1. Density estimation

a. (i) Beta distribution

For a Beta random variable, the PDF is given by:

$$f(x) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha,\beta)}$$

$$B(\alpha,\beta) = \frac{\Gamma\alpha\Gamma\beta}{\Gamma(\alpha+\beta)}$$

We know, $\Gamma\alpha = \int_0^\infty x^{t-1}\, e^{-x} dx$

And, $\Gamma\alpha = (\alpha-1)!$

Here α is unknown, $\beta = 1$ given.

$$B(\alpha,\beta) = \frac{\Gamma\alpha\Gamma\beta}{\Gamma(\alpha+\beta)}$$

$$B(\alpha,\beta) = \frac{\Gamma\alpha\Gamma 1}{\Gamma(\alpha+1)}$$

$$B(\alpha,\beta) = \frac{(\alpha-1)!}{(\alpha)!} = \frac{(\alpha-1)!}{(\alpha)*(\alpha-1)!} = \frac{1}{(\alpha)}$$

Also

$$f(x) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha,\beta)} = \frac{x^{(\alpha-1)}(1-x)^{(1-1)}}{1/\alpha} = \alpha x^{(\alpha-1)}$$

We have log likelihood defined as,

$L(\alpha) = \sum_{i=1}^{n} \log(f(x_i,\alpha))$

$L(\alpha) = \sum_{i=1}^{n} \log(f(x_i,\alpha)\,)$

We can find the derivative of above function and equate it to zero to find the maximum of likelihood.

(Name)Ravishankar Sivaraman (USC ID)6370-0913-06 (Email)rsivaram@usc.edu

$$\frac{\partial L(\alpha)}{\partial \alpha} = \frac{\partial}{\partial \alpha}\sum_{i=1}^{n} \log(f(x_i, \alpha) = \frac{\partial}{\partial \alpha}\sum_{i=1}^{n} \log\left(\alpha {x_i}^{(\alpha-1)}\right) = \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}\left(\log(\alpha) + \log\left({x_i}^{(\alpha-1)}\right)\right)$$

$$= \sum_{i=1}^{n}\frac{1}{\alpha} + \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}(\,\alpha \log(x_i) - \log(x_i)) = \frac{n}{\alpha} + \sum_{i=1}^{n}\log(x_i) = 0$$

$$0 = \frac{n}{\alpha} + \sum_{i=1}^{n}\log(x_i)$$

Hence we have

$$\boxed{\hat{\alpha} = \frac{-n}{\sum_{i=1}^{n}\log(x_i)}}$$

1. **a. (ii) Normal Distribution**

$$\mu = \theta, \sigma = \theta^2$$

$$f(x) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-\frac{(x-\theta)^2}{2\theta}}$$

We have the likelihood function defined for the I.I.D random variables as the product of each pdf.

From the definition of Log likelihood, we can write.

$$L(\theta) = \sum_{i=1}^{n}\log(f(x_i, \theta)\,)$$

$$L(\theta) = \frac{\partial}{\partial \theta}\sum_{i=1}^{n} -\log\left(\sqrt{2\pi\theta}\,\right) - \frac{(x_i-\theta)^2}{2\theta}$$

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{\partial}{\partial \theta}\left(\frac{n}{2}\log(2\pi\,)\right) - \frac{\partial}{\partial \theta}\left(\frac{n}{2}\log(\theta)\right) - \frac{\partial}{\partial \theta}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\theta}$$

We can calculate the Maximum Log likelihood of the function by equating its derivative (w.r.t θ) to zero.

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{n}{2\theta} - \frac{\partial}{\partial \theta}\sum_{i=1}^{n}\frac{(x_i)^2}{2\theta} - x_i + \frac{\theta}{2} = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i)^2 - \frac{n}{2} = 0$$

$$-\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i)^2 - \frac{n}{2} = 0$$

$$\frac{1}{2}\sum_{i=1}^{n}(x_i)^2 = \frac{n\theta}{2} + \frac{n\theta^2}{2}$$

$$n\theta^2 + n\theta - \sum_{i=1}^{n}(x_i)^2 = 0$$

$$\boxed{\hat{\theta} = \frac{-n \pm \sqrt{n^2 + 4n\sum_{i=1}^{n}(x_i)^2}}{2n}}$$

(i)

$$E\left[\hat{f}(x)\right] = \frac{1}{N}\sum_{i=1}^{N} E\left[\frac{1}{h}K\left(\frac{x - X_i}{h}\right)\right] = \frac{1}{N}N\int_{-\infty}^{\infty}\frac{1}{h}K\left(\frac{x-z}{h}\right)f(z)dz$$

$$= \int_{-\infty}^{\infty}\frac{1}{h}K(u)f(u - uh)du$$

(ii) Taylors expansion

$$f(x - uh) = f(x) + f^{(1)}(x)(-uh) + \frac{1}{2}f^{(2)}(x)(-uh)^2 + \cdots + \frac{1}{n!}f^{(n)}(x)(-uh)^n + O(h^{n+1})$$

$$\int_{-\infty}^{\infty}K(u)f(x - uh)du$$

$$= f(x) + f^{(1)}(x)(-h)\int_{-\infty}^{\infty}K(u)udu + \cdots + \frac{1}{2}f^{(2)}(x)(-h)^2\int_{-\infty}^{\infty}K(u)u^2du + O(h^3)$$

$$\int_{-\infty}^{\infty}K(u)f(x - uh)du = f(x) + \frac{1}{2}f^{(2)}(x)h^2\int_{-\infty}^{\infty}K(u)u^2du + O(h^3)$$

(iii) Bias

$$Bias\left(\hat{f}(x)\right) = E\left[\hat{f}(x)\right] - f(x) = \frac{1}{2}f^{(2)}(x)h^2\int_{-\infty}^{\infty}K(u)u^2du + O(h^3)$$

2) Naive Bayes

a.

Given, $P(Y = 1) = \pi$, we can imply that $P(Y = 0) = 1 - \pi$

We know:

$$P(Y = 1\,|X) = \frac{P(X\,|Y = 1)\ \cdot\ P(Y = 1)}{P(X\,|Y = 1)\ \cdot\ P(Y = 1) + \ P(X\,|Y = 0)\ \cdot\ P(Y = 0)}$$

Dividing numerator and denominator by $P(X\,|Y=1)\ \cdot\ P(Y=1)$, we get:

$$P(Y=1\,|X) = \frac{1}{1+\frac{P(X\,|Y=0)\ \cdot\ P(Y=0)}{P(X\,|Y=1)\ \cdot\ P(Y=1)}}$$

We can express this as

$$P(Y=1\,|X) = \frac{1}{1+e^{\ln(\frac{P(X\,|Y=0)\,\cdot\,P(Y=0)}{P(X\,|Y=1)\,\cdot\,P(Y=1)})}}$$

Expanding vector $X=\{X_1, X_2, \dots, X_D\}$, and also under the assumption that the dimensions are independent, we can write

$$P(X\,|Y=0) = \prod_{i=1}^{D} P(X_i\,|Y=0)$$

and

$$P(X\,|Y=1) = \prod_{i=1}^{D} P(X_i\,|Y=1)$$

Taking ln of the above 2 expressions gives

$$\ln(P(X\,|Y=0)) = \sum_{i=1}^{D} \ln(P(X_i\,|Y=0))$$

$$\ln(P(X\,|Y=1)) = \sum_{i=1}^{D} \ln(P(X_i\,|Y=1))$$

Substituting these values,

$$P(Y=1\,|X) = \frac{1}{1+e^{\ln\left(\frac{P(Y=0)}{P(Y=1)}\right)+\sum_{i=1}^{D}\ln(\frac{P(X_i\,|Y=0)}{P(X_i\,|Y=1)})}}$$

Substituting for $P(Y=0)\ and\ \ P(Y=1)$

$$P(Y=1\,|X) = \frac{1}{1+e^{\ln\left(\frac{1-\pi}{\pi}\right)+\sum_{i=1}^{D}\ln(\frac{P(X_i\,|Y=0)}{P(X_i\,|Y=1)})}}$$

Now, let at look at $\sum_{i=1}^{D}\ln(\frac{P(X_i\,|Y=0)}{P(X_i\,|Y=1)})$ and apply the Gaussian PDF for this.

$$\sum_{i=1}^{D} \ln\left(\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}\right) = \sum_{i=1}^{D} \ln\left(\frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(X_i-\mu_{i0})^2}{2\sigma_{i0}^2}}}{\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}}}\right)$$

Cancelling $\frac{1}{\sqrt{2\pi\sigma_i^2}}$ from the numerator and denominator and simplifying

$$\sum_{i=1}^{D} \ln\left(\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}\right) = \sum_{i=1}^{D} \ln\left(e^{\frac{(X_i-\mu_{i1})^2}{2\sigma_i^2} - \frac{(X_i-\mu_{i0})^2}{2\sigma_i^2}}\right)$$

$$\sum_{i=1}^{D} \ln\left(\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}\right) = \sum_{i=1}^{D} \frac{(2(\mu_{i0}-\mu_{i1})X_i + (\mu_{i1}^2 - \mu_{i0}^2)}{2\sigma_i^2}$$

$$\sum_{i=1}^{D} \ln\left(\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}\right) = \sum_{i=1}^{D} \frac{(\mu_{i1}^2 - \mu_{i0}^2)}{2\sigma_i^2} + \sum_{i=1}^{D} \frac{(\mu_{i0}-\mu_{i1})}{\sigma_i^2} X_i$$

From earlier, we have

$$P(Y=1 \mid X) = \frac{1}{1 + e^{\ln\left(\frac{1-\pi}{\pi}\right) + \sum_{i=1}^{D} \ln\left(\frac{P(X_i \mid Y=0)}{P(X_i \mid Y=1)}\right)}}$$

To represent this in the form

$$P(Y=1 \mid X) = \frac{1}{1 + e^{w_0 + w^T X}}$$

We can substitute the above simplification to get

$$P(Y=1\,|X) = \frac{1}{1+e^{\ln\left(\frac{1-\pi}{\pi}\right)+\sum_{i=1}^{D}\frac{(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}+\sum_{i=1}^{D}\frac{(\mu_{i0}-\mu_{i1})}{\sigma_i^2}X_i}}$$

Therefore we have

$$\boxed{w_0 = \ln\left(\frac{1-\pi}{\pi}\right)+\sum_{i=1}^{D}\frac{(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}}$$

And

$$\boxed{w_j = \frac{(\mu_{i0}-\mu_{i1})}{\sigma_i^2}}$$

2) Naive Bayes

b.

$$Log\ Likelihood\ (LL) = \log\prod_{c=0}^{1}\prod_{\substack{i=1\\y_i=c}}^{N}\prod_{j=1}^{D}\left[\left(2\pi\sigma_j^2\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_j^2}\left(x_{ij}-\mu_{jc}\right)^2\right\}\right]$$

$$= -\sum_{c=0}^{1}\sum_{\substack{i=1\\y_i=c}}^{N}\sum_{j=1}^{D}\left[\frac{1}{2}\log\left(2\pi\sigma_j^2\right)-\frac{1}{2\sigma_j^2}\left(x_{ij}-\mu_{jc}\right)^2\right]$$

$$\frac{\delta}{\delta\mu_{jc}}LL = \sum_{\substack{i=1\\y_i=c}}^{N}\frac{x_{ji}-\mu_{jc}}{\sigma_j^2} = 0$$

Equate to 0 and split the summation.

$$\sum_{\substack{i=1\\y_i=c}}^{N}x_{ji} = \sum_{\substack{i=1\\y_i=c}}^{N}\mu_{jc} = N_c\mu_{jc}$$

$$\mu_{jc} = \frac{1}{N}\sum_{\substack{i=1\\y_i=c}}^{N}x_{ji}$$

For σ_j:

$$\frac{\delta}{\delta(\sigma_j^2)}LL = \sum_{c=0}^{1}\left[-\sum_{\substack{i=1\\y_i=c}}^{N}\frac{1}{2\sigma_j^2} + \sum_{\substack{i=1\\y_i=c}}^{N}\frac{1}{2\sigma_j^4}\left(x_{ij}-\mu_{jc}\right)^2\right]$$

Equate to 0 and separate:

$$\sum_{c=0}^{1}\sum_{\substack{i=1\\y_i=c}}^{N}\frac{1}{\sigma_j^2} = \sum_{c=0}^{1}\sum_{\substack{i=1\\y_i=c}}^{N}\frac{1}{\sigma_j^4}\left(x_{ij}-\mu_{jc}\right)^2$$

$$\sigma_j^2 = \frac{1}{N}\sum_{c=0}^{1}\sum_{\substack{i=1\\y_i=c}}^{N}\frac{1}{\sigma_j^4}\left(x_{ij}-\mu_{jc}\right)^2$$

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{c=0}^{1}\sum_{\substack{i=1\\y_i=c}}^{N}\frac{1}{\sigma_j^4}\left(x_{ij}-\mu_{jc}\right)^2}$$

3) Nearest Neighbor

a.

(i) Normalized L2 distances for the student (20,7) are

Neighbor	L2 Distance
Mathematics(0,49)	1.885585211
Mathematics(-7,32)	1.62112587
Mathematics(-9,47)	2.08303616
Electrical Engineering (29,12)	0.475296728
Electrical Engineering (49,31)	1.67813313
Electrical Engineering (37,38)	1.450024856
Computer Science (8,9)	0.584348182
Computer Science (13,-1)	0.457547526
Computer Science (-6,-3)	1.312925396
Computer Science (-21,12)	1.988426411
Economics (27,-32)	1.541500231
Economics (19,-14)	0.811290371
Economics (27,-20)	1.094690895

Therefore the prediction of major using K-Nearest Neighbors for K = 1 will be the one with the shortest L2 distance which is Computer Science (13,-1) which is at a distance of 0.457547526

(ii) For K = 5

Based on L2 Distances

5 Nearest Neighbors	L2 Distance
Computer Science (13,-1)	0.45754753
Electrical Engineering (29,12)	0.47529673
Computer Science (8,9)	0.58434818
Economics (19,-14)	0.81129037
Economics (27,-20)	1.0946909

`

Therefore the prediction from above will be Computer Science (based on the tie breaker which picks the one with the shortest L2 distance)

(iii) Normalized L1 distances for the student (20,7) are

Neighbor	L1 Distance
Mathematics(0,49)	2.585099025
Mathematics(-7,32)	2.267391087
Mathematics(-9,47)	2.94239689
Electrical Engineering (29,12)	0.627248912
Electrical Engineering (49,31)	2.325365926
Electrical Engineering (37,38)	2.016081326
Computer Science (8,9)	0.656364518
Computer Science (13,-1)	0.646402943
Computer Science (-6,-3)	1.640654921
Computer Science (-21,12)	2.171877304
Economics (27,-32)	1.841900435
Economics (19,-14)	0.858122777
Economics (27,-20)	1.379127212

The prediction of major using K-Nearest Neighbors for K = 1 will be the one with the shortest L1 distance which can be either of Electrical Engineering (29,12) which is at a distance of 0.627248912.

(iv) For K = 5

Based on L1 Distances

5 Nearest Neighbors	L1 Distance
Electrical Engineering (29,12)	0.62724891
Computer Science (13,-1)	0.64640294
Computer Science (8,9)	0.65636452
Economics (19,-14)	0.85812278
Economics (27,-20)	1.37912721

Therefore the prediction from above will be Computer Science (based on the tie breaker which picks the one with the shortest L1 distance)

(v) Performance comparison

We have the following results

Metric	Major Chosen
L2 Distance, K = 1	Computer Science
L2 Distance, K = 5	Computer Science
L1 Distance, K = 1	Electrical Engineering
L1 Distance, K = 5	Computer Science

While the student is assigned an Electrical Major in K = 1 with L1 distances, I believe looking at a broader data set around it helps it to get classified correctly (as Computer Science). All the other predictions seem to agree.

3) Nearest Neighbor

b. (i) Unconditional density.

Sphere volume is V,

Total data points are =$\sum N_c = N$

Class prior is given by $\mathrm{P}(\mathrm{Y} = \mathrm{c}) = \frac{N_c}{N}$

conditional probability is associated with each class is

$$\mathrm{P}(\mathrm{x}|\mathrm{Y} = \mathrm{c}) = \frac{K_c}{NV}$$

unconditional density is given by $\mathrm{P}(\mathrm{x}) = \sum_{i=1}^{c} \mathrm{P}(\mathrm{x}|\mathrm{Y} = \mathrm{i})\, P(\mathrm{Y} = i)$

Where i=1,2,…C are values taken by class Y.

$$\mathrm{P}(\mathrm{x}) = \sum_{i=1}^{c} \mathrm{P}(\mathrm{x}|\mathrm{Y} = \mathrm{i})\, P(\mathrm{Y} = i)$$

Given $\sum K_c = K$,

$$\mathrm{P(x)} = \frac{K_1}{N_1 V} * \frac{N_1}{N} + \frac{K_2}{N_2 V} * \frac{N_2}{N} + \frac{K_3}{N_3 V} * \frac{N_3}{N} + \cdots \frac{K_c}{N_c V} * \frac{N_c}{N}$$

$$\mathrm{P(x)} = \frac{K_1}{NV} + \frac{K_2}{NV} + \frac{K_3}{NV} + \cdots \frac{K_c}{NV} = \frac{\sum K_i}{NV} = \frac{K}{NV}$$

b. (ii) Class membership probability.

By using the Bayes formula,

$$\mathrm{P(Y = c|x)} = \frac{P(\mathrm{x|Y = c})P(Y = c)}{P(\mathrm{x})}$$

From above part we can write $P(\mathrm{x}) = \frac{K}{NV}$

$$\mathrm{P(Y = c|x)} = \frac{\frac{K_c}{N_c V} * \frac{N_c}{N}}{\frac{K}{NV}} = \frac{K_c}{K}$$

4) Decision Tree

(a) For the given set of accident data we get higher information gain if the entropy is low if we select "Traffic" as the predictor variable to get the best prediction.

(b) The trees T1 and T2 provide us with the same kind of information. If the given trees T1 and T2 have features which are continuous we can transform them by taking into account the decision boundary, subtracting the mean and dividing by variance associated with it. We get to see that they have the same structure and accuracy, hence the same information.

(c) Consider the difference between Gini Index and Cross Entropy,

$$G - CE = \sum_{k=1}^{K} [p_k(1 - p_k)] + \sum_{k=1}^{K} [p_k \, log \, p_k]$$
$$= \sum_{k=1}^{K} [p_k(1 - p_k + log \, p_k)]$$

Examining the function $f(x) = 1 - x + \log(x)$, where the base is less than or equal to e. On a positive real line the function f is continuous.

Taking the derivative of f, we get

$$\frac{d}{dx} f = -1 + \frac{1}{x \, log(a)}$$

Here, a is the base of the log. We see that this function is also continuous on a positive real line. $\forall\ a \leq e, \log(a) \leq 1 \implies \frac{1}{x\log(a)} \leq 1\ \ \forall\ x \in (0,1),$ and for $x = 1, \frac{1}{x\log(a)} = 1$. This implies that $\frac{d}{dx} f(x) > 0\ \forall x \in (0,1), a < e\ so\ f$ has no critical points in (0,1).
We see that $f(x) \to -\infty\ as\ x \to 0+$ and consider x =1. $f(x) = 0$, and has no previous critical points so it cannot have any positive points to $f(0)$. If it were to have a positive point it must have decreased to $f(0)$ since its continuous but it then must have a negative derivative , meaning its derivative must have a zero, meaning it must have a critical point which is a contradiction.

Thus, $1 - p_k + log\ p_k < 0$, meaning that $G - CE < 0$, which means that Gini Index is always less than Cross Entropy.

5) Programming

a. Data Inspection

How many attributes are?

The data has a total of 11 attributes in total. But two of those are not usable for classification. Those are ID and Type (which is the classification itself). The nine other attributes can be used as features to generate the feature vector.

Do you think that all attributes are meaningful for the classification? If not, explain why.

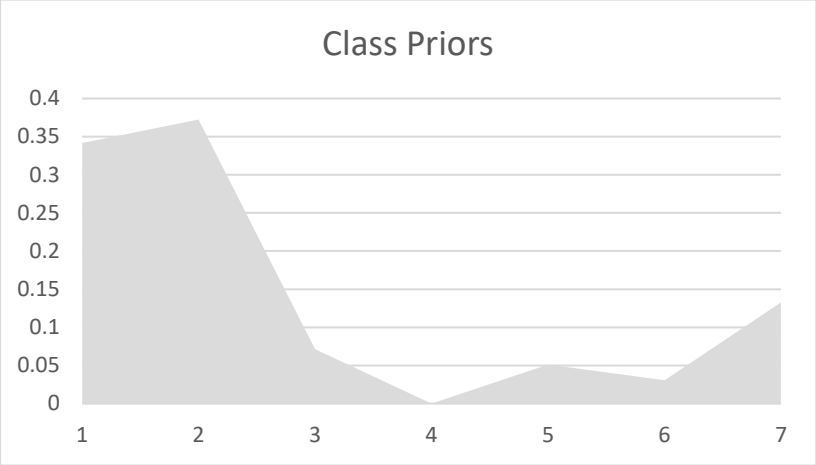
All attributes are not meaningful. The ID doesn't convey any information towards the type of the glass it is. It is just a key attribute which is usually unique and doesn't convey any information. Also Type of the glass is its label or classification. Statistically it doesn't have any information to contribute to the dataset.

The 9 other features features, can be used to train and classify the data. Some contribute less like Refractive index and Silicon and Fe (Refractive index, Si and Fe have very low variance and changes very little with each class). While the others contribute meaningful information to classify the data

How many classes are? Class is a type of a glass

There are totally 7 types of glass (as per the specification). However, Type 4 is absent in the training data set, so we effectively have only 6 classes for classification.

Please explain the class distribution. Which class is majority? Do you think that it can be considered as a uniform distribution?



As you can see from the above plot of priors of each class, class 2 is the majority in this distribution. Class 1 is second majority in this dataset. Class 4 is absent, and remaining classes are minority compared to Class 1 and Class2. This dataset can't be considered as uniform distribution. The priors for 7 classes are as below. Type4 is absent in the dataset.

Looking at the graph we can say that this can't be considered as uniform distribution. Because for ideal uniform distribution all the probabilities have to be equal. In practical consideration at least they need to be close to be considered as uniform distribution.

d. Performance Comparison:

k- Nearest Neighbors

	Testing with Training Data:
For k = 1	L1 Accuracy = 74.4897959184
	L2 Accuracy = 73.9795918367
For k = 3	L1 Accuracy = 71.9387755102
	L2 Accuracy = 68.8775510204
For k = 5	L1 Accuracy = 72.4489795918
	L2 Accuracy = 68.3673469388
For k = 7	L1 Accuracy = 68.8775510204
	L2 Accuracy = 66.8367346939
	Testing with Test Data:
For k = 1	L1 Accuracy = 61.1111111111
	L2 Accuracy = 61.1111111111
For k = 3	L1 Accuracy = 55.5555555556
	L2 Accuracy = 61.1111111111
For k = 5	L1 Accuracy = 50.0
	L2 Accuracy = 50.0
For k = 7	L1 Accuracy = 44.4444444444
	L2 Accuracy = 44.4444444444

NAIVE BAYES:
Testing with Training Data:
Accuracy = 53.5714285714
Testing with Test Data:
Accuracy = 38.8888888889

The naive bayes is a probabilistic classifier. While in this case it seems to be giving a lower accuracy against both test and training data, given a large enough data set, will more likely produce the better results. However, in this particular case, the kNN has a better accuracy in classifying data.

In both, the training data as expected has better accuracy.

In kNN, the boundaries are not very clear in classification. And can get more vague when the value of k increases.

(Name)Ravishankar Sivaraman (USC ID)6370-0913-06 (Email)rsivaram@usc.edu