Tentamen 3 Februari 2017, antwoorden

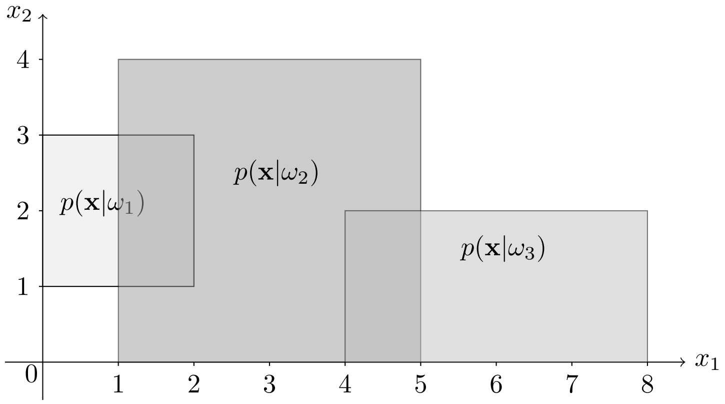
Pattern Recognition (Technische Universiteit Delft)

IN4085
Pattern Recognition

Written examination
3-02-2017, 9:00-12:00

- **There are 5 questions**
- **You have 45 minutes to answer the first question (Answer sheets 1 and 2), during which you cannot consult the book (or something else). Circle the right answers, and hand in this answer sheet.**
- **After you have handed in the Answer sheets, you are allowed to consult the book by Theodoridis and Koutroumbas to answer the following questions in the remaining time.**
- **Answer EACH QUESTION on a SEPARATE SHEET OF PAPER.**
- **As much as possible, include the CALCULATIONS you made to get to an answer.**
- **Do not forget to put your NAME and STUDENT NUMBER on top of every sheet**

2 Classifiers *(10 points)*



Let us consider the above two-dimensional, three-class classification problem, with two class-conditional distributions $p(\mathbf{x}|\omega_1)$, $p(\mathbf{x}|\omega_2)$ and $p(\mathbf{x}|\omega_3)$. Class ω_1 has a uniform distribution for $0 < x_1 < 2$, $1 < x_2 < 3$, class ω_2 has a uniform distribution for $1 < x_1 < 5$, $0 < x_2 < 4$ and class ω_3 has a uniform distribution for $4 < x_1 < 8$, $0 < x_2 < 2$.

Assume that all classes are equally likely.

a. Determine the decision boundary of the Bayes classifier. *(2 points)*
Solution: First we need to determine the heights of the densities. For $p(\mathbf{x}|\omega_1)$, the height

is 1/4, for $p(\mathbf{x}|\omega_2)$, the height is 1/16, for $p(\mathbf{x}|\omega_3)$, the height is 1/8. Therefore the whole support of ω_1 is assigned to ω_1, and the whole support of ω_3 is assigned to ω_3. The decision boundary is $(1,1)-(2,1)-(2,3)-(1,3)$ and $(4,0)-(4,2)-(5,2)$.

b. What are the posterior probabilities for $\mathbf{x} = [4.5, 1]$? *(2 points)*
Solution: First, ω_1 is not here, so $p(\omega_1|x) = 0$. $p(\omega_2|x) = \frac{1/16 \cdot 1/3}{1/16 \cdot 1/3 + 1/8 \cdot 1/3} = 1/3$, and $p(\omega_3|x) = 2/3$.

c. Determine the classification error that the Bayes classifier is making, i.e. the Bayes error. *(2 points)*
Solution: $\varepsilon^* = p(\varepsilon|\omega_1)p(\omega_1) + p(\varepsilon|\omega_2)p(\omega_2) + p(\varepsilon|\omega_3)p(\omega_3) = 0 + 4/16 \times 1/3 + 0 = 1/12 = 0.0833$

d. Determine the decision boundary of the nearest mean classifier when it is trained on a extremely large dataset drawn from this distribution. Give an explicit formula that shows which $\mathbf{x}$ are part of this boundary. *(2 points)*
Solution: Means at $\mu_1 = (1,2)$, $\mu_2 = (3,2)$, $\mu_3 = (6,1)$. Boundary between class 1 and 2 at $x_1 = 2$. Boundary between class 2 and 3 has normal $\mathbf{w} = \mu_3 - \mu_2 = (3,-1)$, and the mid point $(4.5, 1.5)$ should get $w_1x_1 + w_2x_2 + b = 0$, so $b = -12$. If you are very smart you should also mention that for very negative values of x_2 we will at a certain moment also have the decision boundary between class 1 and 3.

e. Assume now that the class priors are changed to: $P(\omega_1) = P(\omega_3) = 1/5$ and $P(\omega_2) = 3/5$. How does the Bayes classifier change? How large does its error become? *(2 points)*
Solution: Now the left decision boundary does not change (still, $1/4 \cdot 1/5 > 1/16 \cdot 3/5$), but the right decision boundary does ($1/16 \cdot 3/5 > 1/8 \cdot 1/5$)! So still no error on class 1. Error on class 2 is 2/16 and on class 3 1/4. Total error is $0 + 2/16 \cdot 3/5 + 1/4 \cdot 1/5 = 0.125$.

3 Dissimilarity-based Classification *(9 points)*

Let us consider a two-class classification problem in two dimensions. Rather than constructing a classifier directly in this feature space, we decide to use a dissimilarity-based approach. The dissimilarity we use is simply the Euclidean distance (so *not* the squared one) between points in feature space.

To start with, we consider the dissimilarity representation based on a single prototype p in this two-dimensional feature space, which means all data points are represented by the distance to this single prototype.

a. Let us say we determine a nearest mean classifier (NMC) in this 1D dissimilarity space. Describe and/or draw how the decision boundary looks in the original feature space; what is the geometric shape of the decision boundary? *(2 points)*
Solution: The dissimilarity space is 1D and NMC just thresholds it at some distance. So, points in feature space within a particular circle around p are going to be assigned to the one class, while everything outside that circle are assigned to the other class.

b. What happens if we apply a quadratic discriminant (QDA or qdc) in the dissimilarity space? Describe what shapes and forms the decision boundary now can take on in the feature space. *(2 points)*
Solution: Similar reasoning as under a, but now we possibly get up to two concentric circles, which regions are alternately assigned to the one or the other class.

Let us make things a bit more complicated. We are now going to take two distinct prototypes $p = (-1, 0)$ and $q = (+1, 0)$, i.e., two points in the 2D feature space. As we consider a dissimilarity representation based on the Euclidean distance, not every feature pair in the dissimilarity space can be reached. For instance, we will never see a feature pair $(-2, 3)$, because a negative dissimilarity is not possible. The set of all attainable dissimilarity pairs is, however, much smaller than the positive quadrant. Using prototypes p and q, you are going to draw and explain what is the set of all points that can be attained by mapping 2D feature vectors into the dissimilarity space considered.

c. Determine the 2D coordinates in the dissimilarity space to which p and q are mapped. *(1 points)*
Solution: Prototype p ends up at $(0, \|p - q\|)$ and q at $(\|p - q\|, 0)$ or vice versa.

d. Consider the straight line going through p and q in the feature space. Onto which curve are the feature vectors along this line mapped in the dissimilarity space? Make a drawing and also clearly indicate where p and q end up. *(2 points)*
Solution: The sought-after curve is given by the line between $(0, \|p - q\|)$ and $(\|p - q\|, 0)$. And the two lines in the positive qudrant that leave from $(0, \|p - q\|)$ and $(\|p - q\|, 0)$, respectively, in the direction (1,1) (i.e., "under 45 degrees").

e. All point within the strip defined by the foregoing curve can be reached. Explain why a point like $(\frac{1}{2}, 3)$ can never be reached. *(1 points)*
Solution: A reference to a violation of the triangular inequality with some additional hand waving should do.

f. Find a classifier that is linear in this 2D dissimilarity space, which corresponds to a linear classifier in feature space as well. *(1 points)*
Solution: Corresponding to the line $x - y = 0$ in the dissimilarity space, we find the perpendicular bisector in the original feature space. Both are linear.

4 Feature Extraction *(10 points)*

Consider 1000 samples from a 3D Gaussian distribution. Assume this sample has zero mean and a 3 × 3-covariance matrix C given by

$$C = \begin{pmatrix} 99 & 0 & 0 \\ 0 & 99 & 0 \\ 0 & 0 & 124 \end{pmatrix} .$$

a. Which direction gives the first principle component for this data set? *(2 points)*
Solution: This should be a vector pointing in the direction of the third dimension.

It turns out, that these 1000 samples are in fact the class means of 1000 different classes of bird species described by three different features. Assume that all class priors are $\frac{1}{1000}$ and that all these classes are normally distributed with covariance matrix equal to the identity matrix.

b. Give the total covariance matrix (also called the mixture scatter matrix) for this data set. *(1 points)*
Solution: The total covariance is given by the sum of the 3D covariance of the Gaussian plus the identity matrix, which gives diag(100,100,125) as the solution.

c. The Fisher mapping determines the Eigenvectors with the largest Eigenvalues of the matrix $\Sigma_W^{-1}\Sigma_T$ (or, if you prefer, $\Sigma_W^{-1}\Sigma_B$.) Σ_W, Σ_B, and Σ_T are the within-class covariance, the between-class, and the total covariance matrices, respectively. Which direction gives the

optimal Fisher mapping if we want to reduce the feature space to one dimension? (So, we look for a single 3D vector here.) *(2 points)*
Solution: One can do the explicit calculations and find the Eigenvector with the largest Eigenvalue, but one can also see that, as the mean within class is spherical, we will find the same component as the one under a.

In a next step, a linear feature transformation T is applied to the original 3D space. The 3 × 3-matrix describing this transformation is given by

$$T = \begin{pmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

d. Recall your answer under b. and show that the total covariance matrix of the data after the transformation equals

$$\begin{pmatrix} 125 & 100 & 0 \\ 100 & 125 & 0 \\ 0 & 0 & 125 \end{pmatrix} .$$

(2 points)
Solution: One should work out the product $T'CT$ which leads to the asked for solution.

e. Are the first two features correlated and, if so, are they positively or negatively correlated? *(1 points)*
Solution: Considering the off diagonals, they are positively correlated.

f. Determine the first principal component for this new feature space. *(2 points)*
Solution: Realize that all direction along the diagonals have the same variance and that the first two features are positively correlated. This means, the largest for these two features is in the direction (1,1). Then realize that the variance in this direction must be larger than the individual variance. But this mean that in 3D, as the third feature is not correlated to the first two, that (1,1,0) has the largest variance and therefore is the first PC.

5 Clustering *(10 points)*

Assume we have a one-dimensional dataset with five objects: $x_1 = -0.5$, $x_2 = +0.5$, $x_3 = +3.0$, $x_4 = +3.5$ and $x_5 = 5.5$. First, we are clustering this dataset with a Mixture of Gaussians with $k = 2$ clusters. A mixture of 2 Gaussians models the data with the following probability density function:

$$p(x) = \sum_{i=1}^{2} P_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \tag{1}$$

with $\sum_i P_i = 1$.

a. Compute the loglikelihood of the data, given the cluster parameters: $\mu_1 = 0$, $\mu_2 = 3.5$, $\sigma_1 = \sigma_2 = 1$ and $P_1 = P_2 = 0.5$. *(2 points)*
Solution: First write:

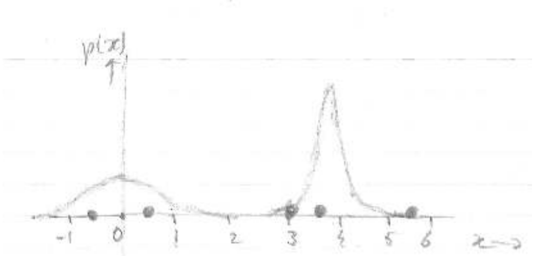
$$LL = \log(\prod_j p(x_j)) = \sum_{j=1}^{5} \log(p(x_j)) \tag{2}$$

Now for one object x_j fill in the σ's and P_i's:

$$\begin{aligned} \log(p(x_j)) &= \log(\sum_{i=1}^{2} \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_j-\mu_i)^2\right)) && (3) \\ &= \log(\frac{1}{2\sqrt{2\pi}}) + \log\left(\exp(-\frac{1}{2}x_j^2) + \exp(-\frac{1}{2}(x_j-3.5)^2)\right) && (4) \end{aligned}$$

This is already great. Now sum over the 5 objects and in the end I get $LL = -3.9595$.

b. Now change the standard deviation of class two to $\sigma_2 = 0.1$ (the standard deviation of class one stays the same, i.e. 1). Sketch this probability density and indicate where the 5 objects are located. Argue if the loglikelihood of this mixture is smaller, equal, or larger than the loglikelihood from question (a). *(2 points)*



Solution: The second gaussian is tighter and the peak is much higher. The (log)likelihood is higher: the likelihood for object x_4 increases tremendously.

c. **THIS QUESTION IS NOT GRADED!** Give a solution of the Mixture model, in terms of $\mu_1, \mu_2, \sigma_1, \sigma_2, P_1$ and P_2 for which the loglikelihood is maximal. *(2 points)*
Solution: If you keep $\mu_2 = 3.5$, you can $\sigma_2 \to 0$ to make the loglikelihood go to infinity!

Consider now the situation that we are clustering using hierarchical clustering, with average linkage.

d. Perform the hierarchical clustering, and draw the fusion graph (also called the threshold dendrogram). *(2 points)*
Solution: Fusion levels are 0.5, 1, 2.25 and 4 for average linking.

e. Compute the cophenetic correlation coefficient (also known as the Pearson correlation) between the true distances and the dendrogram distances of objects x_3, x_4, and x_5. *(2 points)*
Solution: The true distances are

$$D = \begin{pmatrix} 0 & 1 & 3.5 & 4 & 6 \\ 1 & 0 & 2.5 & 3 & 5 \\ 3.5 & 2.5 & 0 & 0.5 & 2.5 \\ 4 & 3 & 0.5 & 0 & 2 \\ 6 & 5 & 2.5 & 2 & 0 \end{pmatrix} \tag{5}$$

The distances over the dendrogram are:

$$D = \begin{pmatrix} 0 & 1 & 4 & 4 & 4 \\ 1 & 0 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0.5 & 2.25 \\ 4 & 4 & 0.5 & 0 & 2.25 \\ 4 & 4 & 2.25 & 2.25 & 0 \end{pmatrix} \tag{6}$$

So for the full correlation, we need to compute the correlation coefficient between $d_1 = [1, 3.5, 4, 6, 2.5, 3, 5, 0.5, 2.5, 2]$ and $d_2 = [1, 4, 4, 4, 4, 4, 4, 0.5, 2.25, 2.25]$. It appears $\rho = 0.8175$.

For the subset of the three objects we only have coefficient between $d_1 = [0.5, 2.5, 2]$ and $d_2 = [0.5, 2.25, 2.25]$. Then $\rho = 0.97$.

The correlation in itself is not so important, but the fact that you compute it over the distances *is*.

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. The Bayes error for a two-class classification problem is smaller than the Bayes error of a three-class problem.
   TRUE FALSE
   Solution: FALSE

2. To train the quadratic classifier, you need to estimate the class prior probabilities.
   TRUE FALSE
   Solution: TRUE

3. For a k-nearest neighbor classifier you need to estimate the class prior probabilities.
   TRUE FALSE
   Solution: FALSE

4. To train the logistic classifier, you need to estimate the class conditional probabilities.
   TRUE FALSE
   Solution: FALSE

5. Rescaling the features of a classification problem may improve the classification performance of the nearest mean classifier.
   TRUE FALSE
   Solution: TRUE

6. If no features are available, the best classification in a two-class problem is realized by assigning all objects to the class with the largest prior probability.
   TRUE FALSE
   Solution: TRUE

7. The classification error found for an infinite training set is for the Bayes classifier a non-increasing function of the number of features.
   TRUE FALSE
   Solution: TRUE

8. Due to their immense flexibility, deep nets (i.e., an artificial neural network with a lot of hidden layers) typically outperform all other classifiers.
   TRUE FALSE
   Solution: FALSE

9. If one knows the AUC (the area under the ROC) associated with a classifier, then one can also determine the classification error rate that classifier gives.
   TRUE FALSE
   Solution: FALSE

10. Using the reject option typically leads to improved error rates.
TRUE FALSE
Solution: TRUE

Answer sheet 2

Name :
Student number :

Clustering

1. Clustering is a supervised technique, because it needs the number of clusters.
   TRUE FALSE
   Solution: FALSE

2. The within-scatter criterion value typically becomes smaller when you increase the number of clusters k.
   TRUE FALSE
   Solution: TRUE

3. The result of hierarchical clustering with complete linkage depends on the scaling of the features.
   TRUE FALSE
   Solution: TRUE

4. When the number of objects in a clustering problem is very large, the best clustering algorithm is the k-means clustering.
   TRUE FALSE
   Solution: FALSE

5. To find the optimum number of clusters for a k-means clustering, the log-likelihood has to be optimized.
   TRUE FALSE
   Solution: FALSE

Feature Selection and Extraction

1. If a kernel mapping of an initial feature space is complex enough, we can reduce the Bayes error of our problem.
   TRUE FALSE
   Solution: FALSE

2. Even if one uses error rate as performance criterion and does exhaustive search over all feature combinations, it remains impossible to guarantee finding the best features due to the finite sample size one is dealing with.
   TRUE FALSE
   Solution: TRUE

3. Individual feature selection is a powerful and convenient method because it takes the correlation between features into account.
   TRUE FALSE
   Solution: FALSE

4. Prototype selection for dissimilarity representations can be dealt with through feature selection.
   TRUE FALSE
   Solution: TRUE

5. For a three dimensional feature space, feature selection admits 6 essentially different proper subspaces. Proper means that you should not count the three-dimensional and the zero-dimensional subspaces.
TRUE FALSE
Solution: TRUE