Homework I – Group 017

(ist1106992, ist1106154)

I. Pen-and-paper

1.

	y_1	y_2	
x_1	A	0	P
x_2	B	1	P
x_3	A	1	P
x_4	A	0	P
x_5	B	0	N
x_6	B	0	N
x_7	A	1	N
x_8	B	1	N

$$\text{F1-score} = \frac{2PR}{P+R}$$

$$P = \frac{TP}{TP+FP} \qquad R = \frac{TP}{TP+FN}$$

Hamming Distance c/ 5NN

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

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Target	P	P	P	P	N	N	N	N
Predicted	N	N	P	N	P	P	P	N

Predicted \ Target	P	N
P	1	3
N	3	1

$$P = \frac{1}{1+3} = \frac{1}{4} \qquad R = \frac{1}{1+3} = \frac{1}{4}$$

$$F1\text{-score} = \frac{2 \times \frac{1}{4} \times \frac{1}{4}}{\frac{1}{4} + \frac{1}{4}} = \frac{\frac{1}{8}}{\frac{1}{2}} = 0{,}25$$

2.

Using 3NN and weighted distances:

$$d(u_i, u_j) = \alpha\, d_{y_1}(u_i, u_j) + \beta\, d_{y_2}(u_i, u_j)$$

For $\alpha = 0$ and $\beta = 1$

$$d(u_i, u_j) = d_{y_2}(u_i, u_j) = \begin{cases} 0 & \text{if } y_1\, u_i = y_2\, u_j \\ 1 & \text{else} \end{cases}$$

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
u_1	—	1	0	0	1	1	0	1
u_2	1	—	1	1	0	0	1	0
u_3	0	1	—	0	1	1	0	1
u_4	0	1	0	—	1	1	0	1
u_5	1	0	1	1	—	0	1	0
u_6	1	0	1	1	0	—	1	0
u_7	0	1	0	0	1	1	—	1
u_8	1	0	1	1	0	0	1	—

x_1 : mode (x_3 , x_4 , x_7) = mode (P, P, N) = P

x_2 : mode (x_5 , x_6 , x_8) = mode (N, N, N) = N

x_3 : mode (x_1 , x_4 , x_7) = mode (P, P, N) = P

x_4 : mode (x_1 , x_3 , x_7) = mode (P, P, N) = P

x_5 : mode (x_2 , x_6 , x_8) = mode (P, N, N) = N

x_6 : mode (x_2 , x_5 , x_8) = mode (P, N, N) = N

x_7 : mode (x_1 , x_3 , x_4) = mode (P, P, P) = P

x_8 : mode (x_2 , x_5 , x_6) = mode (P, N, N) = N

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Target	P	P	P	P	N	N	N	N
Predicted	P	N	P	P	N	N	P	N

Target

Predicted	P	N
P	3	1
N	1	3

$P = \frac{3}{1+3} = \frac{3}{4}$ $\quad R = \frac{3}{1+3} = \frac{3}{4}$

$$F1\text{-score} = \frac{2 \times \frac{3}{4} \times \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{3}{4} = 0{,}75$$

3.

	y_1	y_2	y_3	Target
x_1	A	0	1,1	P
x_2	B	1	0,8	P
x_3	A	1	0,5	P
x_4	A	0	0,9	P
x_5	B	0	1	N
x_6	B	0	0,9	N
x_7	A	1	1,2	N
x_8	B	1	0,9	N
x_9	B	0	0,8	P

Bayes Rule : $P(C \mid y_1, y_2, y_3) = \dfrac{P(C) \times P(y_1, y_2 \mid C) \times P(y_3 \mid C)}{P(y_1, y_2, y_3)}$

$(C = P \vee C = N)$

$P(y_1, y_2 \mid C)$:

$C = P$:

$P(A, 0 \mid P) = \frac{2}{5}$

$P(A, 1 \mid P) = \frac{1}{5}$

$P(B, 0 \mid P) = \frac{1}{5}$

$P(B, 1 \mid P) = \frac{1}{5}$

$C = N$:

$P(A, 0 \mid N) = 0$

$P(A, 1 \mid N) = \frac{1}{4}$

$P(B, 0 \mid N) = \frac{1}{2}$

$P(B, 1 \mid N) = \frac{1}{4}$

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

$P(y_3 \mid C)$: assuming normal distribution

$$P(y_3 \mid C) = N(y_3 \mid \mu, \sigma) = \frac{e^{-\frac{(u-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}$$

$C = P$:

$$\mu = \frac{1{,}1 + 0{,}8 + 0{,}5 + 0{,}9 + 0{,}8}{5} = 0{,}82$$

$$\sigma = \sqrt{\frac{1}{5-1} \times \left((1{,}1-0{,}82)^2 + (0{,}8-0{,}82)^2 \times 2 + (0{,}5-0{,}82)^2 + (0{,}9-0{,}82)^2 \right)}$$

$= 0{,}217$

$$P(y_3 \mid C=P) = N(y_3 \mid 0{,}82 \,;\, 0{,}217)$$

$C = N$:

$$\mu = \frac{1 + 0{,}9 + 1{,}2 + 0{,}9}{4} = 1$$

$$\sigma = \sqrt{\frac{1}{4-1} \times \left((1-0{,}975)^2 + (0{,}9-0{,}975)^2 + (1{,}2-0{,}975)^2 + (0{,}9-0{,}975)^2 \right)}$$

$= 0{,}144$

$$P(y_3 \mid C=N) = N(y_3 \mid 1 \,;\, 0{,}144)$$

$P(y_1, y_2, y_3)$ calculations are superfluous because it's a constant

4.

$$P(P \mid A;1;0{,}8) = \frac{P(P) \times P(A,1 \mid P) \times P(y_3 = 0{,}8 \mid \mu = 0{,}82,\ \sigma = 0{,}217)}{P(A;1,0{,}1)}$$

$$= \frac{\frac{5}{9} \times \frac{1}{5} \times 1{,}83}{P(A;1;0{,}8)} = \frac{0{,}203}{P(A,1,0{,}8)}$$

$$P(N \mid A;1;0{,}8) = \frac{P(N) \times P(A,1 \mid N) \times P(y_3 = 0{,}8 \mid \mu = 1;\ \sigma = 0{,}144)}{P(A,1;0{,}1)}$$

$$= \frac{\frac{4}{9} \times \frac{1}{4} \times 1{,}03}{P(A;1;0{,}8)} = \frac{0{,}114}{P(A,1,0{,}8)}$$

$P(P \mid A;1;0{,}8) > P(N \mid A;1;0{,}8)$ so $(A;1;0{,}8)$ classifies as P

$$P(P \mid B,1,1) = \frac{P(P) \times P(B,1 \mid P) \times P(y_3 = 1 \mid \mu = 0{,}82,\ \sigma = 0{,}217)}{P(B,1,1)}$$

$$= \frac{\frac{5}{9} \times \frac{1}{5} \times 1{,}30}{P(B,1,1)} = \frac{0{,}144}{P(B,1,1)}$$

$$P(N \mid B,1,1) = \frac{P(N) \times P(B,1 \mid N) \times P(y_3 = 1 \mid \mu = 1,\ \sigma = 0{,}144)}{P(B,1,1)}$$

$$= \frac{\frac{4}{9} \times \frac{1}{4} \times 2{,}85}{P(B,1,1)} = \frac{0{,}317}{P(B,1,1)}$$

$P(N \mid B,1,1) > P(P \mid B,1,1)$ so $(B,1,1)$ classifies as N

$$P(P \mid B;0;0{,}9) = \frac{P(P) \times P(B,0 \mid P) \times P(y_3 = 0{,}9 \mid \mu = 0{,}82,\ \sigma = 0{,}217)}{P(B;0;0{,}9)}$$

$$= \frac{\frac{5}{9} \times \frac{1}{5} \times 1{,}72}{P(B;0;0{,}9)} = \frac{0{,}191}{P(B;0,0{,}9)}$$

$$P(N \mid B;0;0{,}9) = \frac{P(N) \times P(B,0 \mid N) \times P(y_3 = 0{,}9 \mid \mu = 1,\ \sigma = 0{,}144)}{P(B;0;0{,}9)}$$

$$= \frac{\frac{4}{9} \times \frac{1}{2} \times 2{,}21}{P(B;0,0{,}9)} = \frac{0{,}491}{P(B;0,0{,}9)}$$

$P(N \mid B;0;0{,}9) > P(P \mid B;0;0{,}9)$ so $(B;0,0{,}9)$ classifies as N

5.

$$P(t_i \mid c) = \frac{freq(t_i) + 1}{N_c + V} \qquad t = \text{"I like to run"}$$

$$P(N) = P(P) = \frac{1}{2} \qquad N_P = 5 \quad N_N = 4 \quad V = 8$$

Para $C = P$: independent because of naive Bayes

$$P(t \mid P) = P(\text{"I"} \mid P) \times P(\text{"like"} \mid P) \times P(\text{"to"} \mid P) \times P(\text{"run"} \mid P) =$$

$$= \frac{1+1}{5+8} \times \frac{1+1}{5+8} \times \frac{0+1}{5+8} \times \frac{1+1}{5+8} = \frac{2}{13} \times \frac{2}{13} \times \frac{1}{13} \times \frac{2}{13} = 2,8 \times 10^{-4}$$

$$P(P \mid t) = \frac{\frac{1}{2} \times 2,8 \times 10^{-4}}{P(t)} = \frac{1,4 \times 10^{-4}}{P(t)}$$

Para $C = N$:

$$P(t \mid C) = P(\text{"I"} \mid C) \times P(\text{"like"} \mid C) \times P(\text{"to"} \mid C) \times P(\text{"run"} \mid C) =$$

$$= \frac{0+1}{4+8} \times \frac{0+1}{4+8} \times \frac{0+1}{4+8} \times \frac{1+1}{4+8} = \frac{1}{12} \times \frac{1}{12} \times \frac{1}{12} \times \frac{2}{12} = 9,65 \times 10^{-5}$$

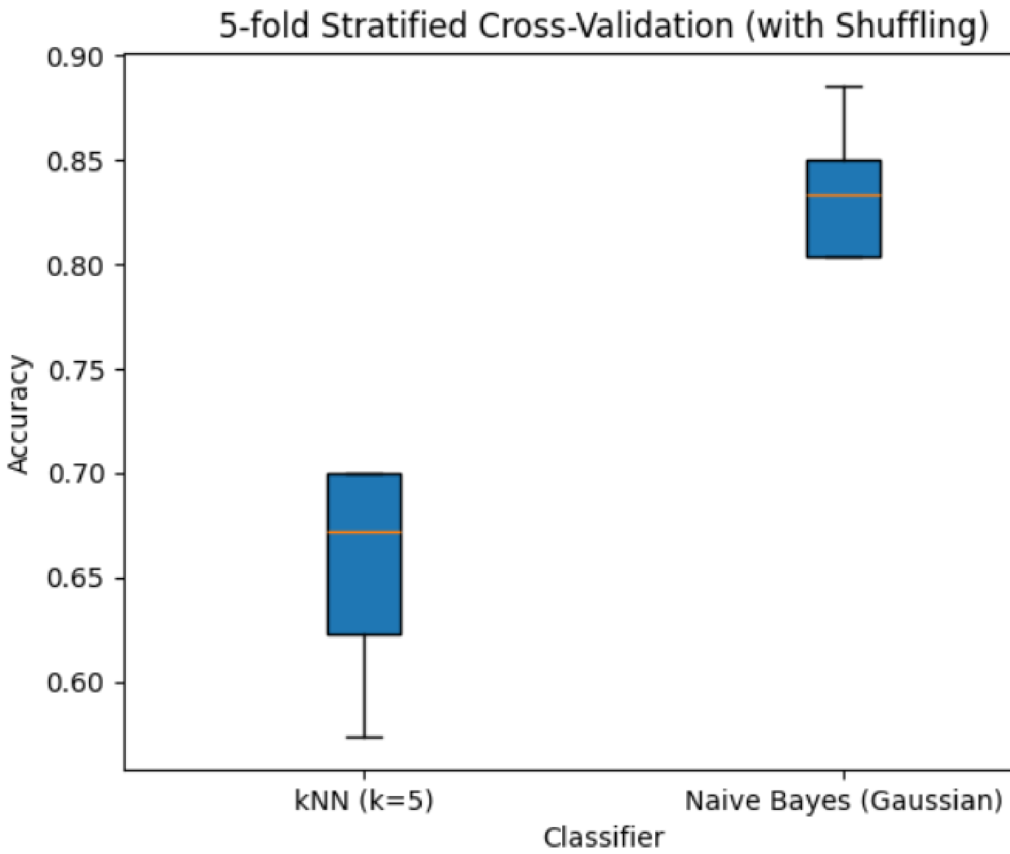
$$P(C \mid t) = \frac{\frac{1}{2} \times 9,65 \times 10^{-5}}{P(t)} = \frac{4,83 \times 10^{-5}}{P(t)}$$

$P(P \mid t) > P(C \mid t)$ so "I like to run" is classified as P.

II. Programming and critical analysis

1.

a.



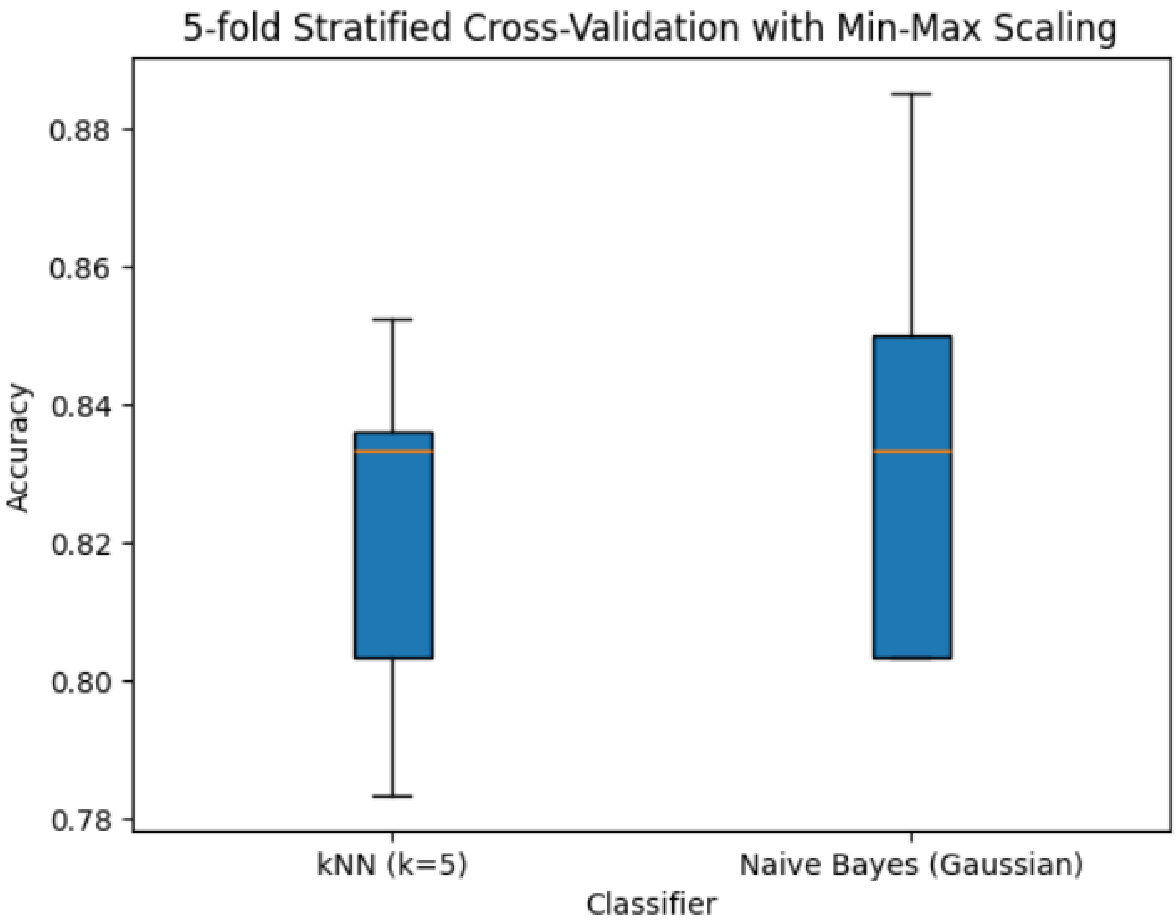
The boxplot above shows that the Naive Bayes classifier has a smaller interquartile range and doesn't show such extreme outliers when compared to the kNN classifier. On the other hand, the kNN classifier exhibits a larger variance in performance, as indicated by the wider range of accuracy values.

Naive Bayes is by definition more stable because it assumes the independence of features and uses probability distributions. This tends to result in a consistent performance and less susceptible to data variations, even with different train-test splits.

For kNN, its stability is easily influenced by the choice of the k parameter and the nature of the data, as the classifier's sensitivity to slight changes in the data is high, because it uses directly the distances between neighboring data points to classify.

In conclusion, the plot's results show that the Naive Bayes classifier is more stable than kNN because of its probabilistic nature and smaller performance variability.

b.



The 5NN model has a mean accuracy of 0.8217 while Naive Bayes' mean accuracy is 0.8350.
The Min-Max Scaling is used to normalize the range of features in a dataset.

This is especially helpful for the kNN model, because it relies on the Euclidean distance between points. When the features are not scaled, attributes with larger numerical ranges dominate the distance calculations, making the kNN classifier biased towards certain features. The scaling ensures that each feature contributes equally to the distance computation, leading to more meaningful neighborhoods and potentially improved classification accuracy, as shown in the generated plot.

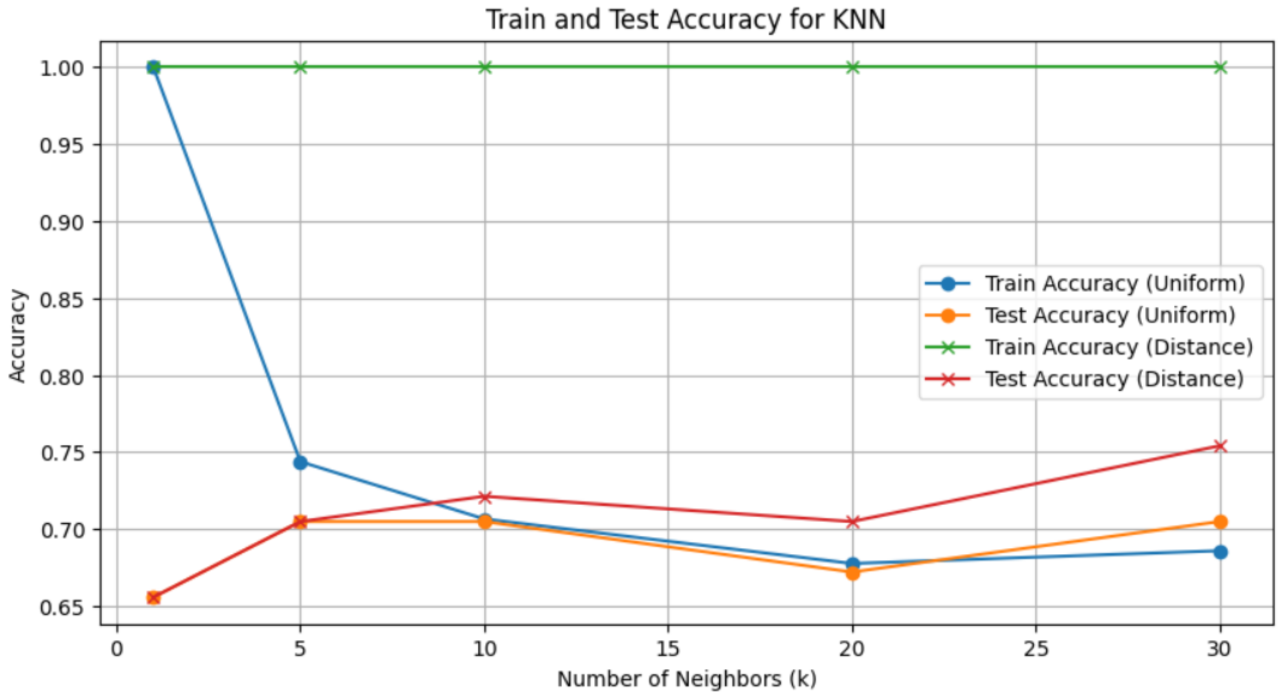
For the Naive Bayes model, a significant change might not happen after scaling, because Gaussian Naive Bayes assumes a normal distribution for each feature and the Min-Max scaling doesn't change the overall structure of a feature's distribution, just its range.

In conclusion, after comparing both, we see that with Min-Max scaling, the kNN model's performance improves because it now works with balanced distances for all features and the Gaussian Naive Bayes model is relatively insensitive since it only works with the probability distributions of the features.

c.

As calculated, the 5NN's model mean accuracy equals 0.6538 and Naive Bayes' is 0.8350. For the t-test, the p-value equals 0.9987 which is above 0.05 (a common alpha for t-tests) which leads to the rejection of the hypothesis "the model kNN is statistically superior to Naïve Bayes regarding accuracy".

2.
a.



b.
The plot shows the relation between the number of neighbors (k) and the training and test accuracies for both uniform and distance weighting strategies.

For uniform weights, all neighbors contribute equally, which can lead to instability when the neighbors have varying distances from the query point. This is confirmed by the training accuracy decreasing significantly and by the testing accuracy remaining below the testing accuracy for weighted distances.
For distance weights, closer neighbors have a higher impact on the decision, which typically improves generalization and robustness. This is shown in the plot, as the training accuracy is always 1 and the testing accuracy increases almost consistently.

In conclusion, as the number of neighbors increases, the model considers a broader neighborhood, leading to a smoother decision boundary that is less overfitted to the training data (for lower values of k, the model gets overfitted because it focuses too much on a smaller part of the training data and can get fitted to local data irregularities). Because of this, the model's testing accuracy tends to be higher, which means the model has a better generalization ability. This only happens up to a certain point, where the testing accuracy may plateau or even decrease because, if the value of k is excessively high, the model may get underfitted as it doesn't learn from the dataset (this last part is just a theoretical prediction as it isn't visible in the plot).

3.
The Naive Bayes model is especially well suited for datasets with independent features. Also, this model has variations for handling categorical or continuous data (separately), but it performs poorly when the dataset has both.

Firstly, in the heart-disease dataset, there are some features that may be correlated, which is common in medical datasets (for example, a person's cholesterol level may affect their blood pressure). This violation of the independence assumption can lead to poor performance, as the model might not capture the underlying relationships between these attributes.

Secondly, the types of variables in the dataset may pose a problem to its performance, because none of the Naive Bayes' variations handles both continuous and categorical data (ex. age and sex, respectively). Also, even if the model only observed continuous data, if the values don't follow a Gaussian/normal distribution or have many extreme outliers, the final model's distributions could get skewed and imprecise.

In summary, the naïve Bayes model's assumption of feature independence and its handling of categorical/continuous variables could limit its effectiveness on the heart-disease dataset.

END