

CS4220  
Machine Learning

Written examination

# Answer sheet 1

Name :  
Student number :

## 1 Statements

*(10 points)*

Circle the correct statement, i.e. TRUE or FALSE. If a statement does not hold in general, but only under certain conditions that are not mentioned, then the statement should be marked as FALSE. 10 correct answers will give you 0 points; each additional correct answer gives you 1 point.

### Classification

1. To obtain a classifier with a high classification performance, you need to estimate the class-conditional probabilities well.  
TRUE                      FALSE
2. For a given dataset, the Bayes' classifier has the lowest classification error.  
TRUE                      FALSE
3. The linear discriminant analysis `ldc` is scale insensitive.  
TRUE                      FALSE
4. A confusion matrix shows how expensive it is to misclassify an object from one class to another class.  
TRUE                      FALSE
5. The logistic classifier minimizes the number of erroneously classified objects of the training set.  
TRUE                      FALSE
6. The decision boundary of a two-class classifier is defined as all points  $\mathbf{x}$  for which holds:  $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$ .  
TRUE                      FALSE
7. The Receiver Operating Characteristic Curve is independent of class priors in the test set.  
TRUE                      FALSE
8. When the classes in a classification problem have a Gaussian distribution, the optimal classifier is the quadratic classifier `qdc`.  
TRUE                      FALSE
9. When  $k$  in the  $k$ -nearest neighbor classifier is set to the number of training examples,  $k = n$ , the classifier becomes independent of class distributions.  
TRUE                      FALSE
10. Increasing the number of folds in crossvalidation reduces the variance in the error estimate.  
TRUE                      FALSE

# Answer sheet 2

Name :  
Student number :

## Clustering

1. Hierarchical clustering makes no assumptions about the underlying distribution of the data.  
TRUE FALSE
2. A hierarchical clustering result is sensitive to its random initialization.  
TRUE FALSE
3. The larger  $K$  is in  $K$ -means clustering the better the model typically fits to the training data.  
TRUE FALSE
4. If single linkage clustering fails to work on a particular data set, complete linkage will work (and vice versa).  
TRUE FALSE
5. To avoid local minima, the  $K$ -means algorithm should always be started with the  $K$  means centered on randomly selected objects.  
TRUE FALSE

## Feature Extraction and Selection

1. Reducing the feature vector dimensionality by means of principle component analysis (PCA) will lead to worse classification performance because it is an unsupervised technique.  
TRUE FALSE
2. Extracting features can never lead to a decrease of the Bayes error.  
TRUE FALSE
3. In feature selection, the number of different feature sets of size 7 that one can select from 1,000 dimensions is larger than the number of different feature sets of size 993 that one can select.  
TRUE FALSE
4. PCA assumes that the underlying data distribution is Gaussian.  
TRUE FALSE
5. Even by extracting more than 5 (linear or nonlinear) features from a 5-dimensional classification problem, one cannot reduce the Bayes error of the original problem.  
TRUE FALSE

## 2 Classification: probabilities

(10 points)

A classifier obtained the following class-conditional probabilities for ten test objects:

object nr	Class conditional probabilities		true class
	class 1	class 2	
1	0.07	0.02	1
2	0.05	0.04	1
3	0.1	0.2	1
4	0.3	0.35	1
5	0.8	0.07	1
6	0.02	0.01	1
7	0.2	0.5	2
8	0.4	0.6	2
9	0.03	0.8	2
10	0.01	0.2	2

- Assume equal class priors. How many objects are misclassified? (1 points)
- Assume equal class priors. What is the classification error of this test set? (2 points)
- Assume that class 1 is ten times as likely as class 2,  $p(\omega_1) = 10p(\omega_2)$ . What is the classification error now? (3 points)
- Assume still the case that  $p(\omega_1) = 10p(\omega_2)$ . Will the rejection of the most uncertain object improve the performance of the classifier? (Assume that the cost of rejection is lower than the cost of misclassification) (2 points)
- Another classifier obtained the following class-conditional probabilities for the ten objects:

object nr	Class conditional probabilities		true class
	class 1	class 2	
1	0	0	1
2	0	0	1
3	0	0	1
4	0.0005	0	1
5	0	0	1
6	0	0	1
7	0	0	2
8	0	0	2
9	0	0	2
10	0	0	2

Explain what happened here.

(2 points)

### 3 Classification: curves

(10 points)

- a. Assume we train a Parzen classifier on some training data, and we vary the width parameter  $h$ . First make a plot of the likelihood of the training data as a function of the width parameter  $h$ . Second, explain how you can optimize the width parameter  $h$ . (1 points)
- b. Draw a learning curve. Show both the true error as the apparent error. Name the axes, and list the important characteristics of the curve. Mention also the words "Bayes' error" and "overfitting". (2 points)
- c. Assume we have a training set consisting of  $n$  objects and  $d$  features with 2 classes. How many parameters have to be estimated for the quadratic classifier? And how many parameters have to be estimated for the linear support vector classifier? And how many for the Parzen classifier? (2 points)
- d. Assume we have a two-class classification problem, with a fixed number of  $n$  training points, but with a very large set of weakly informative features. Assume we randomly select features from the large feature set. Make a plot of the classification error as a function of the number of features, for the quadratic classifier, the linear support vector classifier and the Parzen classifier. Draw them all in one plot, and explain the similarities and differences between the curves. (3 points)
- e. Assume we are in the same situation as in previous question d. and we decide to combine the three classifiers using *majority voting*. Again plot the error as a function of the number of features, for all classifiers and the combined classifier. (2 points)

## 4 Feature Extraction and Selection

(11 points)

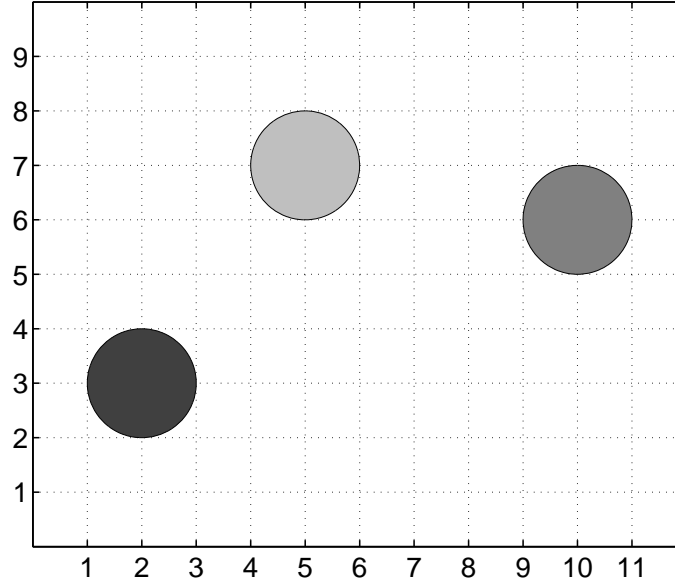


Figure 3: Example configuration of three classes.

Let us consider a two-dimensional, three-class problem in which all classes are uniformly distributed in a disc with radius 1. All class means are variable but differ so that there is no class overlap in the 2D feature space. The prior probabilities of all classes are assumed equal. Figure 3 above gives an example configuration in which the three class means are given by  $(2, 3)$ ,  $(5, 7)$ , and  $(10, 6)$ , respectively.

- Assume we reduce the feature dimensionality from 2 to 1 by means of the Fisher mapping (**fisherm**). Configure the three class means such that in the 2D space there is *no overlap* but in the 1D space found by the Fisher mapping two of the three classes *completely* overlap. You are only allowed to put the class means in the *grid points* given in the figure. Provide the three coordinates of your solution. (3 points)
- Is it possible to give three means for which there is no overlap in 2D but for which all three classes overlap completely in 1D when using the Fisher mapping? (2 points)

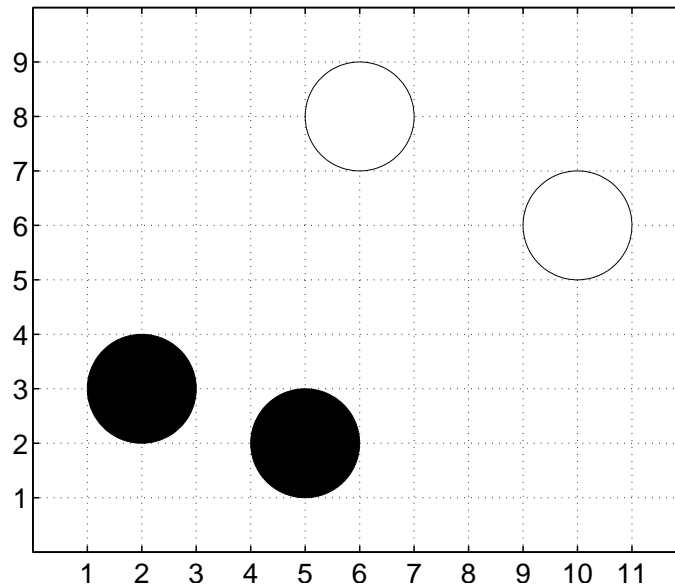


Figure 4: Example configuration of two classes (both of which in turn consist of two clusters).

We now consider a similar setting as exemplified in Figure 3, but now we have *two* classes (one black one white) both consisting of two clusters. All four clusters are uniformly distributed in a disc with radius 1. (The four clusters have equal priors.) An example configuration is given in Figure 4.

- c. Configure the four class means such that if one would look at either of the two features, both classes would perfectly *overlap*, while in the original 2D space the two classes are perfectly non-overlapping. In other words, construct an example for which selecting a single feature would give a Bayes error of 0.5, while the Bayes error in the original space is 0.

Again, you are only allowed to put the means in the *grid points* given in the figure. Provide the four coordinates of the clusters of your solution. Make sure you make clear where the black clusters go and where the white go. (2 points)

- d. Would a linear feature extraction technique be able to provide a better one-dimensional subspace for the problem created in a. than feature selection is able to do? That is, can feature extraction for a 1D feature for which the two classes do not fully overlap? (2 points)
- e. Is it possible to create a classification problem, again positioning the four clusters, such that feature selection will outperform any kind of feature extraction? Explain your answer. (2 points)

## 5 Clustering

(9 points)



Figure 5: A scatter plot of three hundred feature vectors in a two-dimensional feature space.

Figure 5 above displays three hundred points in 2D drawn from three normal distributions. We want to cluster the points in *two* clusters by means of  $K$ -means clustering with  $K = 2$ , i.e., we want to divide the data set up in two groups.

- If we initialize  $K$ -means clustering with the two crosses indicated in the same figure, how will the final clustering look like (indicate this clustering in figure 6 or describe it in some other unambiguous way)? (2 points)

Now consider all possible random initializations for  $K$ -means clustering (with  $K$  still equal to 2) for the problem in the figure.

- How many different clusterings consisting of *two nonempty* clusters can be found when randomly initializing the two initial cluster centers? (3 points)
- Will every random initialization, with two initial means, result in a solution with two clusters? Explain your answer. (2 points)
- If we choose to take four cluster centers ( $K = 4$ ), how can one detect that this is probably a wrong number of clusters? What measure or test reveals this? (2 points)



# Answer sheet 3

Name :  
Student number :

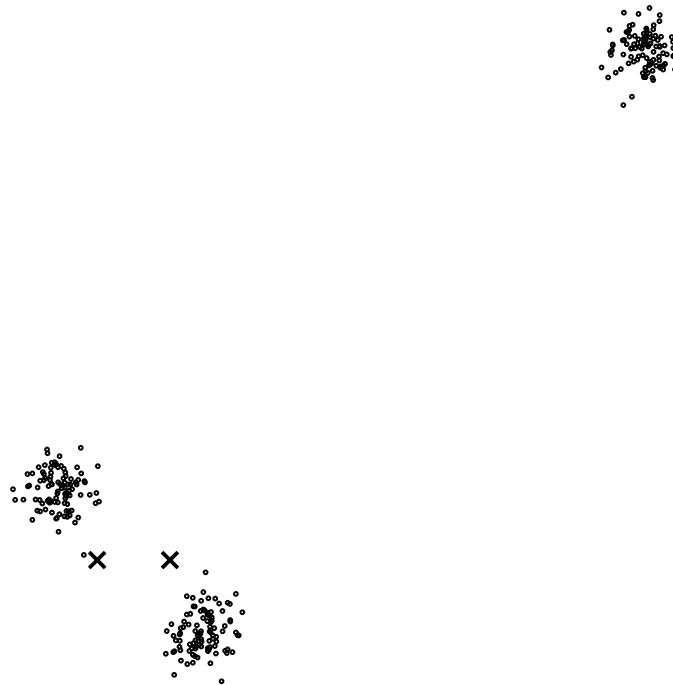


Figure 6: A scatter plot of three hundred feature vectors in a two-dimensional feature space.