AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Yusuf Brima | Assignment Number: 2

Course: Statistical Machine Learning for Data Science | Date: January 17, 2021

Exercise 1

(1) The empirical Probability of Correct Classification on the test set $\widehat{\text{PCC}}_{te}$ of the learning machine $\widehat{f}$ is thus:

$$\widehat{\text{PCC}}_{te}(\widehat{f}) = \frac{\widehat{\text{TPR}}(\widehat{f}) + \widehat{\text{TNR}}(\widehat{f})}{2}$$

Where

$$\begin{aligned}\widehat{\text{TPR}}(\widehat{f}) &= \frac{\widehat{\text{TP}}(\widehat{f})}{\widehat{\text{TP}}(\widehat{f}) + \widehat{\text{FN}}(\widehat{f})}\\ &= \frac{7}{8}\end{aligned}$$

And

$$\begin{aligned}\widehat{\text{TNR}}(\widehat{f}) &= \frac{\widehat{\text{TN}}(\widehat{f})}{\widehat{\text{TN}}(\widehat{f}) + \widehat{\text{FP}}(\widehat{f})}\\ &= \frac{7}{8}\end{aligned}$$

Therefore

$$\begin{aligned}\widehat{\text{PCC}}_{te}(\widehat{f}) &= \frac{\frac{7}{8} + \frac{7}{8}}{2}\\ &= 0.875\end{aligned}$$

Whilst the theoretical/ideal Probability of Correct Classification on the test set PCC_{te} of the learning machine $\widehat{f}$ is thus:

$$\text{PCC}_{te}(\widehat{f}) = \frac{\text{TPR}(\widehat{f}) + \text{TNR}(\widehat{f})}{2}$$

(2) $\widehat{\mathrm{R}}_{te}(\widehat{f})$ is the empirical Risk Functional on the test set $\mathscr{D}_{te}$ for the learning function $\widehat{f}$.

$$\begin{aligned}\widehat{\mathrm{R}}_{te}(\widehat{f}) &= \frac{\widehat{\mathrm{FN}}(\widehat{f}) + \widehat{\mathrm{FP}}(\widehat{f})}{\widehat{\mathrm{FN}}(\widehat{f}) + \widehat{\mathrm{FP}}(\widehat{f}) + \widehat{\mathrm{TN}}(\widehat{f}) + \widehat{\mathrm{TP}}(\widehat{f})} \\ &= \frac{2}{16}\end{aligned}$$

Whilst the true/theoretical Risk Functional $\mathrm{R}_{te}(\widehat{f})$ on the test set $\mathscr{D}_{te}$ for the learning function $\widehat{f}$ is thus:

$$\mathrm{R}_{te}(\widehat{f}) = \frac{\mathrm{FN}(\widehat{f}) + \mathrm{FP}(\widehat{f})}{\mathrm{FN}(\widehat{f}) + \mathrm{FP}(\widehat{f}) + \mathrm{TN}(\widehat{f}) + \mathrm{TP}(\widehat{f})}$$

(3) $\widehat{M_{te}}$, the confusing matrix of $\widehat{f}$

		-1	+1
y	-1	7	1
	+1	1	7

(4)

$$\begin{aligned}\frac{\mathrm{trace}(\widehat{M_{te}})}{|\mathrm{D}_{te}|} &= \frac{\widehat{\mathrm{TN}}(\widehat{f}) + \widehat{\mathrm{TP}}(\widehat{f})}{|\mathrm{D}_{te}|} \\ &= \widehat{A}(\widehat{f}) = \frac{14}{16}\end{aligned}$$

This indicates the proportion of correctly classified positive (+1) test set data, in this case, the learning machine correctly classified 87.5% of the test set correctly.

(5) The True Positive Rate is a measure of the ratio of True Positives over the Positive Ground Truth. In other words $\mathbb{P}(f(\mathrm{x}_i) = 1|\mathrm{y}_i = 1)$ .The empirical True Positive Rate $\widehat{\mathrm{TPR}}(\widehat{f})$ of the learning function on $\mathscr{D}_{te}$ is defined below :

$$\begin{aligned}\widehat{\mathrm{TPR}}(\widehat{f}) &= \frac{\widehat{\mathrm{TP}}(\widehat{f})}{\widehat{\mathrm{TP}}(\widehat{f}) + \widehat{\mathrm{FN}}(\widehat{f})} \\ &= \frac{7}{8}\end{aligned}$$

And the theoretical/ideal True Positive Rate $\mathrm{TPR}(\widehat{f})$ of the learning function on $\mathscr{D}_{te}$ is thus:

$$\mathrm{TPR}(\widehat{f}) = \frac{\mathrm{TP}(\widehat{f})}{\mathrm{TP}(\widehat{f}) + \mathrm{FN}(\widehat{f})}$$

(6) This indicates the proportion of incorrectly classified negative (-1) test set data, in this case, the learning machine correctly classified 12.5% negative (-1) of the test set incorrectly.

(5) The False Positive Rate is a measure of the ratio of False Positives over the Positive Ground Truth. In other words $\mathbb{P}(f(\mathrm{x}_i) = 1|\mathrm{y}_i = -1)$ .The empirical True Positive Rate $\widehat{\mathrm{TPR}}(\widehat{f})$ of the learning function on $\mathscr{D}_{te}$ is defined below :

$$\begin{aligned}\widehat{\mathrm{FPR}}(\widehat{f}) &= \frac{\widehat{\mathrm{FP}}(\widehat{f})}{\widehat{\mathrm{FP}}(\widehat{f}) + \widehat{\mathrm{TN}}(\widehat{f})}\\ &= \frac{1}{8}\end{aligned}$$

And the theoretical/ideal False Positive Rate $\mathrm{FPR}(\widehat{f})$ of the learning function on $\mathscr{D}_{te}$ is thus:

$$\mathrm{FPR}(\widehat{f}) = \frac{\mathrm{FP}(\widehat{f})}{\mathrm{FP}(\widehat{f}) + \mathrm{TN}(\widehat{f})}$$

(7)

$$\begin{aligned}\widehat{F}_{measure} &= \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}) = \frac{\widehat{\mathrm{TP}}(\widehat{f})}{\widehat{\mathrm{TP}}(\widehat{f}) + \frac{1}{2}(\widehat{\mathrm{FP}}(\widehat{f}) + \widehat{\mathrm{FN}}(\widehat{f}))}\\ &= \frac{7}{7 + \frac{1}{2}(1+1)}\\ &= \frac{7}{8}\end{aligned}$$

Exercise 2

We'll consider 1-NN, 2-NN and 3-NN, and use two weighting schemes:

- Uniform weighting: the weight of each member of the neighborhood is simply $\frac{1}{k}$
- Inverse distance weighting: the weight of each member of the neighborhood is

$$w_j = \frac{\frac{1}{d_j}}{\sum_{l=1}^{k} \frac{1}{d_l}} \tag{1}$$

(1) Considering 1-NN to determine $\widehat{y_{\text{new}}} = \widehat{f_{kNN}(x_{new})}$.

$$\begin{aligned}\widehat{y_{\text{new}}} &= \widehat{f_{1NN}(x_{new})}\\ &= \underset{c\in\{1,2\}}{argmax}\{\sum_{i=1}^{n} \mathbb{1}(y_i = c)\mathbb{1}(x_i \in V(x))\}\end{aligned}$$

When $c = 1$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 1|x_{new}) &= \sum_{i=1}^{n} \mathbb{1}(y_i = 1)\mathbb{1}(x_i \in V_1)\\ &= \mathbb{1}(y_1 = 1)\mathbb{1}(x_1 \in V_1) + \mathbb{1}(y_2 = 1)\mathbb{1}(x_2 \in V_1) + \mathbb{1}(y_3 = 1)\mathbb{1}(x_3 \in V_1)\\ &= (0 \times 1) + (1 \times 0) + (1 \times 0)\\ &= 0\end{aligned}$$

When $c = 2$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 2|x_{new}) &= \sum_{i=1}^{n} \mathbb{1}(y_i = 1)\mathbb{1}(x_i \in V_2)\\ &= \mathbb{1}(y_1 = 1)\mathbb{1}(x_1 \in V_2) + \mathbb{1}(y_2 = 1)\mathbb{1}(x_2 \in V_2) + \mathbb{1}(y_3 = 1)\mathbb{1}(x_3 \in V_2)\\ &= (0 \times 1) + (0 \times 1) + (1 \times 1)\\ &= 1\end{aligned}$$

Therefore, $\widehat{y_{new}} = 2 = \widehat{f_{1NN}}$

(2) Considering 2-NN.

(1) Determining $\widehat{y_{\text{new}}} = \widehat{f_{kNN}(x_{new})}$ under the uniform weighting scheme.

$$\begin{aligned}\widehat{y_{\text{new}}} &= \widehat{f_{2NN}(x_{new})}\\ &= \underset{c\in\{1,2\}}{argmax}\{\sum_{i=1}^{n} \mathbb{1}(y_i = c)\mathbb{1}(x_i \in V(x))w_i\}\end{aligned}$$

$V_2(x_{new}\{x_i, d(x_{new}, x_i) \leq d_2 = 2\})$
$\{x_1 \in V_2, x_2 \in v_2, x_3 \notin v_3\}$
When $c = 1$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 1|x_{new}) &= \frac{1}{2}\sum_{i=1}^{3} \mathbb{1}(y_i = 1)\mathbb{1}(x_i \in V_2)\\ &= \frac{1}{2}[(0 \times 1) + (1 \times 1) + (1 \times 0)]\\ &= \frac{1}{2}\end{aligned}$$

When $c = 2$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 2|x_{new}) &= \frac{1}{2}\sum_{i=1}^{3} \mathbb{1}(y_i = 2)\mathbb{1}(x_i \in V_2)\\ &= 1 - \mathbb{P}(\widehat{y_{new}} = 1|x_{new})\\ &= 1 - \frac{1}{2}\\ &= \frac{1}{2}\end{aligned}$$

(2) Determining $\widehat{y_{\text{new}}} = \widehat{f_{kNN}(x_{new})}$ under the inverse distance weighting scheme as stated in equation (1).

$$\begin{aligned}\widehat{y_{\text{new}}} &= \widehat{f_{2NN}(x_{new})}\\ &= \underset{c\in\{1,2\}}{argmax}\{\sum_{i=1}^{3}\mathbb{1}(y_i = c)\mathbb{1}(x_i \in V(x))w_i\}\end{aligned}$$

For $i = 1$

$$\begin{aligned}w_1 &= \frac{\frac{1}{d_1}}{\frac{1}{d_1}+\frac{1}{d_2}}\\ &= \frac{\frac{1}{1}}{\frac{1}{1}+\frac{1}{2}}\\ &= \frac{2}{3}\end{aligned}$$

For $i = 2$

$$\begin{aligned}w_2 &= \frac{\frac{1}{d_2}}{\frac{1}{d_1}+\frac{1}{d_2}}\\ &= \frac{1}{3}\end{aligned}$$

For $i = 3$

$$\begin{aligned}w_3 &= \frac{\frac{1}{d_3}}{\frac{1}{d_1}+\frac{1}{d_2}}\\ &= \frac{2}{15}\end{aligned}$$

When $c = 1$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 1|x_{new}) &= \sum_{i=1}^{3}\mathbb{1}(y_i = 1)\mathbb{1}(x_i \in V_2)w_i\\ &= (0\times 1)\times\frac{2}{3} + (1\times 1)\times\frac{1}{3}\\ &= \frac{1}{3}\end{aligned}$$

When $c = 2$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 2|x_{new}) &= \sum_{i=1}^{3}\mathbb{1}(y_i = 2)\mathbb{1}(x_i \in V_2)w_i\\ &= 1 - \mathbb{P}(\widehat{y_{new}} = 1|x_{new})\\ &= 1 - \frac{1}{3}\\ &= \frac{2}{3}\end{aligned}$$

Therefore, $\widehat{y_{new}} = 2 = \widehat{f_{2NN}}$

(2) Considering 3-NN.

(1) Determining $\widehat{y_{\text{new}}} = \widehat{f_{kNN}(x_{new})}$ under the uniform weighting scheme.

$$\begin{aligned}\widehat{y_{\text{new}}} &= \widehat{f_{3NN}(x_{new})} \\ &= \underset{c\in\{1,2\}}{argmax}\{\frac{1}{3}\sum_{i=1}^{n}\mathbb{1}(y_i = c)\mathbb{1}(x_i \in V(x))\}\end{aligned}$$

$V_3(x_{new}\{x_i, d(x_{new}, x_i) \leq d_3 = 3\})$
When $c = 1$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 1|x_{new}) &= \frac{1}{3}\sum_{i=1}^{3}\mathbb{1}(y_i = 1)\mathbb{1}(x_i \in V_3) \\ &= \frac{1}{3}[(0\times 1) + (1\times 1) + (1\times 1)] \\ &= \frac{2}{3}\end{aligned}$$

When $c = 2$

$$\begin{aligned}\mathbb{P}(\widehat{y_{new}} = 2|x_{new}) &= \frac{1}{2}\sum_{i=1}^{3}\mathbb{1}(y_i = 2)\mathbb{1}(x_i \in V_3) \\ &= 1 - \mathbb{P}(\widehat{y_{new}} = 1|x_{new}) \\ &= 1 - \frac{2}{3} \\ &= \frac{1}{3}\end{aligned}$$

Therefore, $\widehat{y_{new}} = 1 = \widehat{f_{3NN}}$

(2) Determining $\widehat{y_{\text{new}}} = \widehat{f_{kNN}(x_{new})}$ under the inverse distance weighting scheme as stated in equation (1).

$$\begin{aligned}\widehat{y_{\text{new}}} &= \widehat{f_{3NN}(x_{new})} \\ &= \underset{c\in\{1,2\}}{argmax}\{\sum_{i=1}^{3}\mathbb{1}(y_i = c)\mathbb{1}(x_i \in V(x))w_i\}\end{aligned}$$

For $i = 1$

$$\begin{aligned}w_1 &= \frac{\frac{1}{d_1}}{\frac{1}{d_1} + \frac{1}{d_2}} \\ &= \frac{\frac{1}{1}}{\frac{1}{1} + \frac{1}{2} + \frac{1}{5}} \\ &= \frac{10}{17}\end{aligned}$$

For $i = 2$

$$\begin{aligned} w_1 &= \frac{\frac{1}{d_2}}{\frac{1}{d_1} + \frac{1}{d_2}} \\ &= \frac{\frac{1}{2}}{\frac{1}{1} + \frac{1}{2} + \frac{1}{5}} \\ &= \frac{5}{17} \end{aligned}$$

For $i = 3$

$$\begin{aligned} w_1 &= \frac{\frac{1}{d_3}}{\frac{1}{d_1} + \frac{1}{d_2}} \\ &= \frac{\frac{1}{5}}{\frac{1}{1} + \frac{1}{2} + \frac{1}{5}} \\ &= \frac{2}{17} \end{aligned}$$

When $c = 1$

$$\begin{aligned} \mathbb{P}(\widehat{y_{new}} = 1 | x_{new}) &= \sum_{i=1}^{3} \mathbb{1}(y_i = 1)\mathbb{1}(x_i \in V_2) w_i \\ &= (0 \times 1) \times \frac{10}{17} + (1 \times 1) \times \frac{5}{17} + (1 \times 1) \times \frac{2}{17} \\ &= \frac{7}{17} \end{aligned}$$

When $c = 2$

$$\begin{aligned} \mathbb{P}(\widehat{y_{new}} = 2 | x_{new}) &= \sum_{i=1}^{3} \mathbb{1}(y_i = 2)\mathbb{1}(x_i \in V_2) w_i \\ &= 1 - \mathbb{P}(\widehat{y_{new}} = 1 | x_{new}) \\ &= 1 - \frac{7}{17} \\ &= \frac{10}{17} \end{aligned}$$

Therefore, $\widehat{y_{new}} = 2 = \widehat{f_{3NN}}$

(4) The Inverse Distance Weighting (IDW) has a superior measure of neighborhood closeness for class labels compared to uniform weighting scheme.

Exercise 3

(1) Display both your training confusion matrix and your test confusion matrix

(1) 1NN

```
         y.tr.hat
ytrain     1    7
     1  5941    0
     7     0 5437

        y.te.hat
ytest     1    7
    1  1936    0
    7    24 1832
```

(2) 7NN

```
         y.tr.hat
ytrain     1    7
     1  5926   15
     7    50 5387

        y.te.hat
ytest     1    7
    1  1936    0
    7    34 1822
```

(3) 9NN

```
         y.tr.hat
ytrain     1    7
     1  5928   13
     7    59 5378

        y.te.hat
ytest     1    7
    1  1936    0
    7    36 1820
```

(2) ROC Curves for the three models

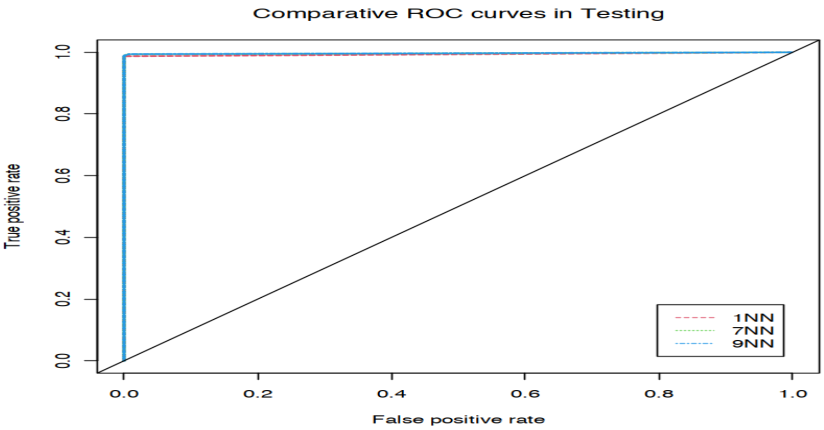

Figure 1: Comparative ROC Curves in Testing

(3) Solution in the r file

(4) From the ROC Curve, it is clear that 1NN has the highest model complexity whilst 7NN has the least model complexity. Therefore, according to Ocam's Razor, the model with the least complexity must be preferred.

(5) Solution in r file

Exercise Bonus 1

The point-wise bias variance decomposition when $\hat{f}$ is the k Nearest Neighbors regression learner.

Let $\mathscr{D}_n := \{(\mathrm{x}_i, yi) \overset{\text{iid}}{\sim} p_{\mathrm{x},y}(\mathrm{x}, y), \mathrm{x}_i \in \mathbb{R}, y_i \in \mathbb{R}\}_{i=1}^{n}$.

$$\widehat{f}_{kNN}(x) = \frac{1}{k}\sum_{i=1}^{n} y_i \mathbb{1}(x_i \in V_k(x))$$

where

$$V_k(x) := \{x_i \in \mathscr{D}_n \quad \text{s.t. } d(x, x_j) \leq d(k)\}$$

where $d(k)$ being the shortest k-th distance from the new data-point x.

Thus, for the variance of $\widehat{f}_{kNN}$:

$$\begin{aligned}\text{Variance}\left(\widehat{f}_{kNN}(x)\right) &= \mathbb{V}(\frac{1}{k}\sum_{i=1}^{n} y_i \mathbb{1}(x_i \in V_k(x)) \\ &= \frac{1}{k^2}\sum_{i=1}^{n} \mathbb{V}[y_i|x_i]\mathbb{1}(x_i \in V_k(x)) \\ &= \frac{1}{k^2}\sum_{i=1}^{n} \sigma^2 \mathbb{1}(x_i \in V_k(x)) \\ &= \frac{k\sigma^2}{k^2} \\ &= \frac{\sigma^2}{k}\end{aligned}$$

For the bias of $\widehat{f}_{kNN}$:

$$\begin{aligned}\text{Bias}\left(\widehat{f}_{kNN}(x)\right) &= \mathbb{E}[\widehat{f}_{kNN}(x)] - f(x) \\ &= \frac{1}{k}\sum_{i=1}^{n} y_i \mathbb{1}(x_i \in V_k(x)) - f(x) \\ &= \frac{1}{k}\sum_{i=1}^{n} \mathbb{E}[y_i|x_i]\mathbb{1}(x_i \in V_k(x)) - f(x)\end{aligned}$$

Given that

$$\mathbb{E}[y_i|x_i] = (fx_i)$$

Therefore

$$\begin{aligned}\text{Bias}\left(\widehat{f}_{kNN}(x)\right) &= \frac{1}{k}\sum_{i=1}^{n} f(x_i)\mathbb{1}(x_i \in V_k(x)) - f(x) \\ &= \frac{1}{k}\sum_{x_i \in V_k(x)} f(x_i) - f(x)\end{aligned}$$

Finally, the point wise decomposition of MSE $\left(\widehat{f}_{kNN}(x)\right)$ is:

$$\text{MSE}\left(\widehat{f}_{kNN}(x)\right) = \sigma^2 + \frac{\sigma^2}{k} + \left(\frac{1}{k}\sum_{x_i \in V_k(x)} f(x_i) - f(x)\right)^2$$

Exercise Bonus 2

Solution in the r file.