Aprendizagem 2024/2025
Homework II – Group 115
(ist106485, ist107219)

Part I: Pen and paper

We collected four positive (P) observations, $\{x_1 = (A, 0), x_2 = (B, 1), x_3 = (A, 1), x_4 = (A, 0)\}$, and four negative (N) observations, $\{x_5 = (B, 0), x_6 = (B, 0), x_7 = (A, 1), x_8 = (B, 1)\}$. Consider the problem of classifying observations as positive or negative.

1. Compute the F1-measure of a kNN with $k = 5$ and Hamming distance using a leave-one-out evaluation schema. Show all calculus

   (a) Let us classify each observation based on the mentioned model, where $d(x_i, x_j)$ is the Hamming distance between the two observations

      i. For x_1:

      $$d(x_1, x_4) = 0; \quad d(x_1, x_3) = 1; \quad d(x_1, x_5) = 1; \quad d(x_1, x_6) = 1; \quad d(x_1, x_7) = 1$$

      The majority class out of $\{x_4, x_3, x_5, x_6, x_7\}$ is Negative. Therefore we classify x_1 as **Negative**.

      ii. For x_2:

      $$d(x_2, x_8) = 0; \quad d(x_2, x_3) = 1; \quad d(x_2, x_5) = 1; \quad d(x_2, x_6) = 1; \quad d(x_2, x_7) = 1$$

      The majority class out of $\{x_8, x_3, x_5, x_6, x_7\}$ is Negative. Therefore we classify x_2 as **Negative**.

      iii. For x_3:

      $$d(x_3, x_1) = 1; \quad d(x_3, x_2) = 1; \quad d(x_3, x_4) = 1; \quad d(x_3, x_7) = 0; \quad d(x_3, x_8) = 1$$

      The majority class out of $\{x_1, x_2, x_4, x_7, x_8\}$ is Positive. Therefore we classify x_3 as **Positive**.

      iv. For x_4:
      $x_4 = x_1$, Therefore we also classify x_4 as **Negative**.

v. For x_5:

$$d(x_5, x_1) = 1; \quad d(x_5, x_2) = 1; \quad d(x_5, x_4) = 1; \quad d(x_5, x_6) = 0; \quad d(x_5, x_8) = 1$$

The majority class out of $\{x_4, x_3, x_5, x_6, x_7\}$ is Positive. Therefore we classify x_5 as **Positive**.

vi. For x_6:
$x_6 = x_5$, Therefore we also classify x_6 as **Positive**.

vii. For x_7:
$x_7 = x_3$, Therefore we also classify x_7 as **Positive**.

viii. For x_8:
$x_8 = x_2$, Therefore we also classify x_6 as **Negative**.

(b) Finally we can calculate the F1-measure with the help of the confusion matrix

Actual \ Predicted	**Positive**	**Negative**
Positive	1	3
Negative	3	1

$$\mathrm{P} = \frac{TP}{TP + FP} = \frac{1}{1+3} = \frac{1}{4}$$
$$\mathrm{R} = \frac{TP}{TP + FN} = \frac{1}{1+3} = \frac{1}{4}$$

$$\mathbf{F1} = 2\frac{P \cdot R}{P + R} = 2\frac{\frac{1}{16}}{\frac{1}{2}}$$
$$= \boxed{0.25}$$

where, P and R are the precision and recall of the model respectively.

2. Propose a new metric (distance and/or k) that improves the latter's performance (i.e., the F1-measure) by three fold.

(a) Let us repeat the last exercise, this time in order to improve the F1-measure by 3 fold, we propose using $k = 3$ and to use Hamming distance only between the y_1 features, essentially giving 0 weight to y_2

i. For x_1:

$$d(x_1, x_3) = 0; \quad d(x_1, x_4) = 0; \quad d(x_1, x_7) = 0$$

The majority class out of $\{x_3, x_4, x_7\}$ is Positive. Therefore we classify x_1 as **Positive**.

ii. For x_2:

$$d(x_2, x_5) = 0; \quad d(x_2, x_6) = 0; \quad d(x_2, x_8) = 0$$

The majority class out of $\{x_5, x_6, x_8\}$ is Negative. Therefore we classify x_2 as **Negative**.

iii. For x_3:

$$d(x_3, x_1) = 0; \quad d(x_3, x_4) = 0; \quad d(x_3, x_7) = 0$$

The majority class out of $\{x_1, x_4, x_7\}$ is Positive. Therefore we classify x_3 as **Positive**.

iv. For x_4:
$x_4 = x_1$, Therefore we also classify x_4 as **Positive**.

v. For x_5:

$$d(x_5, x_2) = 0; \quad d(x_5, x_6) = 0; \quad d(x_5, x_8) = 0$$

The majority class out of $\{x_2, x_6, x_8, \}$ is Negative. Therefore we classify x_5 as **Negative**.

vi. For x_6:
$x_6 = x_5$, Therefore we also classify x_6 as **Negative**.

vii. For x_7:
$x_7 = x_3$, Therefore we also classify x_7 as **Positive**.

viii. For x_8:
$x_8 = x_2$, Therefore we also classify x_6 as **Negative**.

3. Once again computing the confusion matrix to calculate the F1-measure

Actual \ **Predicted**	**Positive**	**Negative**
Positive	3	1
Negative	1	3

$$\mathrm{P} = \frac{TP}{TP + FP} = \frac{3}{3+1} = \frac{3}{4}$$
$$\mathrm{R} = \frac{TP}{TP + FN} = \frac{3}{3+1} = \frac{3}{4}$$

$$\mathbf{F1} = 2\frac{P \cdot R}{P + R} = 2\frac{\frac{9}{16}}{\frac{3}{2}}$$
$$= \boxed{0.75}$$

Using the proposed metrics, we managed to increase the F1-measure by 3x.

An additional positive observation was acquired, $x_9 = (B, 0)$, and a third variable y_3 was independently monitored, yielding estimates,

$$y_3 \mid P = \{1.1, 0.8, 0.5, 0.9, 0.8\} \text{ and } y_3 \mid N = \{1, 0.9, 1.2, 0.9\}$$

4. 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.

   (a) First we calculate the class priors:

   $$P(\text{Pos}) = \frac{5}{9}; \quad P(\text{Neg}) = \frac{4}{9}$$

   (b) Next, we find the likelihoods of the features for each class:

      i. y_1, y_2 (dependent features)

      $$P(y_1 = A, y_2 = 0 \mid \text{Pos}) = \frac{2}{5}; \quad P(y_1 = A, y_2 = 1 \mid \text{Pos}) = \frac{1}{5};$$
      $$P(y_1 = B, y_2 = 0 \mid \text{Pos}) = \frac{1}{5}; \quad P(y_1 = B, y_2 = 1 \mid \text{Pos}) = \frac{1}{5};$$

      $$P(y_1 = A, y_2 = 1 \mid \text{Neg}) = \frac{1}{4}; \quad P(y_1 = B, y_2 = 0 \mid \text{Neg}) = \frac{1}{2};$$
      $$P(y_1 = B, y_2 = 1 \mid \text{Neg}) = \frac{2}{5};$$

ii. y_3 (normally distributed)

$$\mu_{|\text{Pos}} = \frac{1}{5} \sum_{y \in y_3|P} y = 0.82$$
$$\mu_{|\text{Neg}} = \frac{1}{4} \sum_{y \in y_3|N} y = 1$$

$$\sigma^2_{|\text{Pos}} = \frac{1}{5-1} \sum_{y \in y_3|P} (y - \mu_{|\text{Pos}}) = 0.0470$$
$$\sigma^2_{|\text{Neg}} = \frac{1}{4-1} \sum_{y \in y_3|N} (y - \mu_{|\text{Neg}}) = 0.0200$$

$$P(y_3 \mid \text{Pos}) \sim \mathcal{N}(\mu_{|\text{Pos}}, \sigma^2_{|\text{Pos}})$$
$$P(y_3 \mid \text{Neg}) \sim \mathcal{N}(\mu_{|\text{Neg}}, \sigma^2_{|\text{Neg}})$$

$$P(y_3 = k \mid \text{Pos}) = \phi_{Pos}(k) = \frac{1}{\sqrt{2\pi\sigma^2_{|\text{Pos}}}} \exp\left(-\frac{(k - \mu^2_{|\text{Pos}})}{2\sigma^2_{|\text{Pos}}}\right)$$
$$P(y_3 = k \mid \text{Neg}) = \phi_{Neg}(k) = \frac{1}{\sqrt{2\pi\sigma^2_{|\text{Neg}}}} \exp\left(-\frac{(k - \mu^2_{|\text{Neg}})}{2\sigma^2_{|\text{Neg}}}\right)$$

(c) Finally, we define the posterior probability of any new observation $x_{new} = (y_1, y_2, y_3)$ to be classified as h

$$P(h \mid x_{new}) = \frac{P(y_1, y_2 \mid h)P(y_3 \mid h)P(h)}{P(y_1, y_2)P(y_3)}$$

Consider now three testing observations,

$$\{x_a = (A, 1, 0.8), x_b = (B, 1, 1), x_c = (B, 0, 0.9)\}$$

5. Under a MAP assumption, classify each testing observation showing all your calculus.

(a) To classify each testing observation we must determine the maximum *a posteriori* hypothesis by maximizing:

$$h_{MAP} = \arg\max_h P(h \mid D)$$
$$h_{MAP} \overset{\text{Bayes Theorem}}{=} \arg\max_h \frac{P(D \mid h)P(h)}{P(D)}$$

(b) For each of the testing observations x_i, $i \in \{a, b, c\}$ we can calculate the hypothesis (Positive

/ Negative) like this:

$$P(\text{Pos} \mid x_i) = \frac{P(y_1, y_2 \mid \text{Pos})P(y_3 \mid \text{Pos})P(\text{Pos})}{P(y_1, y_2)P(y_3)}$$
$$\overset{\text{L. Total Prob.}}{=} \frac{P(y_1, y_2 \mid \text{Pos})P(y_3 \mid \text{Pos})P(\text{Pos})}{P(y_1, y_2 \mid \text{Pos})P(y_3 \mid \text{Pos})P(\text{Pos}) + P(y_1, y_2 \mid \text{Neg})P(y_3 \mid \text{Neg})P(\text{Neg})}$$

$$P(\text{Neg} \mid x_i) = 1 - P(\text{Pos} \mid x_i)$$

The Law of Total Probability can be applied since the events *Positive* and *Negative* are mutually exclusive and exhaustive.

(c) Let us now classify our testing observations using the calculated values obtained in the previous exercise, remembering that $\{y_1, y_2\}$, $\{y_3\}$ variable sets are independent and equally important; and y_3 is normally distributed

i. x_a

$$P(y_3 = 0.8 \mid \text{Pos}) = \phi_{Pos}(0.8) \approxeq 1.8324$$
$$P(y_3 = 0.8 \mid \text{Neg}) = \phi_{Neg}(0.8) \approxeq 1.0378$$

$$P(\text{Pos} \mid x_a) = \frac{P(y_1 = A, y_2 = 1 \mid \text{Pos})P(y_3 = 0.8 \mid \text{Pos})P(\text{Pos})}{P(y_1 = A, y_2 = 1 \mid \text{Pos})P(y_3 = 0.8 \mid \text{Pos})P(\text{Pos}) + P(y_1 = A, y_2 = 1 \mid \text{Neg})P(y_3 = 0.8 \mid \text{Neg})P(\text{Neg})}$$
$$= \frac{\frac{1}{5}\Phi_{Pos}(0.8)\frac{5}{9}}{\frac{1}{5}\Phi_{Pos}(0.8)\frac{5}{9} + \frac{1}{4}\Phi_{Neg}(0.8)\frac{4}{9}}$$
$$\approxeq \boxed{0.6384}$$

$$P(\text{Neg} \mid x_a) = 1 - P(\text{Pos} \mid x_a)$$
$$\approxeq 1 - 0.6384$$
$$\approxeq \boxed{0.3616}$$

$P(\text{Pos} \mid x_a) > P(\text{Neg} \mid x_a)$, therefore we classify x_a as **Positive**

ii. x_b

$$P(y_3 = 1 \mid \text{Pos}) = \Phi_{Pos}(1) \approxeq 1.3037$$
$$P(y_3 = 1 \mid \text{Neg}) = \Phi_{Neg}(1) \approxeq 2.8209$$

$$P(\text{Pos} \mid x_b) = \frac{P(y_1 = B, y_2 = 1 \mid \text{Pos})P(y_3 = 1 \mid \text{Pos})P(\text{Pos})}{P(y_1 = B, y_2 = 1 \mid \text{Pos})P(y_3 = 1 \mid \text{Pos})P(\text{Pos}) + P(y_1 = B, y_2 = 1 \mid \text{Neg})P(y_3 = 1 \mid \text{Neg})P(\text{Neg})}$$
$$= \frac{\frac{1}{5}\Phi_{Pos}(1)\frac{5}{9}}{\frac{1}{5}\Phi_{Pos}(1)\frac{5}{9} + \frac{1}{4}\Phi_{Neg}(1)\frac{4}{9}}$$
$$\approxeq \boxed{0.3161}$$

$$P(\text{Neg} \mid x_b) = 1 - P(\text{Pos} \mid x_b)$$
$$\approxeq 1 - 0.3161$$
$$\approxeq \boxed{0.6839}$$

$P(\text{Pos} \mid x_b) < P(\text{Neg} \mid x_b)$, therefore we classify x_b as **Negative**

iii. x_c

$$P(y_3 = 0.9 \mid \text{Pos}) = \Phi_{Pos}(0.9) \cong 1.7191$$
$$P(y_3 = 0.9 \mid \text{Neg}) = \Phi_{Neg}(0.9) \cong 2.1970$$

$$\begin{aligned} P(\text{Pos} \mid x_c) &= \frac{P(y_1 = B, y_2 = 0 \mid \text{Pos})P(y_3 = 0.9 \mid \text{Pos})P(\text{Pos})}{P(y_1 = B, y_2 = 0 \mid \text{Pos})P(y_3 = 0.9 \mid \text{Pos})P(\text{Pos}) + P(y_1 = B, y_2 = 0 \mid \text{Neg})P(y_3 = 0.9 \mid \text{Neg})P(\text{Neg})} \\ &= \frac{\frac{1}{5}\Phi_{Pos}(0.9)\frac{5}{9}}{\frac{1}{5}\Phi_{Pos}(1)\frac{5}{9} + \frac{1}{2}\Phi_{Neg}(0.9)\frac{4}{9}} \\ &\cong \boxed{0.2812} \end{aligned}$$

$$\begin{aligned} P(\text{Neg} \mid x_c) &= 1 - P(\text{Pos} \mid x_c) \\ &\cong 1 - 0.2812 \\ &\cong \boxed{0.7188} \end{aligned}$$

$P(\text{Pos} \mid x_c) < P(\text{Neg} \mid x_c)$, therefore we classify x_c as **Negative**

At last, consider only the following sentences and their respective connotations,

$$\{(\text{"}\textit{Amazing run}\text{"}, \text{Pos}), (\text{"}\textit{I like it}\text{"}, \text{Pos}), (\text{"}\textit{Too tired}\text{"}, \text{Neg}), (\text{"}\textit{Bad run}\text{"}, \text{Neg})\}$$

6. Using a naïve Bayes under a ML assumption, classify the new sentence "*i like to run*". For the likelihoods calculation consider the following formula,

$$P(t_i \mid c) = \frac{freq(t_i) + 1}{N_c + V}$$

where t_i represents a certain term i, V the number of unique terms in the vocabulary, and N_c the total number of terms in class c. Show all calculus.

(a) Firstly, let us calculate the prior probabilities

$$P(\text{Pos}) = \frac{\text{Positive sentences}}{\text{Total sentences}} = \frac{2}{4} = \frac{1}{2}$$

$$P(\text{Neg}) = \frac{\text{Negative sentences}}{\text{Total sentences}} = \frac{2}{4} = \frac{1}{2}$$

(b) We can also obtain the frequencies of each term and total number of terms per class c and the number of unique terms in the vocabulary

$$\begin{aligned} V &= \#\{\text{"Amazing"}, \text{"run"}, \text{"I"}, \text{"like"}, \text{"it"}, \text{"Too"}, \text{"tired"}, \text{"Bad"}\} = 8 \\ N_{Pos} &= \#\{\text{"Amazing"}, \text{"run"}, \text{"I"}, \text{"like"}, \text{"it"}\} = 5 \\ N_{Neg} &= \#\{\text{"Too"}, \text{"tired"}, \text{"Bad"}, \text{"run"}\} = 4 \end{aligned}$$

Term / Class	Amazing	run	I	like	it	Too	tired	Bad
Positive	1	1	1	1	1	0	0	0
Negative	0	1	0	0	0	1	1	1

Table 1: Frequency of each term per class

(c) Finally let us classify the new sentence S_{new} = *"I like to run"* under ML assumption $h_{ML} = \arg\max_h P(S \mid h)$

$$\begin{aligned} P(\text{"I like to run"} \mid \text{Pos}) &= P(\text{"I"} \mid \text{Pos})P(\text{"like"} \mid \text{Pos})P(\text{"to"} \mid \text{Pos})P(\text{"run"} \mid \text{Pos}) \\ &= \frac{1+1}{5+8} \cdot \frac{1+1}{5+8} \cdot \frac{0+1}{5+8} \cdot \frac{1+1}{5+8} \\ &\approx \boxed{2.8 \times 10^{-4}} \end{aligned}$$

$$\begin{aligned} P(\text{"I like to run"} \mid \text{Neg}) &= P(\text{"I"} \mid \text{Neg})P(\text{"like"} \mid \text{Neg})P(\text{"to"} \mid \text{Neg})P(\text{"run"} \mid \text{Neg}) \\ &= \frac{0+1}{4+8} \cdot \frac{0+1}{4+8} \cdot \frac{0+1}{4+8} \cdot \frac{1+1}{4+8} \\ &\approx \boxed{9.6 \times 10^{-5}} \end{aligned}$$

$P(\text{"I like to run"} \mid \text{Pos}) > P(\text{"I like to run"} \mid \text{Neg})$, therefore we classify S_{new} as **Positive**.

Part II: Programming

1. Compare the performance of a kNN with $k = 5$ and a naïve Bayes with Gaussian assumption (consider all remaining parameters as default):

   (a) Plot two boxplots with the fold accuracies for each classifier. Is there one more stable than the other regarding performance? Why do you think that is the case? Explain.

   Answer:

   The model that demonstrates greater stability in performance is naïve Bayes. This stability is shown by the smaller dispersion of accuracy values across different folds, observed in the boxplot through a narrower interquartile range, whereas kNN displays a wider range of values, with some folds resulting in significantly lower accuracies.

   The explanation for this difference in stability lies in the intrinsic characteristics of the algorithms. naïve Bayes relies on the assumption of independence among the variables and uses Gaussian distribution, which is more robust against variations in data distribution. This algorithm does not depend on distance metrics between examples, as is the case with kNN, making it less susceptible to local fluctuations in the data.

   On the other hand, kNN heavily depends on the distances between data points, making it particularly sensitive to the distribution of these points in different samples. This sensitivity translates into greater variance in performance, as the proximity of examples from different classes can significantly influence classification in certain folds.

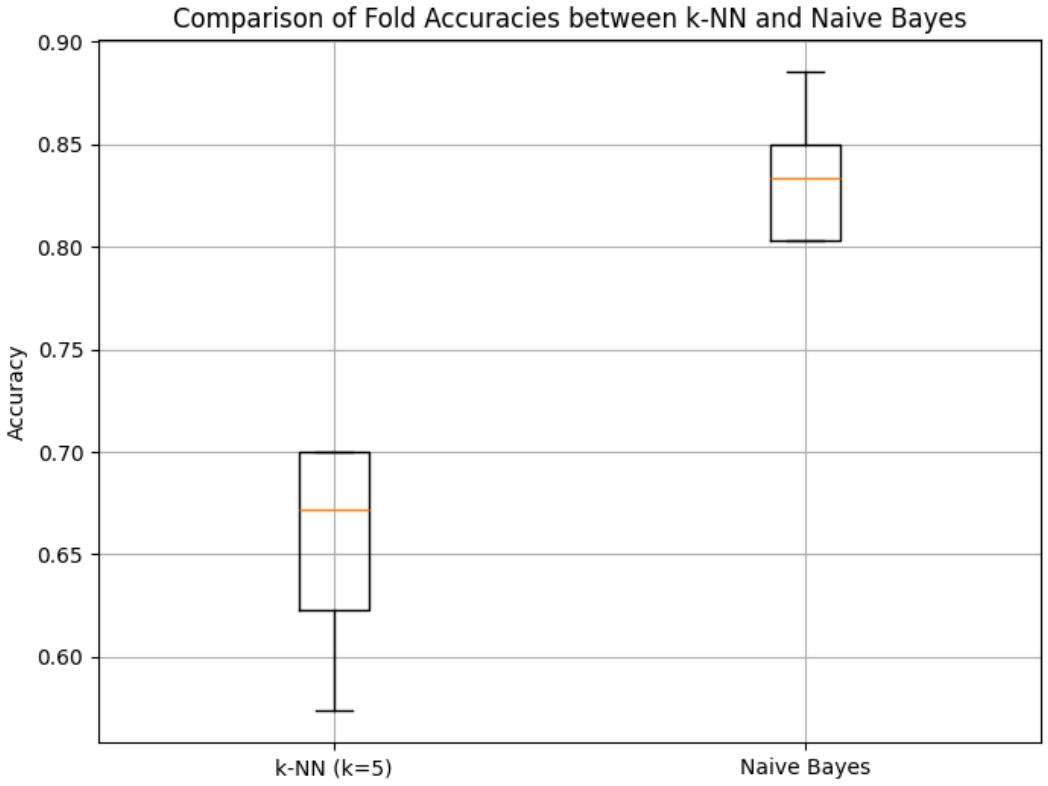

Figure 1: Fold accuracies of kNN vs naïve Bayes with Cross Validation

(b) Report the accuracy of both models, this time scaling the data with a Min-Max scaler before training the models. Explain the impact that this preprocessing step has on the performance of each model, providing an explanation for the results.

Answer:

After applying the Min-Max Scaler, a distinct impact is observed on the performance of the kNN and Naive Bayes models. In the case of kNN, there was a significant improvement, with accuracies ranging from 78.3% to 85.2%. This increase in performance and stability is related with the distance-based approach of kNN. Without normalization, variables with larger ranges tend to dominate the distance calculations, undermining the model's generalization ability. The Min-Max normalization adjusts all attributes to the same range, allowing each variable to have an equitable contribution to the calculation of distances between data points.

On the other hand, naïve Bayes showed almost equal performance to that observed before normalization, with accuracies ranging from 80.3% to 88.5%. This is due to the probabilistic model that assumes independence among variables and does not rely on distance metrics. Therefore, the normalization process has minimal to no impact since the scale of the attributes does not interfere with how the algorithm calculates the conditional probabilities necessary for classification.

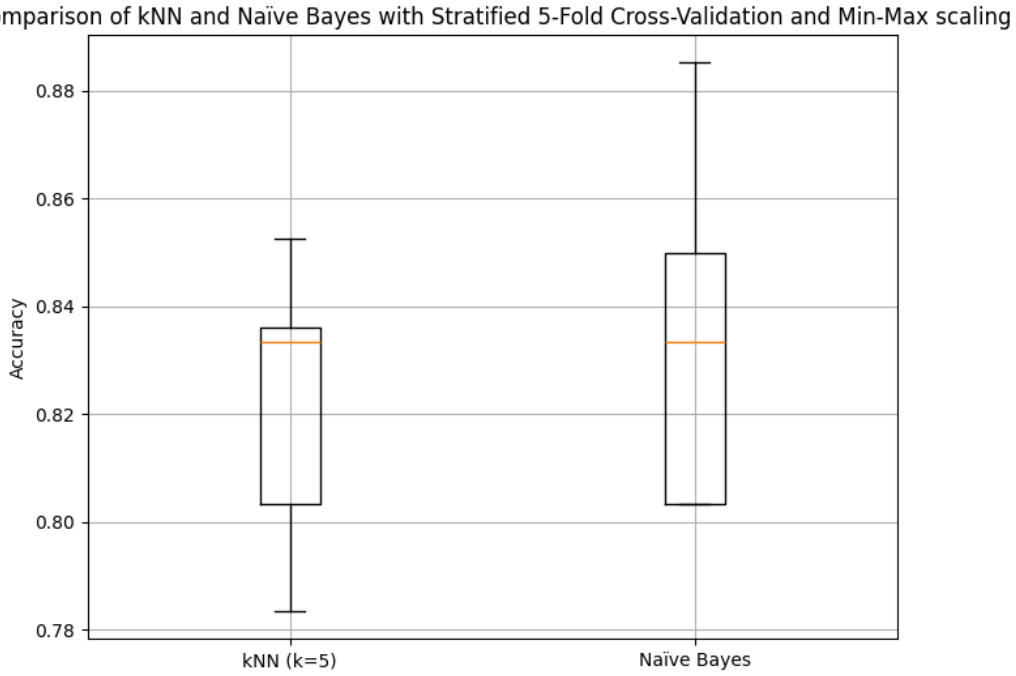

Figure 2: Fold accuracies of kNN vs naïve Bayes with Cross Validation and Min-Max Scaling

(c) Using `scipy`, test the hypothesis "the kNN model is statistically superior to naïve Bayes regarding accuracy", asserting whether it is true.

Answer:

Based on the conducted paired t-test, a t statistic of -6.69 and a p-value of 0.0026 were obtained, which is below the established significance level ($\alpha = 0.05$). Consequently, the null hypothesis is rejected, indicating that there is a statistically significant difference between the accuracies of the kNN and naïve Bayes models. However, given that the t statistic is negative, it can be concluded that naïve Bayes outperformed kNN in terms of accuracy on the analyzed dataset.

2. Using a 80-20 train-test split, vary the number of neighbors of a kNN classifier using $k = \{1, 5, 10, 20, 30\}$. Additionally, for each k, train one classifier using uniform weights and distance weights.

(a) Plot the train and test accuracy for each model.

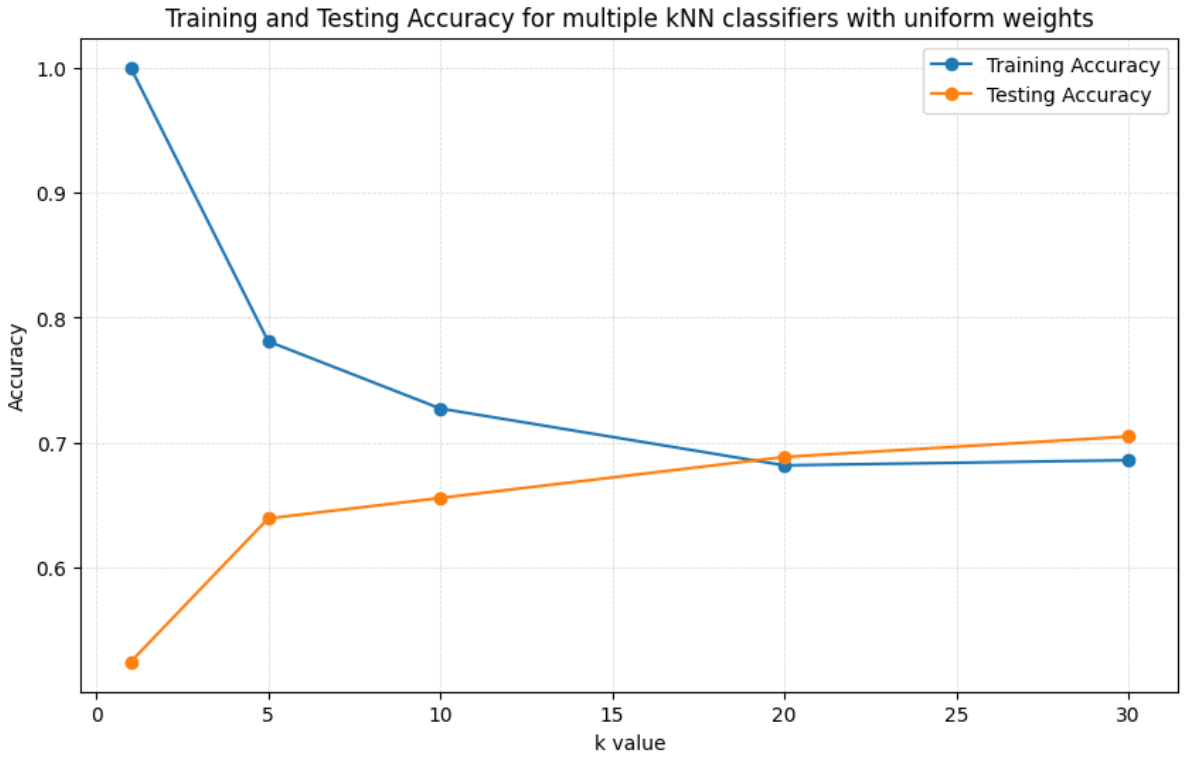

Figure 3: Training vs Testing Accuracy of multiple kNN models with uniform weights

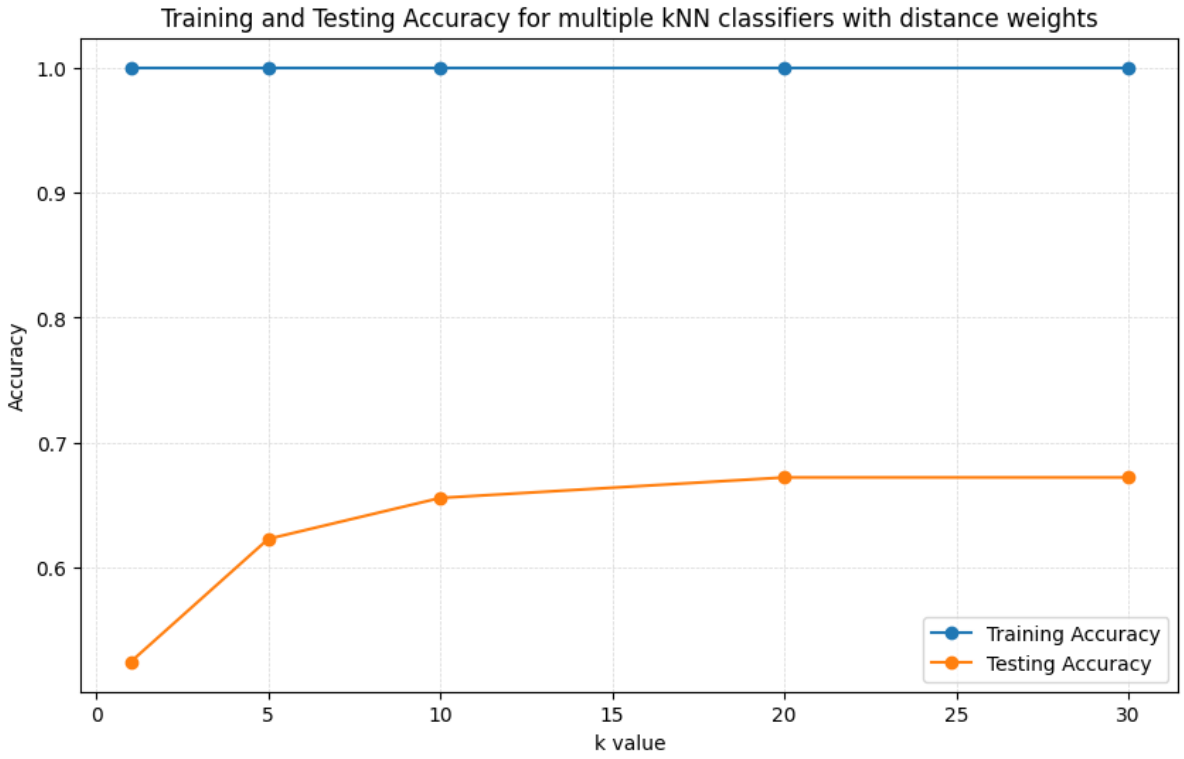

Figure 4: Training vs Testing Accuracy of multiple kNN models with distance weights

(b) Explain the impact of increasing the neighbors on the generalization ability of the models.

Answer:

kNN with uniform weights:

In uniform-weighted kNN, the k nearest neighbors have equal influence on the classification decision.

By increasing the value of k, the model averages over more neighbors, reducing variance and effectively minimizing noise, which enhances generalization. However, if k is too large, the decision boundaries become overly smooth, making the classifier less sensitive to local data structures.

As a result, the generalization ability can degrade due to underfitting when k is too high.

kNN with distance weights:

In distance-weighted kNN, closer neighbours are given more importance than farther ones.

As the value of k increases, the closest neighbors carry more weight in the classification decision, which significantly helps to reduce the risk of underfitting compared to uniform-weighted kNN.

3. Considering the unique properties of the `heart-disease.csv` dataset, identify two possible difficulties of the naïve Bayes model used in the previous exercises when learning from the given dataset.

   (a) naïves Bayes model assumes that features are conditionally independent from each other. However when dealing with medical data, it is most likely that many parameters could be correlated.

   (b) Plotting the sample count per sex group, clearly reveals an overrepresentation of the sex "1" group (2.2x more). This degrades the generalization ability of the model when predicting the minority sex group ("0"). Moreover, using binary variables with Gaussian assumption, further reduces prediction accuracy.

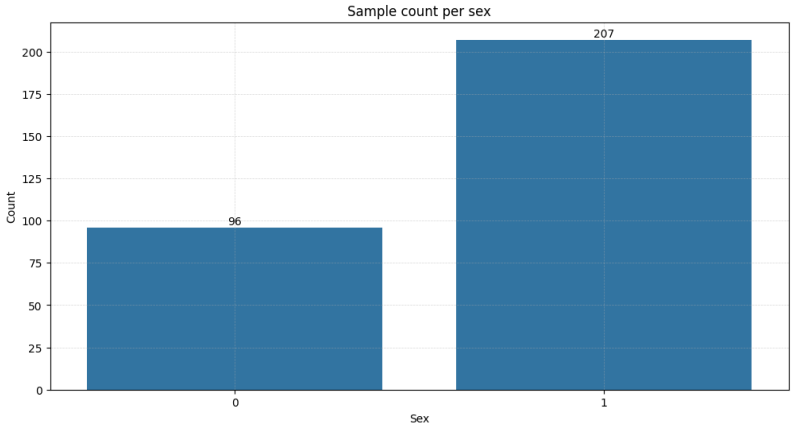

Figure 5: Sample count from `heart-disease.csv` per sex group