TD 1

Exercice 1 – Bayesian Classifier (Author : F. Rossi)

Q 1.1 We consider the dataset of votes cast by members of the United States House of Representatives in 1984 on 16 major proposals. Each individual is a member of the House described by 17 nominal variables. The variable *Party* takes the modalities Democrat and Republican. The other variables, V_1 to V_{16}, represent the votes and take the values YES, NO, and ABSTAIN (for a missed vote). There are 267 Democratic representatives and 168 Republican representatives.

		NO	ABSTAIN	YES
Republicans	V_1	134	3	31
	V_2	73	20	75
	V_3	142	4	22
	V_4	2	3	163
	V_5	8	3	157
	V_6	17	2	149
	V_7	123	6	39
	V_8	133	11	24
	V_9	146	3	19
	V_{10}	73	3	92
	V_{11}	138	9	21
	V_{12}	20	13	135
	V_{13}	22	10	136
	V_{14}	3	7	158
	V_{15}	142	12	14
	V_{16}	50	22	96

Democrats	V_1	102	9	156
	V_2	119	28	120
	V_3	29	7	231
	V_4	245	8	14
	V_5	200	12	55
	V_6	135	9	123
	V_7	59	8	200
	V_8	45	4	218
	V_9	60	19	188
	V_{10}	139	4	124
	V_{11}	126	12	129
	V_{12}	213	18	36
	V_{13}	179	15	73
	V_{14}	167	10	90
	V_{15}	91	16	160
	V_{16}	12	82	173

Q 1.1.1 How many different values are possible for the vote vector ?

Q 1.1.2 Given the vote vector of a representative :
V = (YES, NO, ABSTAIN, YES, NO, YES, YES, YES, NO, NO, YES, NO, NO, NO, NO, YES)
How can we estimate whether they are Republican or Democrat ?

Q 1.2 We consider two populations, men H with an average height of 1.74m and a standard deviation of 0.07m, and women F with an average height of 1.62m and a standard deviation of 0.065m (data from INSEE 2001). The population H contains $|h|$ individuals, and the population F contains $|f|$ individuals. We assume that the height distributions are Gaussian within each subpopulation.
A person is chosen randomly and uniformly from the total population. We want to determine their subpopulation based solely on their height : this is a classification problem based on a continuous variable.

Q 1.2.1 Define the random variable G representing the gender of a randomly chosen person. Provide the distribution of G.

Q 1.2.2 Define the random variable T representing the height of a randomly chosen person. Provide the density of T and $P(G = f|T = t)$.

Q 1.2.3 Provide the optimal Bayesian classifier.

Q 1.2.4 Assume $|h| = |f|$. Specify the decisions made by the optimal classifier and interpret this decision strategy.

Q 1.3 How many parameters does the Bayesian classifier have ?

Exercice 2 – Bayesian Classifier

Let $\mathcal{X}$ be a feature space in $\mathbb{R}^d$ and $\mathcal{Y}$ the set of labels $\{y_1, ..., y_l\}$.

Q 2.1 Recall what a Bayesian classifier is.

Q 2.2 Express the error made by the Bayesian classifier at a point $\mathbf{x}$. Is this error minimal ?

Q 2.3 Let $\lambda(y_j, y_i)$ be the cost of predicting label y_j instead of y_i. What are the values of λ in the case of 0-1 loss ? Provide examples of asymmetric costs and their use cases.

Q 2.4 What is the expression of the risk $R(y_i|\mathbf{x})$ of predicting y_i given $\mathbf{x}$ as a function of λ and posterior probabilities ? What about the 0-1 loss case ?

Q 2.5 Provide the expression for the risk $R(f)$ of a classifier f over $\mathcal{X}$.

Q 2.6 In the binary case, express the decision criterion in terms of λ, posterior probabilities, class distributions, and likelihoods.

Exercice 3 – Density Estimation

Q 3.1 Give the density estimate $p_{\mathcal{B}}$ of a random variable X inside a region of interest $\mathcal{B}$ of volume V, based on a number k of observed samples in this region out of n drawn samples.

Q 3.2 Let X be a random variable in $\mathcal{X}$. We want to estimate the density p_X of this variable from a set of observations $\mathcal{X}_o$. Describe how to proceed to make this estimation using the histogram method.

Q 3.3 Discuss kernel density estimation methods.

Exercice 4 (4 points) – Risk

Let the result of a medical test be expressed as a real number $x \in \mathbb{R}$, and let there be two classes, y_- and y_+, representing "not sick" and "sick" respectively. We know that :

$$P(y = y_+|x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

We use a classifier of the form :

$$f_\theta(x) = \begin{cases} y_+ & \text{if } x > \theta, \\ y_- & \text{if } x \leq \theta. \end{cases}$$

Q 4.1 Give the expression for the 0-1 loss and the risk at a point x^0 as a function of θ.

Q 4.2 Suppose that the result of the test x is uniformly distributed in $[-1, 1]$. What is the optimal classifier ? What is the minimum risk value ?

Q 4.3 It is known that it is three times more costly to classify a sick person as not sick than the reverse. Give the associated cost function, the new risk formulation, and the optimal classifier.

Q 4.4 Could we do better with a Bayesian classifier ? What about a naive Bayesian classifier ?