Aprendizagem 2022

Homework II – Group 019

Diogo Gaspar 99207, Rafael Oliveira 99311

Part I: Pen and paper

1. **Compute the recall of a distance-weighted kNN with $k = 5$ and distance $d(x_1, x_2) = \text{Hamming}(x_1, x_2) + \frac{1}{2}$ using leave-one-out evaluation schema (i.e., when classifying one observation, use all remaining ones).**

   For starters, it's worth noting that, in this context, the **Hamming distance** between two observations x_1 and x_2 is defined as the number of attributes that differ between them.

   Knowing this, we can now create an 8×8 matrix (as can be seen below), where each entry represents the Hamming distance ($+\frac{1}{2}$) between two observations. This matrix is symmetric, of course. Each column i, here, will have $8 - 1 = 7$ associated entries, each representing the distance d between the observation x_i and the remaining 7 observations: we will, then, pick the $k = 5$ nearest neighbors according to said distance, classifying x_i as the class that appears most frequently among them.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	×	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$
x_2	$\frac{5}{2}$	×	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{1}{2}$
x_3	$\frac{3}{2}$	$\frac{3}{2}$	×	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{5}{2}$	$\frac{1}{2}$	$\frac{3}{2}$
x_4	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	×	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$
x_5	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	×	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{3}{2}$
x_6	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	×	$\frac{5}{2}$	$\frac{3}{2}$
x_7	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{5}{2}$	×	$\frac{3}{2}$
x_8	$\frac{5}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	×

Table 1: Distance d between observations - in teal, a given observation's k nearest neighbors

We'll have, given the data gathered above, the following confusion matrix:

Projected \ Real	P	N	
P	1	3	
N	3	1	

Figure 1: Confusion Matrix

Moreover, the **recall** of a classifier is defined as the ratio between the number of true positives and the number of true positives plus the number of false negatives that the classifier makes. Looking at the confusion matrix above, we can assert that the associated recall will, therefore, be:

$$\frac{TP}{TP + FN} = \frac{1}{1+3} = \frac{1}{4} = 0.25$$

2. **Considering the nine training observations, learn a Bayesian classifier assuming: i)** y_1 **and** y_2 **are dependent, ii)** $\{y_1, y_2\}$ **and** $\{y_3\}$ **variable sets are independent and equally important, and iii)** y_3 **is normally distributed. Show all parameters.**

   Considering both variable sets, $\{y_1, y_2\}$ and $\{y_3\}$, to be independent and equally important, it'll make sense to train a Naive Bayes classifier here, such that (and utilizing the Bayes' theorem):

$$P(C =^N_P | y_1, y_2, y_3) = \frac{P(y_1, y_2, y_3 | C =^N_P) P(C =^N_P)}{P(y_1, y_2, y_3)}$$

   More so, since $\{y_1, y_2\}$ and $\{y_3\}$ are independent (and y_1 and y_2 are dependent), we can rewrite the above as:

$$P(C =^N_P | y_1, y_2, y_3) = \frac{P(y_1, y_2 | C =^N_P) P(y_3 | C =^N_P) P(C =^N_P)}{P(y_1, y_2) P(y_3)}$$

   According to the Naive Bayes' assumption, "the presence (or absence) of a particular feature in a class is unrelated to the presence (or absence) of any other feature", so we will, in fact, only need the above equation's numerator to be able to classify a new observation. The goal here is to find:

$$\text{argmax}_{c \in \{N,P\}} P(y_1, y_2 | C = c) P(y_3 | C = c) P(C = c)$$

   For starters, we can note that, from the given training set:

$$P(C = P) = \frac{5}{9}, \quad P(C = N) = \frac{4}{9}$$

   We also know that y_3 is normally distributed, meaning we'll have:

$$P(y_3 | C =^N_P) \sim \mathcal{N}(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

By also looking at the given training set (which includes the new, ninth sample), we'll be able to extrapolate the following probabilities:

$C = N : P(C = N) = \frac{4}{9}$

- $P(y_2 = A, y_2 = 0 | C = N) = 0$
- $P(y_2 = A, y_2 = 1 | C = N) = \frac{1}{4}$
- $P(y_1 = B, y_2 = 0 | C = N) = \frac{2}{4} = \frac{1}{2}$
- $P(y_1 = B, y_2 = 1 | C = N) = \frac{1}{4}$

Regarding y_3 and N labeled observations, we'll have the following parameters:

$$\mu = \frac{1 + 0.9 + 1.2 + 0.8}{4} = 0.975$$

$$\sigma^2 = \frac{1}{4 - 1} \sum_{i=1}^{4} (y_{3,i} - \mu)^2 = 0.029$$

$C = P : P(C = P) = \frac{5}{9}$

- $P(y_1 = A, y_2 = 0 | C = P) = \frac{2}{5}$
- $P(y_2 = A, y_2 = 1 | C = P) = \frac{1}{5}$
- $P(y_1 = B, y_2 = 0 | C = P) = \frac{1}{5}$
- $P(y_1 = B, y_2 = 1 | C = P) = \frac{1}{5}$

Regarding y_3 and P labeled observations, we'll have the following parameters:

$$\mu = \frac{1.2 + 0.8 + 0.5 + 0.9 + 0.8}{5} = 0.84$$

$$\sigma^2 = \frac{1}{5 - 1} \sum_{i=1}^{5} (y_{3,i} - \mu)^2 = 0.063$$

The model is now ready to be used to classify new observations. Applying the model to the nine training observations, we'll have the following results:

	y_1	y_2	y_3	Class	$P(C = N)P(y_1, y_2, y_3 \mid C = N)$	$P(C = P)P(y_1, y_2, y_3 \mid C = P)$	Predicted Class	Veredict
x_1	A	0	1.2	P	0	$\frac{5}{9} \times \frac{2}{5} \times 0.07575$	P	TP
x_2	B	1	0.8	P	$\frac{4}{9} \times \frac{1}{4} \times 0.84794$	$\frac{5}{9} \times \frac{1}{5} \times 0.56331$	N	FN
x_3	A	1	0.5	P	$\frac{4}{9} \times \frac{1}{4} \times 0.99736$	$\frac{5}{9} \times \frac{1}{5} \times 0.91233$	N	FN
x_4	A	0	0.9	P	0	$\frac{5}{9} \times \frac{2}{5} \times 0.40554$	P	TP
x_5	B	0	1	N	$\frac{4}{9} \times \frac{1}{2} \times 0.44164$	$\frac{5}{9} \times \frac{1}{5} \times 0.26191$	N	TN
x_6	A	0	0.9	N	$\frac{4}{9} \times \frac{1}{2} \times 0.67019$	$\frac{5}{9} \times \frac{1}{5} \times 0.40554$	N	TN
x_7	B	1	1.2	N	$\frac{4}{9} \times \frac{1}{4} \times 0.0932$	$\frac{5}{9} \times \frac{1}{5} \times 0.07575$	N	TN
x_8	B	1	0.8	N	$\frac{4}{9} \times \frac{1}{4} \times 0.84794$	$\frac{5}{9} \times \frac{1}{5} \times 0.56331$	N	TN
x_9	B	0	0.8	P	$\frac{4}{9} \times \frac{1}{2} \times 0.84794$	$\frac{5}{9} \times \frac{1}{5} \times 0.56331$	N	FN

Table 2: Classification results

3. **Under a MAP assumption, compute $P(Positive|x)$ of each testing observation.**

   Under the MAP (*maximum a posteriori*) assumption,

4. **Given a binary class variable, the default decision threshold of** $\theta = 0.5$**,**

$$f(x|\theta) = \begin{cases} Positive & \textbf{if } P(positive|x) > \theta \\ Negative & otherwise \end{cases}$$

can be adjusted. Which decision threshold – 0.3, 0.5 or 0.7 – optimizes testing accuracy?

Part II: Programming

5. **Using `sklearn`, considering a 10-fold stratified cross validation (`random=0`), plot the cumulative testing confusion matrices of kNN (uniform weights, $k = 5$, Euclidean distance) and Naïve Bayes (Gaussian assumption). Use all remaining classifier parameters as default.**

6. **Using `scipy`, test the hypothesis "kNN is statistically superior to Naïve Bayes regarding accuracy", asserting whether is true.**

7. **Enumerate three possible reasons that could underlie the observed differences in predictive accuracy between kNN and Naïve Bayes.**

Appendix