Probability and Learning

1. **Bayes Theorem.**

   An economics consulting firm has created a model to predict recessions. The model predicts a recession with probability 90% when a recession is indeed coming and with probability 20% when no recession is coming. The unconditional probability of falling into a recession is 30%. If the model predicts a recession, what is the probability that a recession will indeed come?

2. **Canteen - Data.**

   A student select a food in canteen as follows:

Hunger,H	Diet,D	Time,H	Food,F
Great	*No*	*Early*	*Chips*
Great	*Yes*	*Early*	*Salad*
Middle	*No*	*Early*	*Soup*
Middle	*Yes*	*Early*	*Salad*
Great	*No*	*Late*	*Soup*
Great	*Yes*	*Late*	*Salad*
Small	*Yes*	*Late*	*Salad*
Small	*No*	*Late*	*Salad*
Small	*No*	*Early*	*Salad*
Middle	*Yes*	*Late*	*Salad*

   (a) Predict manually with Naive-Bayes-Classifier a food for following combinations:

      i. H = Great, D = No, Time = Early
      ii. H= Middle, D= Yes, Time = Late

   (b) Comment the results. What problems can arise?

   (c) Proof the results with a program.

3. Apply the logistic regression to the data from Exercise 2. What will be you answer on the questions from Exercise 2.

4. Apply the linear discriminant analysis to the data from Exercise 2. What will be you answer on the questions from Exercise 2 in that case?

5. **Default - Data.**

   Consider the artificial data set shown in following table:

Time with Emp.,T	Size of Loan,S	Home-owner,H	Default,D
5	10.000	1	0
20	10.000	1	0
1	25.000	1	0
1	15.000	3	0
15	2.000	2	0
6	12.000	1	0
1	5.000	2	1
12	8.000	2	1
3	10.000	1	1
1	5.000	3	1

   Construct a rule which will allow prediction of variable D for future customers, where D is default on a bank loan (the last column, labeled 1 for default and 0 for non default).

   The variables which will be used for the prediction are columns 1 to 3:

   - time with the current employeer, T, in years;
   - size of loan requested, S, in dollars;
   - H, where the applicant is a homeowner(1), rental tenant (2), or other (3)

   For T use the classes $(T < 10)$, $(T \geq 10)$

   For S use the classes $(S \leq 10000)$, $(S > 10000)$

   Suppose the new application form is received from an applicant who has been with his or her employer for less then 10 years $(T < 10)$, is seeking a loan of 10.000 $(S \leq 10000)$, and is a home-owner $(H = 1)$

   How will you classify this customer?

6. Apply the logistic regression to the data from Exercise 5. What will be you answer on the question from Exercise 5.

7. Apply the linear discriminant analysis to the data from Exercise 5. What will be you answer on the question from Exercise 5 in that case?

8. **Two 2-dimensional Gaussian**.

   Generate samples of size 100 from each of the two classes. Class 1 is bivariate normal, with zero means and identity covariance matrix:

   $$X_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

   with

   $$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

   Class 2 is bivariate normal, with mean vector $(0, 2)$ and diagonal covariance matrix with leading diagonal $(1, 2)$ and identity covariance matrix:

   $$X_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

   with

   $$\boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

   Apply the logistic regression algorithm to classify this data. Analyse the result.

9. Use the LDA algorithm to fit the data from Exercise 8. Show the results graphically. Use cross-validation and compute ROC curve.

10. Use the LDA algorithm for the dimentionality reduction on a the **spambase** dataset (to load from the **UCI Machine Learning Repository** http://archive.ics.uci.edu/ml/). Show the results graphically.

11. Using the same data set as in Exercise 10 construct the logistic regression model. Show the results graphically. Use cross-validation and compute ROC curve.