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.

Solution:

a) i. **Euclidean distance.** This is a good alternative when standardized, uncorrelated values of metric variables are compared with each other.

ii. **Wasserstein metric.** This divergence can be used to compare probability (density) functions. However, it is not a metric because it is not symmetrical.

iii. **Levenshtein distance.** The most morphologically similar first names can be grouped together in this way.

iv. **Manhattan distance.** Due to the block structure, the Manhattan distance is generally more realistic than the Euclidean distance in such cases.

v. **Hamming distance.** This distance is suitable for measuring the difference between character strings (often binary in the application).

vi. **Cosine distance.** If the content is of interest, for example, the proportion of words in each exam can be compared using cosine similarity; if they are very similar, this could indicate a case of plagiarism

b) (a) Positive definiteness:

$$D_{\text{Euk}}(x,y)^2 \geq 0 \quad \checkmark$$

and $d(x,y) = 0 \Leftrightarrow x = y$?

Consider $x = y$. Then: $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Rightarrow x_1 = y_1 \wedge x_2 = y_2$

$$D_{\text{Euk}}(x,y)^2 = |0| + |0| = 0 \quad \checkmark$$

Next, consider $d(x,y) = 0$. Then $|x_1 - y_1| = 0 \wedge |x_2 - y_2| = 0 \Rightarrow x_1 - y_1 = 0 \wedge x_2 - y_2 = 0$
$\Rightarrow x_1 = y_1 \wedge x_2 = y_2$
$\Rightarrow x = y$ ✓

(b) Symmetry:

$$d(x,y) = d(y,x)$$

$$\begin{aligned} d(x,y) &= |x_1 - y_1|^2 + |x_2 - y_2|^2 \\ &= |-(y_1 - x_1)|^2 + |-(y_2 - x_2)|^2 \\ |-a| = |a| & \\ \Rightarrow d(x,y) &= |y_1 - x_1|^2 + |y_2 - x_2|^2 \\ &= d(y,x) \quad \checkmark \end{aligned}$$

(c) Triangle inequality:

$$d(x,y) \leq d(x,z) + d(z,y)$$

Consider an arbitrary $x \in \mathbb{R}^n \backslash \{0\}$ and set $y = 3x$ and $z = 2x$. Then

$$d(x,y)^2 = \sum_{i=1}^{n} (x_i - 3x_i)^2 = 4 \sum_{i=1}^{n} x_i^2$$

and

$$d(x,z)^2 + d(z,y)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i^2 = 2 \sum_{i=1}^{n} x_i^2$$

Since $4 \sum_{i=1}^{n} x_i^2 > 2 \sum_{i=1}^{n} x_i^2$, we have found a counterexample in which the triangle inequality does not apply.

$\Rightarrow$ Therefore, the squared Euclidean distance isn't a metric.

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`):

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

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:

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

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?

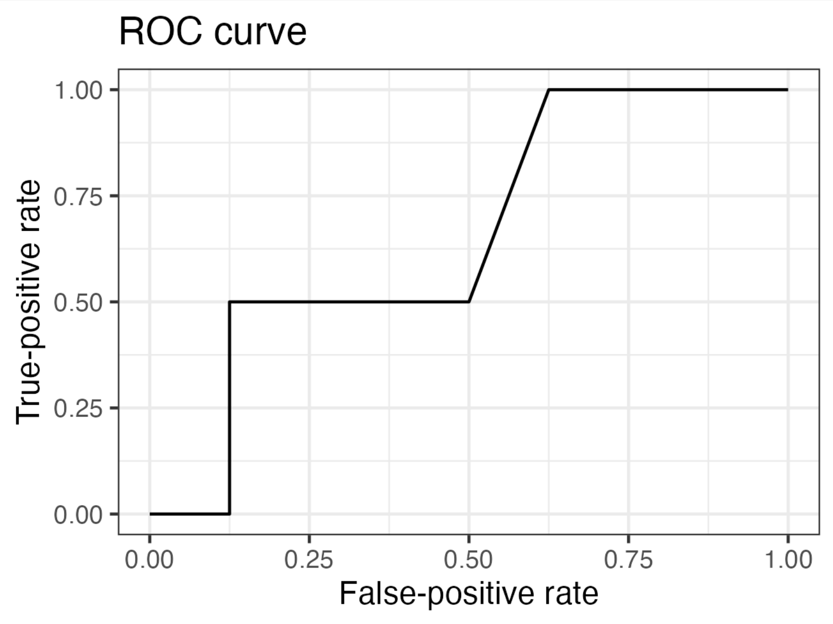
Solution:

a) The following R function returns a data frame with FPF in the first column and TPF in the second:

```
get_fpf_tpf <- function(predicted_probs,actual_outcomes,thresholds){
  output <- data.frame(FPF=numeric(length(thresholds)),
                       TPF=numeric(length(thresholds)))
  for(i in 1:length(thresholds)){
    predicted_classes <- ifelse(predicted_probs >= thresholds[i], 1, 0)
    TP <- sum(predicted_classes == 1 & actual_outcomes == "Yes")
    FP <- sum(predicted_classes == 1 & actual_outcomes == "No")
    TN <- sum(predicted_classes == 0 & actual_outcomes == "No")
    FN <- sum(predicted_classes == 0 & actual_outcomes == "Yes")

    output$FPF[i] <- FP / (FP + TN)
    output$TPF[i] <- TP / (TP + FN)
  }
  return(output)
}
```

b)



c) The AUC gives the area under the curve, i.e. the integral under the ROC on $[0, 1]$. In this case, we have

$$\text{AUC} = (0.375 \cdot 0.5) + 0.5 - (0.125 \cdot 0.25) = 0.65625\,.$$

This is relatively close to 0.5, so the model does not seem to be optimal.

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

Recap:

- **Basic problem of discriminant analysis**:
  With the help of an observed feature vector $\boldsymbol{x} \in \mathbb{R}^p$, determine the class class $Y \in \{1, \ldots, k\}$ from which the observation originates
- The following quantities play a role here:

$f_{X\|Y}(\boldsymbol{x}\|y) \hat{=} f(\boldsymbol{x}\|y)$	sampling distribution	(known, at least as an estimate)
$P(Y = y) \hat{=} p(y)$	a priori-probability	(known, at least as an estimate)
$f(\boldsymbol{x}) = \sum_{y=1}^{K} f(\boldsymbol{x}\|y)p(y)$	mixture distribution	(can be calculated from the two previous items)
$P(Y = y\|X = \boldsymbol{x}) \hat{=} p(y\|\boldsymbol{x})$	a posteriori-probability	(unknown)

- What we want: Classification rule that assigns an observed feature vector $\boldsymbol{x}$ to a class r ($r \in \{1, \ldots, k\}$)
- Approach for Bayesian classification:

$$\delta(\boldsymbol{X} = \boldsymbol{x}) = r \quad \Longleftrightarrow \quad P(Y = r|\boldsymbol{X} = \boldsymbol{x}) = \max_{i=1,\ldots,k} P(Y = i|\boldsymbol{X} = \boldsymbol{x}),$$

  i.e. assign the observation to the class for which the observed feature vector has the highest posteriori probability.
- Problem: $P(Y = r|\boldsymbol{X} = \boldsymbol{x})$ is unknown!
- With the help of Bayes' theorem, though, the following relation can be shown

$$P(Y = i|\boldsymbol{X} = \boldsymbol{x}) = p(i|\boldsymbol{x}) = \frac{f(i, \boldsymbol{x})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|i)p(i)}{f(\boldsymbol{x})} \propto f(\boldsymbol{x}|i)p(i)$$

- Which leads us to the Bayes rule:

$$\delta(\boldsymbol{X} = \boldsymbol{x}) = r \quad \Longleftrightarrow \quad f(\boldsymbol{x}|r)p(r) = \max_{i=1,\ldots,k} f(\boldsymbol{x}|i)p(i).$$

Solution:

a) For $X \in \{G, S\}, Y \in \{1, 2\}$, we start by calculating the following probabilities:

$$\begin{aligned}
P(X = G|Y = 1)P(Y = 1) &= 0.95\pi \\
P(X = G|Y = 2)P(Y = 2) &= 0.1(1 - \pi) \\
P(X = S|Y = 1)P(Y = 1) &= 0.05\pi \\
P(X = S|Y = 2)P(Y = 2) &= 0.9(1 - \pi).
\end{aligned}$$

The decision-rule for $X = G$ is

$$\delta(X = G) = r \quad \Leftrightarrow \quad P(X = G|Y = r)P(Y = r) = \max_i P(X = G|Y = i)P(Y = i).$$

In our case we make the decision $Y = 1$ – in case of equality $Y = 1$ should also be chosen – if

$$\begin{aligned}
0.95\pi &\geq 0.1(1 - \pi) \\
\Leftrightarrow \quad 0.95\pi &\geq 0.1 - 0.1\pi \\
\Leftrightarrow \quad 1.05\pi &\geq 0.1 \\
\Leftrightarrow \quad \pi &\geq \frac{2}{21} \approx 0.0952.
\end{aligned}$$

As a result, the classification rule for $X = G$ is given by

$$\delta(X = G) = \begin{cases} 1, & \pi \geq 2/21, \\ 2, & \pi < 2/21 \end{cases}$$

Therefore, we make the decision $Y = 1$ when $X = S$ exactly when

$$\begin{aligned}
0.05\pi &\geq 0.9(1 - \pi) \\
\Leftrightarrow \quad 0.05\pi &\geq 0.9 - 0.9\pi \\
\Leftrightarrow \quad 0.95\pi \geq 0.9 & \\
\Leftrightarrow \quad \pi &\geq \frac{18}{19} \approx 0.9474
\end{aligned}$$

and the decision rule for $X = S$ is given by

$$\delta(X = S) = \begin{cases} 1, & \pi \geq 18/19, \\ 2, & \pi < 18/19 \end{cases}$$

b) The definition of individual error rates is

$$\epsilon_{rs} = P(\delta(X) = s|Y = r).$$

In our case, we are looking for ϵ_{12} and ϵ_{21} as well as ϵ for $\pi = 0.2$. For ϵ_{12} we get

$$\begin{aligned}
\epsilon_{12} &= P(\delta(X) = 2|Y = 1) = \frac{P(\delta(X) = 2, Y = 1)}{P(Y = 1)} \\
&= \frac{P(\delta(X) = 2, Y = 1, X = G) + P(\delta(X) = 2, Y = 1, X = S)}{P(Y = 1)} \\
&= \frac{P(\delta(X) = 2|Y = 1, X = G)P(Y = 1, X = G) + P(\delta(X) = 2|Y = 1, X = S)P(Y = 1, X = S)}{P(Y = 1)} \\
&= \frac{P(\delta(X) = 2|Y = 1, X = G)P(X = G|Y = 1)P(Y = 1)}{P(Y = 1)} \\
&\quad + \frac{P(\delta(X) = 2|Y = 1, X = S)P(X = S|Y = 1)P(Y = 1)}{P(Y = 1)} \\
&= P(\delta(X) = 2|Y = 1, X = G)P(X = G|Y = 1) + P(\delta(X) = 2|Y = 1, X = S)P(X = S|Y = 1) \\
&= P(\delta(X = G) = 2|Y = 1)P(X = G|Y = 1) + P(\delta(X = S) = 2|Y = 1)P(X = S|Y = 1) \\
&= 0 \cdot 0.95 + 1 \cdot 0.05 = 0.05.
\end{aligned}$$

Analogously, we get

$$\begin{aligned}
\epsilon_{21} &= P(\delta(X = G) = 1|Y = 2)P(X = G|Y = 2) + P(\delta(X = S) = 1|Y = 2)P(X = S|Y = 2) \\
&= 1 \cdot 0.1 + 0 \cdot 0.9 = 0.1.
\end{aligned}$$

For the total error rate, we get

$$\begin{aligned}
\epsilon = \sum_{r=1}^{2} \sum_{s \neq r} \epsilon_{rs} P(Y = r) &= \epsilon_{12} P(Y = 1) + \epsilon_{21} P(Y = 2) \\
&= 0.05 \cdot \pi + 0.1 \cdot (1 - \pi) \\
&= 0.05\pi + 0.1 - 0.1\pi \\
&= 0.1 - 0.05\pi \\
&\overset{\pi=0.2}{=} 0.09.
\end{aligned}$$

c) ML classification is a special case of Bayesian classification in which the priori probabilities are all equal, i.e. $p(1) = p(2) = \ldots = p(g) = 1/g$.
Since all priori probabilities are equal, these can be neglected with regard to the discriminant function, so that only the conditional probabilities of X given Y play a role:

$$\begin{aligned}
P(X = G|Y = 1) &= 0.95 \\
P(X = G|Y = 2) &= 0.1 \\
P(X = S|Y = 1) &= 0.05 \\
P(X = S|Y = 2) &= 0.9.
\end{aligned}$$

Here, it holds that:

$$\begin{aligned}
P(X = G|Y = 1) &> P(X = G|Y = 2) \\
P(X = S|Y = 1) &< P(X = S|Y = 2).
\end{aligned}$$

Therefore, the decision-rule w.r.t. X is

$$\delta(X = G) = 1\ , \quad \delta(X = S) = 2.$$

d) Cost-optimal Bayes classification:

$$\delta(\mathbf{x}) = r \Leftrightarrow \sum_{i=1}^{k} p(i|\mathbf{x}) \cdot c_{ir} = \min_j \sum_{i=1}^{k} p(i|\mathbf{x}) \cdot c_{ij}$$

where c_{ij} denotes the cost when an object that belongs to class i, but gets classified to class j.

Here: $c_{12} = 1, \quad c_{21} = 5, \quad c_{11} = c_{22} = 0$

<u>mit $\pi = 0,2$:</u>

- $X = G$:

$$\begin{aligned}
\min\{ \quad P & \quad (Y = 1|X = G) \cdot c_{11} + P(Y = 2|X = G) \cdot c_{21}; \\
P & \quad (Y = 1|X = G) \cdot c_{12} + P(Y = 2|X = G) \cdot c_{22}\} \\
= & \quad \min\{P(X = G|Y = 2)P(Y = 2) \cdot c_{21}; \\
& \quad P(X = G|Y = 1)P(Y = 1) \cdot c_{12}\} \\
= & \quad \min\{0.1 \cdot 0.8 \cdot 5; 0.95 \cdot 0.2 \cdot 1\} \\
= & \quad \min\{0.4; 0.19\} = 0.19
\end{aligned}$$

  $\Rightarrow \delta(G) = 2$

- $X = S$:

$$\begin{aligned}
\min\{ \quad P & \quad (Y = 1|X = S) \cdot c_{11} + P(Y = 2|X = S) \cdot c_{21}; \\
P & \quad (Y = 1|X = S) \cdot c_{12} + P(Y = 2|X = S) \cdot c_{22}\} \\
= & \quad \min\{P(X = S|Y = 2)P(Y = 2) \cdot c_{21}; \\
& \quad P(X = S|Y = 1)P(Y = 1) \cdot c_{12}\} \\
= & \quad \min\{0.9 \cdot 0.8 \cdot 5; 0.05 \cdot 0.2 \cdot 1\} \\
= & \quad \min\{3.6; 0.01\} = 0.01
\end{aligned}$$

  $\Rightarrow \delta(S) = 2$

$\Rightarrow$ patients are always considered to have an increased risk of heart attack.

<u>Note:</u> Bayes and ML classification are special cases of cost-optimal classification:

- Bayes: $c_{ij} = c$
- ML: $c_{ij} = \frac{c}{p(i)}, \quad i \neq j$

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.

Solution:

a) We use the following discriminant function (important: identical covariance matrices in all classes are assumed!)

$$
\begin{aligned}
d_r(\boldsymbol{x}) &= \log(f(\boldsymbol{x}|r)) + \log(p(r)) \\
&= \log\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_r)^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_r) + \log(p(r)).
\end{aligned}
$$

Since the first term on the right-hand side is identical for all classes, we can neglect it in the discriminant function and get

$$d_r(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_r)^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_r) + \log(p(r)).$$

Multiplication results in

$$d_r(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}_r^\top \Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_r^\top \Sigma^{-1}\boldsymbol{\mu}_r) + \log(p(r)).$$

The first term is again identical for all classes and can be neglected in the following. The same applies to $\log(p(r))$ if identical a priori probabilities are assumed.
This gives us the following discriminant function

$$d_r(\boldsymbol{x}) = \boldsymbol{\mu}_r^\top \Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_r^\top \Sigma^{-1}\boldsymbol{\mu}_r \tag{1}$$

b) For $\Sigma = I$, it now holds that

$$d_r(\boldsymbol{x}) = \boldsymbol{\mu}_r^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_r^\top \boldsymbol{\mu}_r. \tag{2}$$

We get

$$
\begin{aligned}
d_1(\boldsymbol{x}) &= \begin{bmatrix} 4 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 4 & 12 \end{bmatrix} \begin{bmatrix} 4 \\ 12 \end{bmatrix} \\
&= 4x_1 + 12x_2 - 80
\end{aligned}
$$

$$\begin{aligned} d_2(\boldsymbol{x}) &= \begin{bmatrix} 12 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 12 & 8 \end{bmatrix} \begin{bmatrix} 12 \\ 8 \end{bmatrix} \\ &= 12x_1 + 8x_2 - 104 \end{aligned}$$

$$\begin{aligned} d_3(\boldsymbol{x}) &= \begin{bmatrix} 4 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 4 & 8 \end{bmatrix} \begin{bmatrix} 4 \\ 8 \end{bmatrix} \\ &= 4x_1 + 8x_2 - 40. \end{aligned}$$

We decide in favor of category i and against j, if $d_i(\boldsymbol{x}) \geq d_j(\boldsymbol{x})$ applies. For categories 1 and 2 we get

$$\begin{aligned} 4x_1 + 12x_2 - 80 &\geq 12x_1 + 8x_2 - 104 \\ \iff \quad 4x_2 &\geq 8x_1 - 24 \\ \iff \quad x_2 &\geq 2x_1 - 6. \end{aligned}$$

For categories 1 and 3 we get

$$\begin{aligned} 4x_1 + 12x_2 - 80 &\geq 4x_1 + 8x_2 - 40 \\ \iff \quad 4x_2 &\geq 40 \\ \iff \quad x_2 &\geq 10. \end{aligned}$$

For categories 2 and 3 we get

$$\begin{aligned} 12x_1 + 8x_2 - 104 &\geq 4x_1 + 8x_2 - 40 \\ \iff \quad 0 &\geq -8x_1 + 64 \\ \iff \quad x_1 &\geq 8. \end{aligned}$$