8 Naive Bayes Test

Exercise 8.1 (MAP Rule for the Bernoulli Distribution)
Suppose that $X = (X_1, X_2, \ldots, X_n)$ is a random sample of n independent realizations of the Bernoulli distribution with success parameter p described hereafter. The label C of X can take on two values, $C = 0$ or $C = 1$, with probabilities $\Pr(C = 1) = q$ and $\Pr(C = 0) = 1 - q$. We assume that

$$\Pr(X_i = 1|C = k) = p_k, \forall i = 1, \ldots, n,$$

where $k \in \{0, 1\}$ and $0 < p_0 < p_1 < 1$ are some specified values.

1. Calculate $\Pr(X_1 = x_1, \ldots, X_n = x_n)$ where $x = (x_1, x_2, \ldots, x_n)$ is the observed sample.
2. Calculate $\Pr(C = k|X_1 = x_1, \ldots, X_n = x_n)$ for $k \in \{0, 1\}$.
3. Give the MAP rule.
4. Show that the MAP rule can be simplified under the form

$$\begin{cases} \text{assign } C = 0 \text{ to } x \text{ if } & f(x) \leq h, \\ \text{assign } C = 1 \text{ to } x \text{ if } & f(x) > h, \end{cases}$$

   where $h = \frac{1}{n}\text{logit}(1 - q)$, $f(x) = a\bar{x} + b$ is a linear function of $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and (a, b) are some real coefficients to determine.

Exercise 8.2 (Naive Bayes Classification proposed by R)

1. Install, if necessary, the library “e1071”.
2. Execute the commands :

```
data(iris)
m <- naiveBayes(iris[,-5], iris[,5])
t <- table(predict(m, iris), iris[,5])
```

3. Describe precisely the role of each command. You should explain very carefully the output of “naiveBayes” and the links with the lecture.
4. What is the content of the variable “t” ? Interpret it.

Exercise 8.3 (Naive Bayes Classification by Yourselves)

1. Download the data set “Titanic” with “data(Titanic)” and describe briefly its content. If necessary, the data set can be converted into the data frame format with the command :

```
df<-as.data.frame(Titanic)
```

2. Create your own R code to compute the Naive Bayes test which predicts the variable “Survived” from the other variables. The tasks are the followings :
   (a) Describe carefully the features used to make the prediction. How many features do you have ? How many samples do you have ?
   (b) To assess the quality of our test, we use the **k-fold cross-validation**. The original data set is randomly partitioned into k equal sized subsamples (or almost equal sized subsammples). Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k

subsamples used exactly once as the testing data set. The k results from the folds can then be averaged to produce a single estimation.

Cut randomly the data set into $k = 10$ subsamples.

(c) Learn the naive Bayes test from a training data set composed of $k-1$ subsamples (code by yourself the learning step).

(d) Compute the false negative rate and the false positive rate from the training data set.

(e) Compute the false negative rate and the false positive rate from the testing data set composed of the remaining subsamples.

(f) Repeat k times the steps (d)-(e)-(f) to estimate k times the false negative rate and the false positive rate (as explained in step (c)). Plot them on a figure. Discuss the figure.

(g) Compute the average false negative rate and the average false positive rate. Compute the standard deviation of the average estimates.