Hannah Schulz-Kümpel June 6th 2024

Supervised Learning & Distance and Similarity measures

Question 1:

a) As part of a study, objects are to be grouped meaningfully according to similarity criteria. The following objects were observed:

i. Berlin bars (regarding standardized, uncorrelated measurements of average number of visitors per week and time since opening)

ii. Distributions of two random variables X and Y (e.g. two normal distributions with different parameters)

iii. English surnames

iv. Boutiques in Munich (in terms of location/coordinates)

v. Ten bytes (1 byte = 8 bits) e.g. [10001010] vs. [11001010] vs. [00101010] vs. ....

vi. Exam solutions of two high school graduates (plagiarism detection)

Which distance and/or similarity measures would you propose to deal with these kinds of objects?

b) Is the squared Euclidean distance, defined as

$$D_{\text{Euk}}(x, y)^2 = \sum_{i=1}^{p} |x_i - y_i|^2$$

a metric? Prove your answer.

Question 2:

Consider the following subset from the `roc_sim_dat.csv` data set
(Source: `http://static.lib.virginia.edu/statlab/materials/data/roc_sim_dat.csv`):

You may assume that the probabilities were predicted by some logistic model.

a) Write pseudo-code or the code of an R function to calculate the *false positive fraction* (FPF) and *true positive fraction* (TPF) from above data for a set of threshold values.

b) Draw the *receiver operating characteristic* (ROC) for the following thresholds:

$$-\infty\,; \quad 0.115\,; \quad 0.125\,; \quad 0.145\,; \quad 0.185\,; \quad 0.220\,; \quad 0.260\,; \quad 0.325\,; \quad \infty$$

predicted_prob_of_Yes	actual_outcome
0.13	Yes
0.16	No
0.11	No
0.12	No
0.23	No
0.11	No
0.29	Yes
0.13	No
0.21	No
0.36	No

c) Calculate the *area under the curve* (AUC). What would you say about the model that produces the predicted probabilities based on the AUC value?

Question 3:

In this exercise, consider patients from a cardiologist's practice that are divided according to the risk of myocardial infarction (Y). Specifically, the assignment to *class* 1 does not indicates an increased risk, while the assignment to *class* 2 indicated an increased risk. Furthermore, the results of the electrocardiogram (X) are given, which are divided into *good (G)* and *bad (S)*. The conditional distribution $f(x|y)$ and the a priori probabilities for the respective class memberships $Y \in \{1, 2\}$ are given by the following table:

	good Electrocardiogram G	bad Electrocardiogram S	a priori-probabilities
class 1	0.95	0.05	π
class 2	0.10	0.90	$1 - \pi$

a) Determine the Bayesian classification as a function of the parameter π. If no clear assignment is possible, make an assignment to class 1.

b) Determine the error rates ϵ_{12} and ϵ_{21} as well as ϵ for $\pi = 0.2$.

c) What is the difference between Bayesian and ML classification? What would be the decision rule for ML classification?

d) Next, assume that it is worse to assume a patient to be at risk than risk-free (and therefore not to start treatment), than to perform a further and unnecessary examination on a risk-free patient. We can take this fact into account by introducing costs. Which assignments result for $\pi = 0.2$ when additionally taking into account the following cost table

c_{ij}	1	2
1	0	1
2	5	0

Question 4:

Consider a two dimensional feature vector $\boldsymbol{X}$ that is normally distributed in three classes. Specifically

$$\begin{aligned}
\boldsymbol{X} \mid Y = 1 &\sim N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \quad \text{with} \quad \boldsymbol{\mu}_1 = (4, 12)^\top, \\
\boldsymbol{X} \mid Y = 2 &\sim N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \quad \text{with} \quad \boldsymbol{\mu}_2 = (12, 8)^\top, \\
\boldsymbol{X} \mid Y = 3 &\sim N_2(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}) \quad \text{with} \quad \boldsymbol{\mu}_3 = (4, 8)^\top.
\end{aligned}$$

with a priori probabilities $p(1) = p(2) = p(3) = 1/3$.

a) Write out the discriminant function for each class when using *linear discriminant analysis* (LDA) for a general $\boldsymbol{\Sigma}$.

Next let the covariance matrix be equal to the identity matrix, i.e. $\boldsymbol{\Sigma} = \boldsymbol{I}$.

b) Calculate the specific dividing lines between the classes and sketch the areas in which the points classified to each class would have to lie.