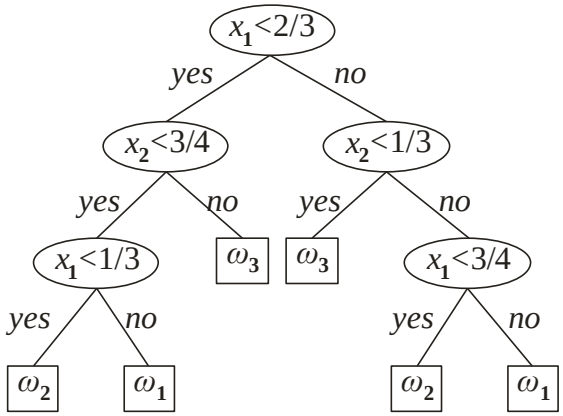




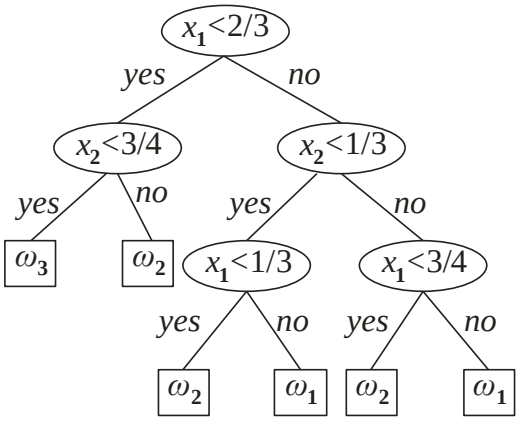
(1)



(2)

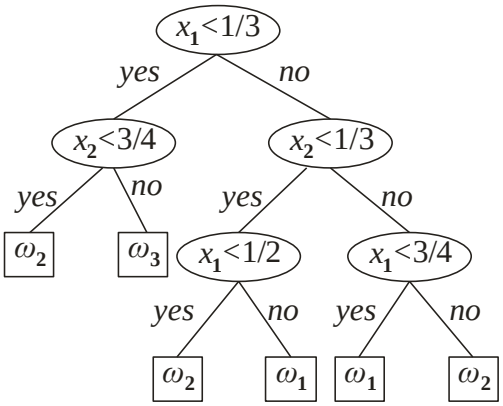


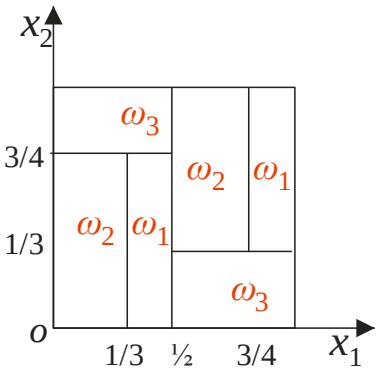
(3)

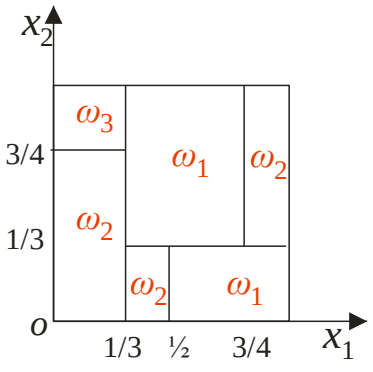


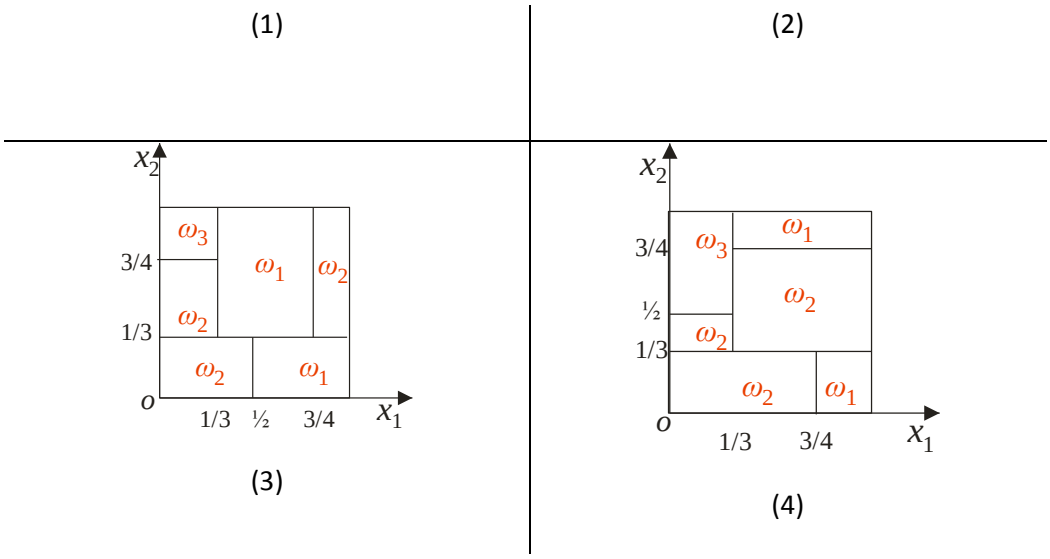
(4)

Exercise 43 (mult. choice question): Consider a three-class two-dimensional classification task, where the classes are denoted by ω_1, ω_2 and ω_3 and the data points are bounded in the $[0,1] \times [0,1]$ unit square. Consider the ordinary binary classification tree (OBCT) shown on the right. Which of the following partitions of the feature space is the one performed by this OBCT?









Exercise 44 (mult. choice question): Consider an M-class classification task, for which a specific data set X of size N is available. Which of the following statements regarding ordinary binary classification trees (OBCTs) are true?

1. The construction of an OBCT takes explicitly into account the probability density functions that model the M classes.

2. The splitting criteria associated with a specific OBCT are as many as its non-leaf nodes.

3. Each leaf node is associated with a single class.

4. Usually, the total number of nodes in an OBCT exceeds N.

Exercise 45 (mult. choice question): Consider a two-class classification task, for which a data set X consisting of twelve points is available. The first six of them, i.e., $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_6$, are from class ω_1, while the next six, i.e., $\boldsymbol{x}_7, \ldots, \boldsymbol{x}_{12}$, are from class ω_2. The aim is to construct an ordinary binary classification tree (OBCT) to solve this problem. The splitting criterion that is to be adopted at the root node splits X into two subsets X_Y and X_N. Which of the following splits is/are valid?

1. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_9, \boldsymbol{x}_{11}\}, X_N = \{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_6, \boldsymbol{x}_8, \boldsymbol{x}_{10}, \boldsymbol{x}_{12}\}$

2. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5, \boldsymbol{x}_7, \boldsymbol{x}_9, \boldsymbol{x}_{11}\}, X_N = \{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_8, \boldsymbol{x}_{10}, \boldsymbol{x}_{12}\}$

3. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_7, \boldsymbol{x}_9, \boldsymbol{x}_{11}\}, X_N = \{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_6, \boldsymbol{x}_8, \boldsymbol{x}_{10}, \boldsymbol{x}_{12}\}$

4. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_7\}, X_N = \{\boldsymbol{x}_6, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}\}$

Exercise 46 (mult. choice question): Consider a three-class two-dimensional classification task, where the classes are denoted as ω_1, ω_2 and ω_3, and for which a data set X that consists of the following six points is available: $\boldsymbol{x}_1 = [1,14]^T, \boldsymbol{x}_2 = [3,12]^T, \boldsymbol{x}_3 = [5,10]^T, \boldsymbol{x}_4 = [6,9]^T, \boldsymbol{x}_5 = [2,13]^T, \boldsymbol{x}_6 = [4,11]^T$. From these points, $\boldsymbol{x}_1, \boldsymbol{x}_2$ are from class ω_1, $\boldsymbol{x}_3, \boldsymbol{x}_4$ are from class ω_2 and $\boldsymbol{x}_5, \boldsymbol{x}_6$ are from class ω_3. Assume that the aim is to construct an ordinary binary classification tree (OBCT) to solve the above task. If we adopt for the root node the splitting criterion $x_1 \leq 3$, what will be the two associated subsets X_Y and X_N to which X is split (x_1 denotes the first feature and it should not be confused with the data vector $\boldsymbol{x}_1$)?

1. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5\}$, $X_N = \{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_6\}$

2. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_5\}$, $X_N = \{\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_6\}$

3. $X_Y = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_5\}$, $X_N = \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_6\}$

4. $X_Y = \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_6\}$, $X_N = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_5\}$

Exercise 47 (mult. choice question): Consider a six-class classification task, where the classes are denoted as $\omega_1, \ldots, \omega_6$. Assume that we construct an ordinary binary classification tree (OBCT) to deal with this task and that we are about to choose the splitting criterion for a specific node t. The associated data set, X_t, consists of eighteen points which belong to only three of the total number of the six classes, so that the first six of them, i.e., $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_6$, are from class ω_1, the next six, i.e., $\boldsymbol{x}_7, \ldots, \boldsymbol{x}_{12}$, are from class ω_2 and the last six, i.e., $\boldsymbol{x}_{13}, \ldots, \boldsymbol{x}_{18}$, are from class ω_3. Assume that, for four candidate splitting criteria, $\mathrm{SC}_1, \mathrm{SC}_2, \mathrm{SC}_3, \mathrm{SC}_4$, the associated subsets X_{tY} and X_{tN} are those shown below.

SC_1: $X_{tY} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{13}, \boldsymbol{x}_{14}, \boldsymbol{x}_{15}\}$, $X_{tN} = \{\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{16}, \boldsymbol{x}_{17}, \boldsymbol{x}_{18}\}$

SC_2: $X_{tY} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9\}$, $X_{tN} = \{\boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}, \boldsymbol{x}_{14}, \boldsymbol{x}_{15}, \boldsymbol{x}_{16}, \boldsymbol{x}_{17}, \boldsymbol{x}_{18}\}$

SC_3: $X_{tY} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_{13}, \boldsymbol{x}_{14}\}$, $X_{tN} = \{\boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_9, \boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{15}, \boldsymbol{x}_{16}, \boldsymbol{x}_{17}, \boldsymbol{x}_{18}\}$

SC_4: $X_{tY} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}$, $X_{tN} = \{\boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}, \boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}, \boldsymbol{x}_{14}, \boldsymbol{x}_{15}, \boldsymbol{x}_{16}, \boldsymbol{x}_{17}, \boldsymbol{x}_{18}\}$

Which of these criteria lead to subsets of X_t that are more class-homogeneous than X_t?

1. SC_1

2. SC_2

3. SC_3

4. SC_4

Exercise 48:

Consider the set-up of the example given in slide 40 of the 8th lecture. Compute the node impurity decrease achieved by the query "$x_2 \leq 8$".

Exercise 49:

Suppose you are given a data set $Y = \{(y_i, \boldsymbol{x}_i'), i = 1, \dots, N\}$ where $y_i \in \{0,1\}$ is the class label for vector $\boldsymbol{x}_i' \in R^l$. Assume that y and $\boldsymbol{x}'$ are related via the following model: $y = f(\boldsymbol{\theta}^T \boldsymbol{x}' + \theta_0)$, where $\boldsymbol{\theta}$ and θ_0 are the model parameters and $f(z) = 1/(1 + exp(-az))$.

(a) **Plot** the function $f(z)$ for various values of the parameter a.
(b) Propose a gradient descent scheme to **train** this model (that is, to estimate the values of the involved parameters), based on the **minimization** of the sum of error squares criterion, using Y.
(c) Can the model ever respond with a **"clear"** 1 or a **"clear"** 0, for a given $\boldsymbol{x}$?
(d) How can we interpret the response of the model for a given $\boldsymbol{x}$?
(e) Propose a way for leading the model responses very close to 1 (for class 1 vectors) or 0 (for class 0 vectors).

Hints:

(a) Use a more compact notation by setting $\boldsymbol{x}_i = [1\ \boldsymbol{x}_i']^T$, $i = 1, \dots, N$, and $\boldsymbol{\theta} = [\theta_0\ \boldsymbol{\theta}]^T$. The model then becomes $y = f(\boldsymbol{\theta}^T \boldsymbol{x})$.
(b) The sum of error squares criterion in this case is $J(\boldsymbol{\theta}) = \sum_{n=1}^{N} \big(y_n - f(\boldsymbol{\theta}^T \boldsymbol{x_n})\big)^2$ (Consult also the relative pdf file about optimization theory basics, uploaded in e-class).

It is $f'(z) = \frac{df(z)}{dz} = af(z)\big(1 - f(z)\big)$.

Exercise 50 (python code + text):

Consider a two-class, two-dimensional classification problem for which you can find attached two sets: one for training and one for testing (file *HW8.mat*). Each of these sets consists of pairs of the form $(y_i, \boldsymbol{x}_i)$, where y_i is the class label for vector $\boldsymbol{x}_i$. Let N_{train} and N_{test} denote the number of training and test sets, respectively. The data are given via the following arrays/matrices:

- ***train_x*** (a $N_{train} \times 2$ **matrix** that contains in its rows the training vectors $\boldsymbol{x}_i$)
- ***train_y*** (a N_{train}–dim. column **vector** containing the **class labels** (1 or 2) of the corresponding training vectors $\boldsymbol{x}_i$ included in ***train_x***).
- ***test_x*** (a $N_{test} \times 2$ **matrix** that contains in its rows the test vectors $\boldsymbol{x}_i$)
- ***test_y*** (a N_{test}–dim. column **vector** containing the **class labels** (1 or 2) of the corresponding test vectors $\boldsymbol{x}_i$ included in ***test_x***).

Assume that the two classes, ω_1 and ω_2 are modeled by normal distributions.

(a) Adopt the **k-nearest neighbor classifier**, for $k = 5$ and estimate the classification error probability.

(b) Depict graphically the training set, using different colors for points from different classes.

(c) Report the classification results obtained by the k-NN classifier and compare them with the results obtained by the Bayes and the naïve Bayes classifier (see relevant exercises in HW7 and HW7a).

Hint: Use the attached Python code in file *HW8.ipynb* (also given in Homework 7).