TÉCNICO LISBOA

Aprendizagem 2024/25

Homework II – Group 014

(ist106266, ist106583)

I. Pen-and-paper

1) We begin by computing the distance matrix between all observations:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0	2	1	0	1	1	1	2
x_2	2	0	1	2	1	1	1	0
x_3	1	1	0	1	2	2	0	1
x_4	0	2	1	0	1	1	1	2
x_5	1	1	2	1	0	0	2	1
x_6	1	1	2	1	0	0	2	1
x_7	1	1	0	1	2	2	0	1
x_8	2	0	1	2	1	1	1	0

Now to implement the leave-one-out strategy, for each observation x_i, we find the five smallest distances in row x_i, ignoring $d(x_i, x_i)$, and classify x_i by taking the mode of those neighbour's classes. By defining $NN_5(x_i)$ as the set that contains x_i's five nearest neighbours and $M(x_i)$ as x_i's class according to our classifier M, we then get:

$$NN_5(x_1) = \{x_3, x_4, x_5, x_6, x_7\} \Rightarrow M(x_1) = mode\{P, P, N, N, N\} = N$$
$$NN_5(x_2) = \{x_3, x_5, x_6, x_7, x_8\} \Rightarrow M(x_2) = mode\{P, N, N, N, N\} = N$$
$$NN_5(x_3) = \{x_1, x_2, x_4, x_7, x_8\} \Rightarrow M(x_3) = mode\{P, P, P, N, N\} = P$$
$$NN_5(x_4) = \{x_1, x_3, x_5, x_6, x_7\} \Rightarrow M(x_4) = mode\{P, P, N, N, N\} = N$$
$$NN_5(x_5) = \{x_1, x_2, x_4, x_6, x_8\} \Rightarrow M(x_5) = mode\{P, P, P, N, N\} = P$$
$$NN_5(x_6) = \{x_1, x_2, x_4, x_5, x_8\} \Rightarrow M(x_6) = mode\{P, P, P, N, N\} = P$$
$$NN_5(x_7) = \{x_1, x_2, x_3, x_4, x_8\} \Rightarrow M(x_7) = mode\{P, P, P, P, N\} = P$$
$$NN_5(x_8) = \{x_2, x_3, x_5, x_6, x_7\} \Rightarrow M(x_8) = mode\{P, P, N, N, N\} = N$$

It is then possible to summarise these results in the following confusion matrix:

$$\begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$$

Where the first and second columns denote the real classifications of P and N, respectively, and the first and second rows denote the predicted classifications of P and N, respectively. All that is left is to calculate the F1-measure:

$$precision = \frac{TP}{TP+FP} = \frac{1}{4} \qquad recall = \frac{TP}{TP+FN} = \frac{1}{4} \qquad F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot \frac{1}{4} \cdot \frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{1}{4}$$

TÉCNICO LISBOA

Aprendizagem 2024/25

Homework II – Group 014

(ist106266, ist106583)

2) By examining the given observations, a clear pattern emerges: when the first feature is 'A', the majority of the observations are classified as positive (P) and when it is 'B' they tend to be negative (N). On the other hand, the second feature (0 or 1) shows no clear relationship with the class labels, as it appears to be equally distributed between positive and negative observations. This indicates that the second feature doesn't provide useful information for classifying the observations. Therefore a weighted Hamming distance might be more suitable:

$$d(x,\ y)\ =\ 1\ \cdot \left(x_1 \neq y_1\right) + \ 0\ \cdot \left(x_2 \neq y_2\right)$$

Here, $x_i \neq\ y_i$ evaluates to 1 if the features differ and 0 otherwise. This metric is non-negative since it is the sum of two non-negative terms. The distance of any point to itself is zero:

$$d(x, x)\ =\ 1\ \cdot \left(x_1 \neq\ x_1\right)\ +\ 0\ \cdot \left(x_2 \neq\ x_2\right)\ =\ 1\ \cdot\ 0\ =\ 0$$

It is also symmetric:

$$d(x,\ y)\ =\ 1\ \cdot \left(x_1 \neq y_1\right) + \ 0\ \cdot \left(x_2 \neq y_2\right)\ =\ d(y,\ x)$$

And it also fulfills the triangular inequality, thus gathering all necessary properties of a valid metric:

$$\forall x, y, z\ \in\ V,\ d(x,\ y) \leq\ d(x,\ z)\ +\ d(z,\ y)\ \Leftrightarrow\ \left(x_1 \neq y_1\right)\ \leq\ \left(x_1 \neq z_1\right)\ +\left(z_1 \neq y_1\right)$$

Case 1, only one of the features is unique. Without loss of generality if $x_1 \neq y_1$ and $y_1 = z_1$:

$$d(x,\ y) = 1\ \wedge\ d(x,\ z) = 1\ \wedge\ d(z,\ y) = 0 \Rightarrow 1 \leq 1\ +\ 0$$

Case 2, all features are the same:

$$d(x,\ y) = 0\ \wedge\ d(x,\ z) = 0\ \wedge\ d(z,\ y) = 0 \Rightarrow 0 \leq 0\ +\ 0$$

Now, using this new distance metric, we recompute the distance matrix between all observations and apply the leave-one-out evaluation as we did before:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0	1	0	0	1	1	0	1
x_2	1	0	1	1	0	0	1	0
x_3	0	1	0	0	1	1	0	1
x_4	0	1	0	0	1	1	0	1
x_5	1	0	1	1	0	0	1	0
x_6	1	0	1	1	0	0	1	0
x_7	0	1	0	0	1	1	0	1
x_8	1	0	1	1	0	0	1	0

By combining this metric and selecting only the three nearest neighbours, it is then possible to increase the F1 measure by three fold.

$$NN_3(x_1)\ =\left\{x_3,\ x_4,\ x_7\right\} \Rightarrow M(x_1)\ =\ mode\{P,\ P,\ N\} =\ P$$
$$NN_3(x_2)\ =\left\{x_5,\ x_6,\ x_8\right\} \Rightarrow M(x_2)\ =\ mode\{N,\ N,\ N\} =\ N$$
$$NN_3(x_3)\ =\left\{x_1,\ x_4,\ x_7\right\} \Rightarrow M(x_3)\ =\ mode\{P,\ P,\ N\} =\ P$$
$$NN_3(x_4)\ =\left\{x_1,\ x_3,\ x_7\right\} \Rightarrow M(x_4)\ =\ mode\{P,\ P,\ N\} =\ P$$
$$NN_3(x_5)\ =\left\{x_2,\ x_6,\ x_8\right\} \Rightarrow M(x_5)\ =\ mode\{P,\ N,\ N\} =\ N$$

$$NN_3(x_6) = \{x_2, x_5, x_8\} \Rightarrow M(x_6) = mode\{P, N, N\} = N$$
$$NN_3(x_7) = \{x_1, x_3, x_4\} \Rightarrow M(x_7) = mode\{P, P, P\} = P$$
$$NN_3(x_8) = \{x_2, x_5, x_6\} \Rightarrow M(x_8) = mode\{P, N, N\} = N$$

It is then possible to summarise these results in the following confusion matrix:

$$\begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

Where the first and second columns denote the real classifications of P and N, respectively, and the first and second rows denote the predicted classifications of P and N, respectively. All that is left is to calculate the F1-measure:

$$precision = \frac{TP}{TP+FP} = \frac{3}{4} \qquad recall = \frac{TP}{TP+FN} = \frac{3}{4} \qquad F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot \frac{3}{4} \cdot \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{3}{4}$$

As expected, the metric and number of neighbours proposed improves the performance by three-fold, increasing the value from $\frac{1}{4}$ to $\frac{3}{4}$.

3) Since y_1 and y_2 are independent from y_3, the naïve Bayes classifier only needs to learn eleven parameters. Only one of the priors needs to be learned, the other can be easily deduced from the first:

$$p(z = P) = \frac{5}{9} \Rightarrow p(z = N) = 1 - p(z = P) = \frac{4}{9}$$

Afterwards, the likelihoods for the conjugation of y_1 and y_2 must be computed:

$$p(y_1 = A, y_2 = 0 | z = P) = \frac{2}{5}$$
$$p(y_1 = A, y_2 = 1 | z = P) = \frac{1}{5}$$
$$p(y_1 = B, y_2 = 0 | z = P) = \frac{1}{5}$$
$$p(y_1 = A, y_2 = 0 | z = N) = 0$$
$$p(y_1 = A, y_2 = 1 | z = N) = \frac{1}{4}$$
$$p(y_1 = B, y_2 = 0 | z = N) = \frac{2}{4}$$

Much like with the priors, only six out of the eight likelihoods needed must be learned, since from the results it is possible to infer that:

$$p(y_1 = B, y_2 = 1 | z = P) = 1 - (\frac{2}{5} + \frac{1}{5} + \frac{1}{5}) = \frac{1}{5}$$
$$p(y_1 = B, y_2 = 1 | z = N) = 1 - (0 + \frac{1}{4} + \frac{2}{4}) = \frac{1}{4}$$

Finally, the model must estimate the parameters of y_3's class conditional normal distributions:

$$\widehat{\mu_{z=P}} = \frac{1}{5}(1.1 + 0.8 + 0.5 + 0.9 + 0.8) = 0.82$$
$$\widehat{\sigma_{z=P}} = \sqrt{\frac{1}{4}\left((1.1 - 0.82)^2 + 2(0.8 - 0.82)^2 + (0.5 - 0.82)^2 + (0.9 - 0.82)^2\right)} = 0.2168$$
$$\widehat{\mu_{z=N}} = \frac{1}{4}(1 + 0.9 + 1.2 + 0.9) = 1$$
$$\widehat{\sigma_{z=N}} = \sqrt{\frac{1}{3}\left((1 - 1))^2 + 2(0.9 - 1)^2 + (1.2 - 1)^2\right)} = 0.1414$$

TÉCNICO LISBOA

Aprendizagem 2024/25

Homework II – Group 014

(ist106266, ist106583)

4) To classify a new observation $x_{new}=(y_{1,new}, y_{2,new}, y_{3,new})$ using the naïve Bayes classifier, under a MAP assumption, it is sufficient to consider the following:

$$h_{MAP} = argmax\{p(z=h)p(y_1=y_{1,new}, y_2=y_{2,new}|z=h)p(y_3=y_{3,new}|z=h)\},\ h\in\{P,N\}$$

The classification of $x_{1,new}=(A, 1, 0.8)$ is as described:

$$p(z=P|x_{1,new}) = p(z=P)p(y_1=A, y_2=1|z=P)p(y_3=0.8|z=P) =$$
$$= \frac{5}{9}\cdot\frac{1}{5}\cdot N(0.8|\mu=0.82, \sigma=0.2168) = 0.2036$$
$$p(z=N|x_{1,new}) = p(z=N)p(y_1=A, y_2=1|z=N)p(y_3=0.8|z=N) =$$
$$= \frac{4}{9}\cdot\frac{1}{4}\cdot N(0.8|\mu=1, \sigma=0.1414) = 0.1153$$

Since $p(z=P|x_{1,new}) > p(z=N|x_{1,new})$, $x_{1,new}$ is classified as P. Now for $x_{2,new}=(B, 1, 1)$:

$$p(z=P|x_{1,new}) = p(z=P)p(y_1=B, y_2=1|z=P)p(y_3=1|z=P) =$$
$$= \frac{5}{9}\cdot\frac{1}{5}\cdot N(1|\mu=0.82, \sigma=0.2168) = 0.14486$$
$$p(z=N|x_{1,new}) = p(z=N)p(y_1=B, y_2=1|z=N)p(y_3=1|z=N) =$$
$$= \frac{4}{9}\cdot\frac{1}{4}\cdot N(1|\mu=1, \sigma=0.1414) = 0.31349$$

Because $p(z=N|x_{2,new}) > p(z=P|x_{2,new})$, $x_{2,new}$ is classified as N. Finally, for $x_{3,new}=(B, 0, 0.9)$:

$$p(z=P|x_{1,new}) = p(z=P)p(y_1=B, y_2=0|z=P)p(y_3=0.9|z=P) =$$
$$= \frac{5}{9}\cdot\frac{1}{5}\cdot N(0.9|\mu=0.82, \sigma=0.2168) = 0.19100$$
$$p(z=N|x_{1,new}) = p(z=N)p(y_1=B, y_2=0|z=N)p(y_3=0.9|z=N) =$$
$$= \frac{4}{9}\cdot\frac{2}{4}\cdot N(0.9|\mu=1, \sigma=0.1414) = 0.24411$$

Seeing as $p(z=N|x_{3,new}) > p(z=P|x_{3,new})$, $x_{3,new}$ is classified as N.

5) Classifying the sentence x=*"I like to run"*, under a ML assumption and considering that there are eight total unique words in the whole vocabulary, five total words in the P class and four total words in the N class, goes as follows:

$$p(x|P) = p("I"|P)p("like"|P)p("to"|P)p("run"|P) = \frac{1+1}{5+8}\cdot\frac{1+1}{5+8}\cdot\frac{0+1}{5+8}\cdot\frac{1+1}{5+8} = 0.0002801$$
$$p(x|N) = p("I"|N)p("like"|N)p("to"|N)p("run"|N) = \frac{0+1}{4+8}\cdot\frac{0+1}{4+8}\cdot\frac{0+1}{4+8}\cdot\frac{1+1}{4+8} = 0.0000965$$

Since p(x|P) is greater than p(x|N), *"I like to run"* should be classified as P. Note that both equations were obtained using the given likelihood calculation rule:

$$p(t_i|c) = \frac{freq(t_i)+1}{N_c+V}$$

Where t_i is a given term i, $freq(t_i)$ is the frequency of the term t_i within the class, V is the number of unique terms in the vocabulary and N_c is the total number of terms in class c.

Homework II – Group 014

(ist106266, ist106583)

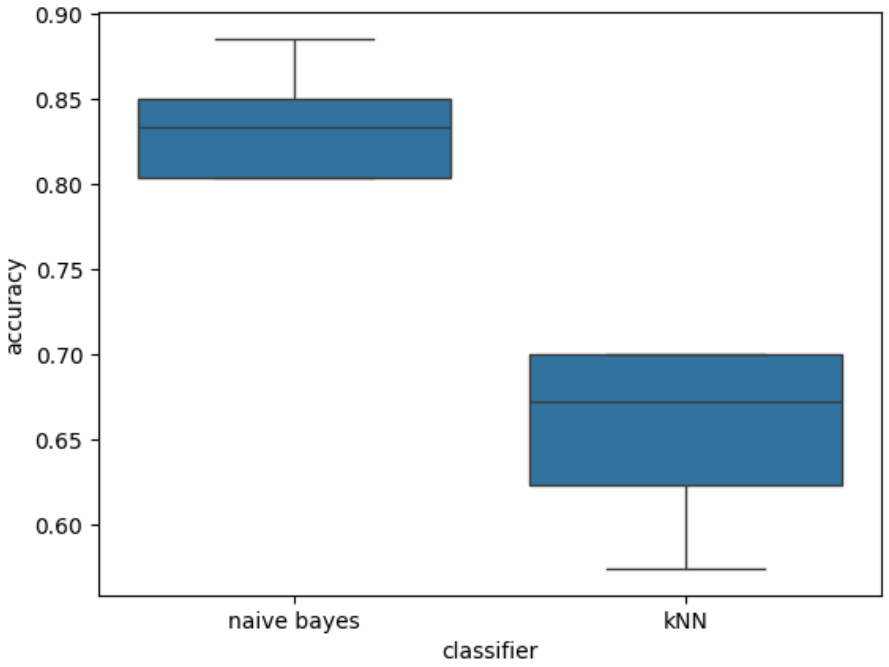
II. Programming and critical analysis

1)

a) As can be observed in the graph below, the naïve Bayes classifier seems to be more stable than its kNN counterpart since the latter's boxplot spans a bigger range than the former's.

The naïve Bayes classifier also seems to have a consistently better accuracy overall. This may be due to the fact that the kNN algorithm is distance-based and there are many different variables in the provided data set, each with their own range of possible values like age and sex, where the first can take values mostly between forty to sixty while the second can only be either zero or one. Considering that the default metric for *Sklearn's* kNeighboursClassifier is the euclidean distance, variables like age will have a much higher influence when compared to others like sex, that take smaller values.

This does not affect the naïve Bayes classifier because it is not distance based like the kNN classifier.



b) After making use of the Min-Max scaler, as suggested, the naïve Bayes model seems to have stayed as accurate as before, with an average accuracy of 84% over all five folds, while the accuracy of the kNN classifier greatly improved, increasing it's average accuracy from around 65% up to 82%.

This seems to corroborate the aforementioned theory: the kNN classifier became much more accurate now that every variable is able to contribute equally. Before, variables like age and sex had wildly different value ranges, making age, which took values from around forty up to sixty, weigh much more than sex, which could only be zero or one. However, after scaling, all variables have the same range, taking values from zero up to one, thus fixing the issue.

The naïve Bayes classifier remains unchanged since it is not distance based, meaning the previous problem did not affect it, nor does the scaling.

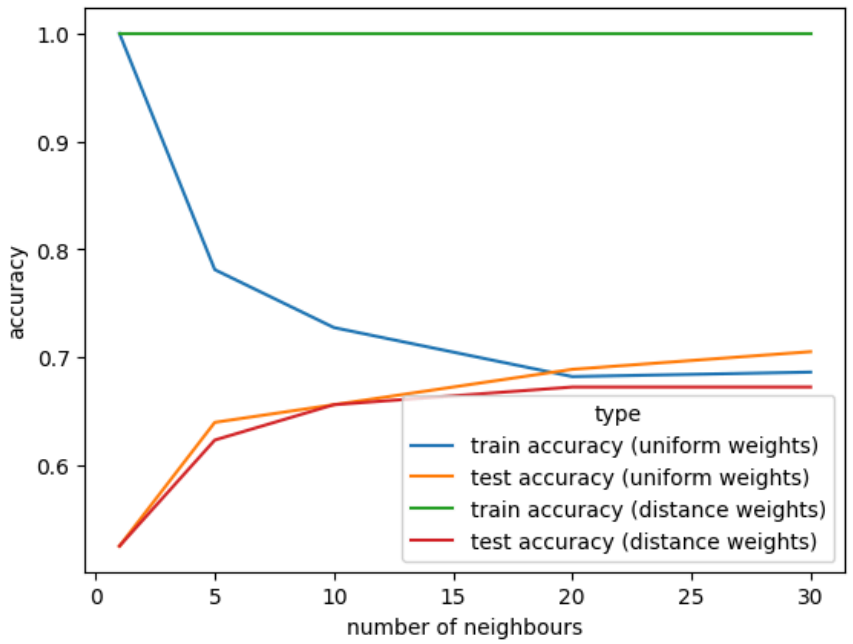
Homework II – Group 014

(ist106266, ist106583)

c) By executing a paired t-test between the accuracies of the folds for each of the trained classifiers, to test the null hypothesis $accuracy_{kNN} \leq accuracy_{naïve\ bayes}$, it is possible to determine that the kNN model is not statistically superior to naïve Bayes regarding accuracy since the t-test's p-value is about 0.746, meaning there is no reason to reject the null hypothesis for the usual significance levels.

2)

a)



b) For the most part, increasing the neighbours has the effect of increasing the generalization ability of the two models.

When considering only one neighbour, the training accuracy is perfect, since the observations will only consider themselves. However, that also makes both models prone to overfitting since there are very few neighbours to take into account when predicting new observations.

By increasing the number of neighbours, both accuracies of the uniform-weighted kNN classifier converge to nearly the same value, which shows its generalization capabilities increase. The distance-weighted kNN classifier's training accuracy and generalization capabilities also increase with the number of neighbours considered. It is important to note that the distance-weighted model's training accuracy stays at 100% since the weight of an observation with itself is one, making it much more likely to be assigned the correct class.

3) When considering the given data set, the naïve Bayes model might run into some issues such as dependencies between variables. Analyzing the features, it is clear that some, like age and cholesterol, share dependencies between each other. This is troublesome for the model, since it assumes independence between features, thus lowering its accuracy.

Another possible cause of error might be using the normal distribution for every single feature, especially when they are discrete. In this case, the model could benefit from using different distributions depending on the feature. For instance, it would be more suitable to model the feature sex using a Bernoulli distribution, considering it can only take two values.

These two problems combined may make the model less accurate than it could have been otherwise.

END