CMPS 242 Homework 2

Jayanth Yetukuri, Weiting Zhan, Jiajie Kong
(jyetukur, wzhan83, jkong7)@ucsc.edu

21-Oct-2018

1 Bayesian decision theory

Solution 1:
Part a. The risk(expected loss) of each of the predictions B,C,T and A are

$$E(\text{loss}|\text{pred } B) = P(B|x) * 0 + P(C|x) * 1 + P(T|x) * 1 = \frac{1}{9} * 0 + \frac{2}{3} * 1 + \frac{2}{9} * 1 = \frac{8}{9}$$

$$E(\text{loss}|\text{pred } C) = P(B|x) * 1 + P(C|x) * 0 + P(T|x) * 1 = \frac{1}{9} * 1 + \frac{2}{3} * 0 + \frac{2}{9} * 1 = \frac{3}{9}$$

$$E(\text{loss}|\text{pred } T) = P(B|x) * 1 + P(C|x) * 1 + P(T|x) * 0 = \frac{1}{9} * 1 + \frac{2}{3} * 1 + \frac{2}{9} * 0 = \frac{7}{9}$$

$$E(\text{loss}|\text{pred } A) = P(B|x) * a + P(C|x) * a + P(T|x) * a = \frac{1}{9} * a + \frac{2}{3} * a + \frac{2}{9} * a = a$$

Part b. We want to minimizing the risk, it means we need to find the minimal loss of $E(\text{loss}|\text{pred } B), E(\text{loss}|\text{pred } C)$,

$E(\text{loss}|\text{pred } T), E(\text{loss}|\text{pred } A)$. i.e. Since

$$\min\left(\frac{8}{9}, \frac{3}{9}, \frac{5}{9}, a\right) = \min\left(\frac{3}{9}, a\right)$$

So, the risk minimizing function depends on the value of a.

$$\min risk = \begin{cases} a \geq \frac{1}{3} & \text{Predict C,} \\ a < \frac{1}{3} & \text{Predict A.} \end{cases}$$

The general prediction rule should be abstaining when $a < \frac{1}{3}$ and predicting C when $a \geq \frac{1}{3}$.
When $a \to 0$, there is no risk (loss) to abstain.

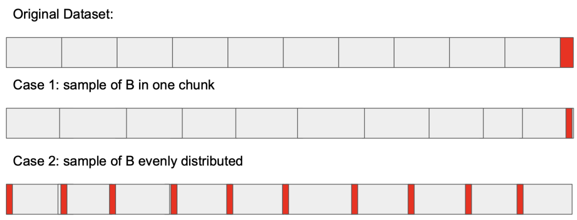

Figure 1: Illustration of the two cases considered

2 10-fold cross validation

Problem 2: Algorithm reaching 99% accuracy under this scenario is **not a big achievement**. It is given that class one have 1000 data points and class two only have 10 data points. Now consider the following two extreme grouping situations and we will calculate the error rate of the algorithm. The illustration of the data distribution can be seen from the above diagram.

Case 1. All 10 points from class II are in the same group, then this group have 10 points from class II and 91 points from class I. Under such circumstances, the error rate of the mentioned an algorithm which will only classify class I and always wrongly classifies class II will be:

$$\frac{1}{10}\left(9 \times 100\% + \frac{91}{101} \times 100\%\right) = 99.0099\% \approx 99\%$$

Case 2. All 10 points from class two are evenly distributed among 10 different groups, then each group have 1 points from class II and 100 points from class I. Under such circumstances, the accuracy of an algorithm which will only classify class I and always wrongly classifies class II will be:

$$\frac{1}{10}\left(\frac{100}{101} \times 100\%\right) = 99.0099\% \approx 99\%$$

More precisely, under such non-technical classifier, the accuracy rate will be exactly 99.0099% using ten fold cross validation, regardless of how we divide all 1010 points because it is not trained for class II. This model is as good as a rule which always spits class I irrespective of the data or model. Hence, the algorithm given by grad student is **not performing well**.

3 Gaussian Generative Model

Solution 3: To show that the *decisive boundary* is a hyperplane, it is equivalent to show that all $d \times 1$ vector $\mathbf{x}$ in decisive boundary only have $d-1$ dimension. The following equation holds for all vector $\mathbf{x}$ in decision boundary:

$$P(T=1)P(\mathbf{x}|T=1) = P(T=0)P(\mathbf{x}|T=0) \tag{1}$$

assume that

$$\begin{aligned} P(T=1) = c_1, &\quad \mathbf{x}|T=1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ P(T=0) = c_0, &\quad \mathbf{x}|T=0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \end{aligned}$$

Substitute back to (1), then

$$\begin{aligned} c_1 \frac{1}{z} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \right\} &= c_0 \frac{1}{z} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) \right\} \\ 2\log(\frac{c_1}{c_0}) &= -(\mathbf{x} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + (\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) \\ 2\log(\frac{c_1}{c_0}) &= -\mathbf{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 \end{aligned}$$

simply the above equation in the form of $\mathbf{w}^T\mathbf{x} = c$, then

$$2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} = \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - 2\log(\frac{c_1}{c_0}) \tag{2}$$

$$2\left[\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right]^T \mathbf{x} = \left\{ \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 - 2\log\left(\frac{P(T=1)}{P(T=0)} \right) \right\} \tag{3}$$

The right hand side of equation (3) is a constant. Since value of vector $\mathbf{x}$ is restrained by equation (3). Vector $\mathbf{x}$ must lose 1 degree of freedom for all the points in decisive boundary to form a **hyperplane**.

4 Naive Bayes

Solution 4: Let us begin with finding the maximum likelihood estimator of μ and σ^2 with respective to GPA, under condition **H** and **N**, respectively. For the Gaussian distribution, the MLE is given by:

$$\hat{\mu}_{MLE} = \bar{x}, \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

Apply such formula to **H** and **N**, respectively, then

$$\begin{aligned} \hat{\mu}_{MLE}|\mathbf{H} = \mu_1 = 3.4, \quad \hat{\sigma}^2_{MLE}|\mathbf{H} = \sigma_1^2 = 0.42 \\ \hat{\mu}_{MLE}|\mathbf{N} = \mu_0 = 3, \quad \hat{\sigma}^2_{MLE}|\mathbf{N} = \sigma_0^2 = 0.243 \end{aligned}$$

So the distribution of GPA will be

$$\begin{aligned} GPA|\mathbf{H} &\sim \mathcal{N}(3.4,\ 0.42) \\ GPA|\mathbf{N} &\sim \mathcal{N}(3,\ 0.243) \end{aligned}$$

Based on the table given by the problems, we also have

$$\begin{aligned} P(AP = 1|\mathbf{H}) = \frac{2}{3}, \qquad P(\mathbf{H}) = \frac{1}{3} \\ P(AP = 1|\mathbf{N}) = \frac{1}{3}, \qquad P(\mathbf{H}) = \frac{2}{3} \end{aligned}$$

Part a. if AP course are taken, based on naive Bayes rule, we predict **H** if:

$$P(AP = 1|\mathbf{H})P(GPA|\mathbf{H})P(\mathbf{H}) > P(AP = 1|\mathbf{N})P(GPA|\mathbf{N})P(\mathbf{N})$$

We must find the decision boundaries

$$\frac{2}{3}\left(2\pi\sigma_1^2\right)^{-1/2}\exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)\frac{1}{3} = \frac{1}{3}\left(2\pi\sigma_0^2\right)^{-1/2}\exp\left(-\frac{1}{2\sigma_0^2}(x-\mu_0)^2\right)\frac{2}{3}$$

$$-\log(\frac{\sigma_1^2}{\sigma_0^2}) = -\frac{1}{\sigma_0^2}(x-\mu_0)^2 + \frac{1}{\sigma_1^2}(x-\mu_1)^2$$

$$-\log(\frac{0.42}{0.243}) = -\frac{1}{0.486}(x-3)^2 + \frac{1}{0.84}(x-3.4)^2$$

we have the quadratic inequality after simplification,

$$1.734 * x^2 - 8.501 * x + 8.966 = 0$$

Solving the quadratic equation and the graphical representation of the Gaussians , we have

$$x < 1.535 \ \text{ or } \ x > 3.366$$

Part b. if AP course are not taken, based on naive Bayes rule, we predict **H** if:

$$P(AP = 0|\mathbf{H})P(GPA|\mathbf{H})P(\mathbf{H}) > P(AP = 0|\mathbf{N})P(GPA|\mathbf{N})P(\mathbf{N})$$

We must find the decision boundaries

$$\frac{1}{3}\left(2\pi\sigma_1^2\right)^{-1/2}\exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)\frac{1}{3} = \frac{2}{3}\left(2\pi\sigma_0^2\right)^{-1/2}\exp\left(-\frac{1}{2\sigma_0^2}(x-\mu_0)^2\right)\frac{2}{3}$$

Upon following the similar process as above we get a quadratic inequality:

$$1.734 * x^2 - 8.501 * x + 6.193 = 0$$

Solving the quadratic equation and the graphical representation of the Gaussians , we have

$$x < 0.890 \text{ or } x > 4.012$$

If AP courses are taken, then predict H if the GPA is less than **1.535** and if GPA is greater than **3.366**

if AP courses are not taken, then predict H if the GPA is less than **0.890** and if GPA is greater than **4.012**

5 Nearest Neighbour

Solution 5. Part a. Find the expected error rate of $3-NN$ Algorithm. Assume that x_{n+1} is the point we want to make prediction and x_i, x_j, x_k are the three nearest points of x_{n+1}. Based on the question, x_i, x_j, x_k are independent and 2/3 chance to be label as "+" and 1/3 chance to be labeled as "-", then we have the following table based on $3-NN$ rules:

Table 1: Output based on 3-NN Rules

x_i	x_j	x_k	**output**	**Probability**
+	+	+	+	$(\frac{2}{3})^3$
-	+	+	+	$(\frac{2}{3})^2(\frac{1}{3})$
+	-	+	+	$(\frac{2}{3})^2(\frac{1}{3})$
+	+	-	+	$(\frac{2}{3})^2(\frac{1}{3})$
-	+	+	+	$(\frac{2}{3})^2(\frac{1}{3})$
-	-	+	+	$(\frac{2}{3})(\frac{1}{3})^2$
-	+	-	+	$(\frac{2}{3})(\frac{1}{3})^2$
+	-	-	+	$(\frac{2}{3})(\frac{1}{3})^2$
-	-	-	+	$(\frac{1}{3})^3$

Then we have

$$P(\text{output='+'}) = (\frac{2}{3})^3 + 3 \times (\frac{2}{3})^2(\frac{1}{3}) = \frac{20}{27} \tag{4}$$

$$P(\text{output='-'}) = (\frac{1}{3})^3 + 3 \times (\frac{1}{3})^2(\frac{2}{3}) = \frac{7}{27} \tag{5}$$

So the expected error rate is

$$\begin{aligned} E\,(\text{error rate}) &= P(\text{output='+'})P(Y_{n+1} = \text{'-'}) + P(\text{output='-'})P(Y_{n+1} = \text{'+'}) \\ &= \frac{20}{27} \times \frac{1}{3} + \frac{7}{27} \times \frac{2}{3} \\ &= \frac{34}{81} \end{aligned}$$

So the expected error rate of $3-NN$ algorithm under such distribution setup is **0.4197**.

Part b. We perform the calculations with noise rate of 0.1

Table 2: Output based on 3-NN Rules

x_i	x_j	x_k	**output**	**Probability**
+	+	+	+	$(\frac{9}{10})^3$
-	+	+	+	$(\frac{9}{10})^2(\frac{1}{10})$
+	-	+	+	$(\frac{9}{10})^2(\frac{1}{10})$
+	+	-	+	$(\frac{9}{10})^2(\frac{1}{10})$
-	+	+	-	$(\frac{9}{10})^2(\frac{1}{10})$
-	-	+	-	$(\frac{9}{10})(\frac{1}{10})^2$
-	+	-	-	$(\frac{9}{10})(\frac{1}{10})^2$
+	-	-	-	$(\frac{9}{10})(\frac{1}{10})^2$
-	-	-	-	$(\frac{1}{10})^3$

Then we have

$$P(\text{output='+'}) = (\frac{9}{10})^3 + 3 \times (\frac{9}{10})^2(\frac{1}{10}) = \frac{972}{1000} \tag{6}$$

$$P(\text{output='-'}) = (\frac{1}{10})^3 + 3 \times (\frac{1}{10})^2(\frac{9}{10}) = \frac{28}{1000} \tag{7}$$

So the expected error rate is

$$\begin{aligned}
E\,(\text{error rate}) &= P(\text{output=`+'})P(Y_{n+1} = \text{`-'}) + P(\text{output=`-'})P(Y_{n+1} = \text{`+'}) \\
&= \frac{972}{1000} \times \frac{1}{10} + \frac{28}{1000} \times \frac{9}{10} \\
&= \frac{1224}{10000}
\end{aligned}$$

So the expected error rate of $3 - NN$ algorithm under such distribution setup is **0.1224**.