Universidad de la República
Facultad de Ingeniería

Practice: Binary Classification

1. From the Bayes Classifier, predict the class for each test data and compute the error.

Training sample								
x_1	a	a	b	a	a	b	b	b
x_2	b	a	a	a	a	b	b	b
y	1	1	1	1	-1	-1	-1	-1

Test sample				
x_1	a	a	b	b
x_2	a	b	a	b
prediction	?	?	?	?
real	1	-1	1	1

2. We have historical data on whether or not a tennis match has been played with a series of very simple meteorological parameters.

Sky	Temp	Humidity	Wind	Game
Sunny	Warm	High	Yes	No
Sunny	Warm	High	No	Yes
Sunny	Tempered	Low	Yes	Yes
Cloudy	Cold	Low	No	Yes
Cloudy	Warm	Normal	No	Yes
Cloudy	Cold	High	Yes	No
Rainy	Temperd	High	No	No

Using Naive Bayes, calculate the probability that there will be a match on a cloudy, cold day, with high humidity and no wind.

3. Suppose that $\pi_1 = \pi_0 = 0{,}5$ and the densities are $g_1 = \mathcal{N}(0,1)$ and $g_0 = 0{,}7\mathcal{N}(0,1) + 0{,}3\mathcal{N}(-1,2)$.

   a) Assuming equal cost find:

      1) Plot the densities and write the Bayes rule for this classification task.

      2) Write the Bayes decision boundary and find its solutions.

   b) Assume that $C(1,0) = 2$ and $C(0,1) = 6$. Repeat questions above.

4. Generate 100 observations from a bivariate Gaussian distribution $\mathcal{N}(\mu_1, \Sigma_1)$ with $\mu_1 = (3,1)'$ and $\Sigma_1 = I$ (identity matrix and label them as 1. Generate another 100 observations from a bivariate Gaussian distribution $\mathcal{N}(\mu_2, \Sigma_2)$ with $\mu_2 = (1,3)'$ and $\Sigma_2 = I$ and label them as 0. Together, these 200 observations constitute the training set.

   a) Write an R code to generate this data set.

   b) Plot this data using different colors for the two classes.

   c) Assuming that priors are equals, find the Bayes Classifier.

   d) Compute the training error.

e) Train a linear regression model, using the function lm(y $\sim$ x), with the training set.

f) Plot the boundary decision of Bayes Classifier and the line obtained by the linear regression model.

g) Generate a test set of 50 observations and compute the test error of Bayes Classifier and the linear model.

5. Consider the following table.

x_1	x_2	x_3	y
0	3	0	Red
2	0	0	Red
0	1	3	Red
0	1	2	Green
-1	0	1	Green
1	1	1	Red

a) With the euclidean distance, what is the prediction with $k = 1$ and with $k = 3$ for the test observation $(0, 0, 0)$?

b) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for k to be large or small? Why?