Homework 02 — Part I — amc1354 & ads798

1. **Exercise 1**

   (a) The True Positive Rate is: $\frac{\#TP}{\#TP+\#FN} = \frac{0}{0+3} = 0.$

   (b) The False Positive Rate is: $\frac{\#FP}{\#FP+\#TN} = \frac{1}{1+2} = 33.33\%.$

   (c) The Accuracy is: $\frac{\#TP+\#TN}{\#TP+\#FN+\#FP+\#TN} = \frac{56+41}{56+2+1+41} = 97\%.$

2. **Exercise 2**
   Let us call the decision function $f(x_1, x_2) = f(g_1(x_1, x_2), g_2(x_1, x_2))$. We know that:

   $$f(x_1,x_2) = \left\{ \begin{array}{ll} C_1 & \text{if} \quad g_1 > g_2 \\ C_2 & \text{if} \quad g_1 \leq g_2 \end{array} \right.$$
   $$= \left\{ \begin{array}{ll} C_1 & \text{if} \quad g_1 - g_2 > 0 \\ C_2 & \text{if} \quad g_1 - g_2 \leq 0 \end{array} \right. ,$$

   where g_i is short for $g_i(x_1, x_2)$.
   Now, defining $g(x_1, x_2) = g_1 - g_2$,

   $$g(x_1, x_2) = 5x_2 + 3x_1 - 4 + 3x_2 - 2x_1 + 6 = 8x_2 + x_1 + 2,$$

   we have a single discriminant function $g(x_1, x_2) = w_2x_2 + w_1x_1 + w_0$ s.t.

   $$f(x_1,x_2) = \left\{ \begin{array}{ll} C_1 & \text{if} \quad g > 0 \\ C_2 & \text{if} \quad g \leq 0 \end{array} \right. ,$$

   and $w_2 = 8$, $w_1 = 1$ and $w_0 = 2$.

3. **Exercise 3**
   Naming the positive class *spam*, we know that the the cost of a positive prediction is $a \cdot P(C = +|X)$. Now, we have $a = 5$ and $b = 2$, and a formula for $P(C = +|X)$, namely:

   $$P(C = spam|x_1, x_2) = \frac{1}{1+e^{-(3x_2-2x_1+2)}}.$$

   So, for $x_1 = 3$ and $x_2 = 2$,

   $$P(C = +|X) = \frac{1}{1+e^{-(3\cdot2-2\cdot3+1)}} = 0.2689414213699951.$$

   $$P(C = -|X) = 1 - P(C = -|X) = 1 - 0.2689414213699951 = 0.7310585786300049.$$

   Now, a false negative prediction (predict not-spam when it is spam) costs 5 units and a false positive prediction (predict spam when it is not-spam) costs 2 units. Hence the expected risk for the not-spam classification is

   $$5 \cdot P(C = +|X) = 5 \cdot 0.2689414213699951 = 1.3447071068499756$$

and the expected risk for the spam classification is

$2 \cdot P(C = -|X) = 2 \cdot 0.7310585786300049 = 1.4621171572600098.$

Therefore, we must conclude that classifying not-spam has smaller expected risk.

4. **Exercise 4**

   (a) The bias $b_\theta(\widehat{\mu})$ of the estimator $\widehat{\mu}$ of the mean μ is:

   $$= \mathbf{E}\left[\frac{\sum x}{N+1}\right] - \mu =$$
   $$= \frac{\sum \mathbf{E}[x]}{N+1} - \mu =$$
   $$= \frac{N \cdot \mu}{N+1} - \mu =$$
   $$= \mu \cdot \left[\frac{N}{N+1} - 1\right] =$$
   $$= -\frac{\mu}{N+1}.$$

   (b) If the distribution is exponential with parameter λ, and $\mathbf{E}[x] = \mu = \dfrac{1}{\lambda}$. Hence, the bias is:

   $b_\theta \widehat{\mu} = \mathbf{E}[\widehat{\mu}] - \mu =$

   $=$ ... same passages as above

   $$= -\frac{\mu}{N+1} = -\frac{1}{N+1} \cdot \frac{1}{\lambda}.$$

5. **Exercise 5**

   (a) Let $\mathbf{X}^+ \sim N(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+)$ denote the matrix of examples with positive class (label "+") $\mathbf{X}^+$, which is normally distributed with parameters mean $\boldsymbol{\mu}^+ = [\mu_1^+, \mu_2^+]^T$ and covariance matrix $\boldsymbol{\Sigma}$, where μ_1^+ is the mean of the examples x_i, $i = 1, 2$.

   For the positive class, we have:

   $$\hat{\boldsymbol{\mu}}^+ = \begin{bmatrix} \hat{\mu}_1^+ \\ \hat{\mu}_2^+ \end{bmatrix} = \begin{bmatrix} \frac{2.7+3.2-0.4}{3} \\ \frac{4.8+5.1-0.3}{3} \end{bmatrix} = \begin{bmatrix} 1.833333 \\ 3.2 \end{bmatrix}, \text{ and}$$

$$\hat{\mathbf{\Sigma}}^{+} = \begin{bmatrix} 3.803333 & 0 \\ 0 & 9.21 \end{bmatrix}.$$

Where the computations for each element of $\hat{\mathbf{\Sigma}}^{+}_{1,1}$ are:

$\hat{\Sigma}^{+}_{1,1} = \frac{(2.7-1.83)^2+(3.2-1.83)^2+(-0.4-1.83)^2}{3} = 3.803333$
$\hat{\Sigma}^{+}_{1,2} = \hat{\Sigma}^{+}_{2,1} = 0$ as for Gaussian Naive Bayes we shall assume independence of x_1, x_2.

$\hat{\Sigma}^{+}_{2,2} = \frac{(4.8-3.2)^2+(5.1-3.2)^2-(0.3-3.2)^2}{3} = 9.21.$

For the negative class, we have:

$$\hat{\boldsymbol{\mu}}^{-} = \begin{bmatrix} \hat{\mu}^{-}_1 \\ \hat{\mu}^{-}_2 \end{bmatrix} = \begin{bmatrix} \frac{0.6+1.8+2.1}{3} \\ \frac{0.5+2.8+4.3}{3} \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2.53333 \end{bmatrix}, \text{ and}$$

$$\hat{\mathbf{\Sigma}}^{-} = \begin{bmatrix} 0.63 & 0 \\ 0 & 3.663333 \end{bmatrix}.$$

Where the computations for each element of $\hat{\mathbf{\Sigma}}^{-}_{1,1}$ are:

$\hat{\Sigma}^{-}_{1,1} = \frac{(0.6-1.5)^2+(1.8-1.5)^2+(2.1-1.5)^2}{3} = 0.63$
$\hat{\Sigma}^{-}_{1,2} = \hat{\Sigma}^{-}_{2,1} = 0$ as for Gaussian Naive Bayes we shall assume independence of x_1, x_2.

$\hat{\Sigma}^{-}_{2,2} = \frac{(0.5-2.53)^2+(2.8-2.53)^2-(4.3-2.53)^2}{3} = 3.663333.$

(b) Using the results from (a), we have for the positive class:

$$p(\mathbf{x}|+) = \frac{1}{2\pi \begin{vmatrix} 3.8 & 0 \\ 0 & 9.21 \end{vmatrix}^{\frac{1}{2}}} exp\left(-\frac{1}{2}\begin{bmatrix} 1.6-1.83 \\ 2.3-3.2 \end{bmatrix}^T \begin{bmatrix} 3.8 & 0 \\ 0 & 9.21 \end{bmatrix}^{-1} \begin{bmatrix} 1.6-1.83 \\ 2.3-3.2 \end{bmatrix}\right)$$
$$= \frac{1}{2\pi(3.8 \cdot 9.21)^{\frac{1}{2}}} exp\left(-\frac{1}{2}\begin{bmatrix} 1.6-1.83 \\ 2.3-3.2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{3.8} & 0 \\ 0 & \frac{1}{9.21} \end{bmatrix} \begin{bmatrix} 1.6-1.83 \\ 2.3-3.2 \end{bmatrix}\right)$$
$$= \frac{1}{2\pi(3.8 \cdot 9.21)^{\frac{1}{2}}} exp\left(-\frac{1}{2 \cdot 3.8}(1.6-1.83)^2 - \frac{1}{2 \cdot 9.21}(2.3-3.2)^2\right)$$
$$= \frac{1}{\sqrt{2\pi \cdot 3.8}} exp\left(-\frac{1}{2 \cdot 3.8}(1.6-1.83)^2\right) \frac{1}{\sqrt{2\pi \cdot 9.21}} exp\left(-\frac{1}{2 \cdot 9.21}(2.3-3.2)^2\right).$$

Taking the *log*, we have:

$log(p(\mathbf{x}|+)) =$

$- \, log(\sqrt{2\pi \cdot 3.8}) - \frac{1}{2 \cdot 3.8}(1.6 - 1.83)^2 - log(\sqrt{2\pi \cdot 9.21}) - \frac{1}{2 \cdot 9.21}(2.3 - 3.2)^2 =$ $-3.666457.//$
Repeating the same calculations for the negative class,

$$p(\mathbf{x}|-) = \frac{1}{2\pi \begin{vmatrix} 0.63 & 0 \\ 0 & 3.66 \end{vmatrix}^{\frac{1}{2}}} exp\left(-\frac{1}{2} \begin{bmatrix} 1.6 - 1.5 \\ 2.3 - 2.53 \end{bmatrix}^T \begin{bmatrix} 0.63 & 0 \\ 0 & 3.66 \end{bmatrix}^{-1} \begin{bmatrix} 1.6 - 1.5 \\ 2.3 - 2.53 \end{bmatrix} \right)$$

Taking the log, we have:

$log(p(\mathbf{x}|-)) =$

$- \, log(\sqrt{2\pi \cdot 0.63}) - \frac{1}{2 \cdot 0.63}(1.6 - 1.5)^2 - log(\sqrt{2\pi \cdot 3.66}) - \frac{1}{2 \cdot 3.66}(2.3 - 2.53)^2$ $= -2.273982.$

As $log(p(\mathbf{x}|-))$ is grater than $log(p(\mathbf{x}|+))$, the ML prediction is the negative class.

(c) A couple of good reason might be a) had we had more data, it might be computationally less expensive; b) with few data points (like here), it may give more more robust parameter estimate as we double the size of the data used for the estimation.

One bad reason is that we are violating a key assumption for the Naive Bayes model - it would be like stating almost the same random process generated classes "-" and "+" and we may throw away discriminative information.

6. **Exercise 6**

(a) From the closed-form formula presented in class, we have:

$\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$, where

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \sum_i r_i \\ \sum_i r_i x_i \end{bmatrix}, i = 1, ..., N.$$

So,

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i r_i \\ \sum_i r_i x_i \end{bmatrix}.$$

If $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$,

$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. Hence,

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \frac{1}{N\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{bmatrix} \begin{bmatrix} \sum_i r_i \\ \sum_i r_i x_i \end{bmatrix}.$$

Solving for w_0 and w_1,

$$w_0 = \frac{\sum_i x_i^2 \sum_i r_i - \sum_i x_i \sum_i r_i x_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} \text{ and}$$

$$w_1 = \frac{-\sum_i x_i \sum_i r_i + N\sum_i r_i x_i}{N\sum_i x_i^2 - (\sum_i x_i)^2}.$$

Now, let's plug the data in all components of these formulae:

$\sum_i x_i^2 = 1,050^2 + 428^2 + ... + 700^2 = 6500115,$

$\sum_i r_i = 57 + 28 + ... + 46 = 527,$

$\sum_i x_i = 1,050 + 428 + ... + 700 = 7221,$

$\sum_i r_i x_i = 57 \cdot 1,050 + 28 \cdot 428 + ... + 46 \cdot 700 = 477908,$

$(\sum_i x_i)^2 = (1,050 + 428 + ... + 700)^2 = 52142841.$

$N = 10.$

Finally,

$$w_0 = \frac{6500115 \cdot 527 - 7221c\dot{4}77908}{10 \cdot 6500115 - 52142841} = -1.97639 \text{ and}$$

$$w_1 = \frac{-7221 \cdot 527 + 10 \cdot 477908)}{10 \cdot 6500115 - 52142841} = 0.07572$$

(b) $y = w_0 + w_1 x = -1.97639 + 0.07572 \cdot 475 = 33.99061$. We predict a 475 feet tall building has 34 floors.

7. **Exercise 7**

(a) $x_2^{min} = -1$

$x_2^{max} = -51$

$x_2^{max} - x_2^{min} = 51 - (-1) = 52$

The scaled column x_2 is then calculated using $\frac{x - x^{min}}{x^{max} - x^{min}}$. E.g., for the first item, 42, we have $\frac{42-(-1)}{52} = 0.83$.

For all the whole dataset, $x_2 = [0.83, 1.00, 0.00, 0.08, 0.52, 0.81]^T$, rounded at two decimal places.

(b) First we scale the new example:

$$x_1^{min} = -0.3$$

$$x_1^{max} = -3.8$$

$$x_1^{max} - x_1^{min} = 3.8 - (-0.3) = 4.1$$

For x_2 we use the result in (a). The new example, scaled, is:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T = \begin{bmatrix} \frac{3.9-(-0.3)}{4.1} \\ \frac{4-(-1)}{52} \end{bmatrix}^T = \begin{bmatrix} 1.02 \\ 0.10 \end{bmatrix}^T$$, rounded at two decimal places.

We then calculate the euclidean distance between the new example $\mathbf{x}$ and each pair of examples in the data set (each row). For example, for the first row:

$Dist = \sqrt{(x_1[1] - x_1)^2 + (x_2[1] - x_2)^2} = \sqrt{(0.68 - 1.02)^2 + (0.83[1] - 0.10)^2} = 0.80$.

The vector of distances between the new example and each pair in the dataset and corresponding classes is (all numbers are rounded at two decimal places) :

$$\begin{bmatrix} Dist & Class \\ 0.80 & + \\ 0.90 & + \\ 1.03 & + \\ 0.78 & - \\ \mathbf{0.70} & - \\ 0.81 & - \end{bmatrix}.$$

In bold we have the shortest distance. Because $k = 1$, we take the label of the first nearest point as the prediction, hence the predicted label for the example would be ”-”.