Homework I

Statistical Machine Learning

CSE 575

Swetaswini Nayak

1. (a) X and Y are independent events.
   $p(Y) > 0,\ p(X) = 0.5$

   $p(X|Y) = \frac{p(X,Y)}{p(Y)} = \frac{p(X)\ p(Y)}{p(Y)} = p(X) = 0.5$

   $\therefore \mathbf{p(X|Y) = 0.5}$

   (b) X and Y are disjoint events, i.e. $p(X,Y) = 0$
   $p(Y) > 0$
   $p(X|Y) = \frac{p(X,Y)}{p(Y)} = 0$

   $\therefore \mathbf{p(X|Y) = 0}$

   (c) Tossing 2 coins C_1, C_2
   $p(C_1 = H) = 0.6 \implies p(C_1 = T) = 1 - 0.6 = 0.4$
   $p(C_2 = H) = 0.4 \implies p(C_2 = T) = 1 - 0.4 = 0.6$

   $p(HT) = p(C_1 = H)\ p(C_2 = T) = 0.6 * 0.6 = 0.36$
   $p(TT) = p(C_1 = T)\ p(C_2 = T) = 0.6 * 0.4 = 0.24$
   Probability to observe HT, HT, TT, TT :
   $p(HT) * p(HT) * p(TT) * p(TT) = 0.36 * 0.36 * 0.24 * 0.24 = 0.00746496$

   $\therefore$ **Probability to observe HT, HT, TT, TT = 0.00746496**

   (d) Coin is tossed 20 times.
   Number of heads = 15
   Number of tails = 5

   The likelihood $p(X|\theta) = \theta^5\ (1-\theta)^5$

   The MLE estimation of the coin toss is:

   $$\theta_{ML} = \frac{15}{20} = 0.75$$

   $\therefore$ **The best estimate of the probability θ of having heads-up = 0.75**

   (e) The estimated probability = θ_{ML}
   Let the true value of coin with heads-up = θ^*
   As per Hoeffding's inequality:

   $$p(|\theta_{ML} - \theta^*| \geq \epsilon)\ \leq\ 2exp\{-2N\epsilon^2\}$$

   $$\implies p(|\theta_{ML} - \theta^*| < \epsilon)\ \geq\ 1 - 2exp\{-2N\epsilon^2\}$$

   To be at least 99% sure that the difference between θ_{ML} and θ^* is no more than ϵ:

   $$p(|\theta_{ML} - \theta^*| < \epsilon) \geq 0.99$$

$$1 - 2exp\{-2N\epsilon^2\} \geq 0.99$$

Here $\epsilon = 0.1$

$$1 - 2exp\{-2N * 0.01\} \geq 0.99$$
$$2exp\{-2N * 0.01\} \leq 0.01$$
$$exp\{-2N * 0.01\} \leq \frac{1}{200}$$
$$-2N * 0.01 \leq -ln200$$
$$N \geq 50 * ln200 \approx 264.9$$

$\therefore$ **The minimum number of tosses = 265**

2. **Discriminant Linear Classifiers**

   Given training data set x_n, t_n of size $N = 21$
   Two target classes C_1 and C_2.

   (a) **least-square linear classifier**
   $t_n = [0,1]^T \; for \; x_n \in C_1$
   $t_n = [1,0]^T \; for \; x_n \in C_2$

   $y(x) = \widetilde{W}^T \widetilde{X}$,
   Here, K = 2, D = 2

   Error Function:

$$E_D(\widetilde{W}) = \frac{1}{2}\sum_{n=1}^{N} ||y(x_n) - t_n||^2$$

$$\widetilde{W} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{T}$$

   $\widetilde{X} = [1 \; X]^T$. $\widetilde{X}$ is a $N * (D+1), \; i.e. \; 21 \; * \; 3$ vector

   $\widetilde{T}$ is a $N * K, \; i.e. \; 21 \; * \; 2$ vector

   $\implies \widetilde{W}$ is a $(D+1) * K, \; i.e. \; 3 \; * \; 2$ vector

$$\widetilde{X}^T\widetilde{X} = \begin{bmatrix} 21 & 345 & 33 \\ 345 & 30165 & 681 \\ 33 & 681 & 71 \end{bmatrix}$$

$$\widetilde{\mathbf{W}} = \begin{bmatrix} \frac{\mathbf{503483}}{\mathbf{393456}} & \frac{\mathbf{-110027}}{\mathbf{393456}} \\ \frac{\mathbf{-27}}{\mathbf{131152}} & \frac{\mathbf{27}}{\mathbf{131152}} \\ \frac{\mathbf{-19477}}{\mathbf{65576}} & \frac{\mathbf{19477}}{\mathbf{65576}} \end{bmatrix} = \begin{bmatrix} \mathbf{1.2796} & \mathbf{-0.2796} \\ \mathbf{-0.0002} & \mathbf{0.0002} \\ \mathbf{-0.297} & \mathbf{0.297} \end{bmatrix}$$

(b) **Fisher's linear discriminant**

Between class covariance matrix: $S_B = (m_1 - m_2)(m_1 - m_2)^T$
Within-class covariance matrix:
$S_W = \sum_{n \in C_1}(x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2}(x_n - m_2)(x_n - m_2)^T$

Maximization of Fisher criterion:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Fisher's linear discriminant:

$$w \propto S_w^{-1}(m_2 - m_1)$$

$\mathbf{w = (w_0, w_1)^T = (0.0001, 0.194)^T}$

3. **Continuous Bayes Classifier**

Bayes classifier for a binary classification task with 2 classes: (y = 1 or y = 2)
Given:
prior: p(y = 1) = 0.6

$$p(x|y=1) = \begin{cases} 0.5, & 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

$$p(x|y=2) = \begin{cases} 0.125, & 0 \leq x \leq 8 \\ 0, & \text{otherwise} \end{cases}$$

(a) Prior for class label y = 2 :
$p(y=2) \ = \ 1 \ - \ p(y=1) \ = \ 1 \ - \ 0.6 \ = \ 0.4$
$\therefore \mathbf{p(y=2) \ = \ 0.4}$

(b) Prior for class label y = 1 given x :

$$p(y=1|x) \ = \ \frac{p(x|y=1) \ p(y=1)}{p(x)}$$

$$= \ \frac{p(x|y=1) \ p(y=1)}{p(x|y=1) \ p(y=1) \ + \ p(x|y=2) \ p(y=2)}$$

For $0 \leq x \leq 2$,

$$p(y=1|x) \ = \ \frac{0.5 * 0.6}{0.5 * 0.6 \ + \ 0.125 * 0.4} \ = \ \frac{6}{7}$$

So, for $2 < x \leq 8$,

$$p(y=1|x) \;=\; \frac{0*0.6}{0*0.6 \;+\; 0.125*0.4} \;=\; 0$$

$$\therefore \mathbf{p(y=1|x)} = \begin{cases} \mathbf{\frac{6}{7}}, & \mathbf{0 \leq x \leq 2} \\ \\ \mathbf{0}, & \mathbf{2 < x \leq 8} \end{cases}$$

(c) For $x = 1$, $p(y=1|x=1) = \frac{6}{7}$
$p(y=2|x=1) \;= 1 - p(y=1|x=1) \;=\; 1 - \frac{6}{7} \;=\; \frac{1}{7}$

Here, $p(y=1|x=1) > p(y=2|x=1)$

$\implies$ **The Bayes classifier will assign $\mathbf{y = 1}$ to $\mathbf{x = 1}$**

$\therefore$ **And the risk of this decision $=$ $\mathbf{p(mistake) = p(y=2|x=1) = \frac{1}{7}}$**

(d) $p(y=2|x) \;= 1 - p(y=1|x)$

$$\mathbf{p(y=1|x)} = \begin{cases} \mathbf{\frac{6}{7}}, & \mathbf{0 \leq x \leq 2} \\ \\ \mathbf{0}, & \mathbf{2 < x \leq 8} \end{cases}$$

$$\therefore \mathbf{p(y=2|x)} = \begin{cases} \mathbf{\frac{1}{7}}, & \mathbf{0 \leq x \leq 2} \\ \\ \mathbf{1}, & \mathbf{2 < x \leq 8} \end{cases}$$

For $0 \leq x \leq 2$,

$$p(y=1|x) > p(y=2|x) \implies \text{Assign } y = 1 \text{ to } x$$

For $2 < x \leq 8$,

$$p(y=2|x) > p(y=1|x) \implies \text{Assign } y = 2 \text{ to } x$$

For $x < 0$ or $x > 8$,
The marginal density can be found from the class conditional densities.

$$p(x) = p(x|y=1)p(y=1) + p(x|y=2)p(y=2)$$

However, the class conditional densities, $p(x|y=1) = 0 \; and \; p(x|y=2) = 0$
$\implies$ $p(x) = 0$. So, we can't do any posterior probability.

We can compare $p(x|y=1) * p(y=1)$ to $p(x|y=2) * p(y=2)$ (i.e likelihood * prior)
As, $p(x|y=1) = 0 \; and \; p(x|y=2) = 0$
$\implies$ Assign either $y = 1$ or $y = 2$ to x

$$\therefore \textbf{The decision regions are: } \begin{cases} \mathbf{y = 1}, & \mathbf{0 \leq x \leq 2} \\ \mathbf{y = 2}, & \mathbf{2 < x \leq 8} \\ \mathbf{y = 1} \text{ or } \mathbf{2}, & \text{Otherwise} \end{cases}$$

4. **Discrete Bayes Classifier**

   Bayes classifier for a binary classification task (y= 1 or y = 2) with input feature x of two binary features (x_1 *and* x_2)
   Given:
   $p(y = 1) = 0.6$

   $p(x_1 = 0, x_2 = 0 \mid y = 1) = 0.3$
   $p(x_1 = 0, x_2 = 1 \mid y = 1) = 0.1$
   $p(x_1 = 1, x_2 = 0 \mid y = 1) = 0.4$
   $p(x_1 = 1, x_2 = 1 \mid y = 1) = 0.2$

   $p(x_1 = 0, x_2 = 0 \mid y = 2) = 0.4$
   $p(x_1 = 0, x_2 = 1 \mid y = 2) = 0.3$
   $p(x_1 = 1, x_2 = 0 \mid y = 2) = 0.2$
   $p(x_1 = 1, x_2 = 1 \mid y = 2) = 0.1$

   (a) Prior for class label y = 2:

   $$p(y = 2) = 1 - p(y = 1)$$
   $$= 1 - 0.6 = 0.4$$

   $\therefore \mathbf{p(y = 2) = 0.4}$

   (b) To find $p(y = 1|x)$, we need to find the posterior probability for each possible x values.

   $$\text{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x \in \{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \}$$

   So, we need to find:
   1. $p(y = 1|x_1 = 0, x_2 = 0)$ 2. $p(y = 1|x_1 = 0, x_2 = 1)$
   3. $p(y = 1|x_1 = 1, x_2 = 0)$ 4. $p(y = 1|x_1 = 1, x_2 = 1)$

   - $p(y = 1|x_1 = 0, x_2 = 0) =$

     $$\frac{p(x_1 = 0, x_2 = 0|\ y = 1)\ p(y = 1)}{p(x_1 = 0, x_2 = 0)}$$

     Here,
     $$p(x_1 = 0, x_2 = 0) = p(x_1 = 0, x_2 = 0|\ y = 1)\ p(y = 1) + p(x_1 = 0, x_2 = 0|\ y = 2)\ p(y = 2)$$

Therefore,
$p(y=1|x_1=0,x_2=0) =$

$$\frac{p(x_1=0,x_2=0|\ y=1)\ p(y=1)}{p(x_1=0,x_2=0|\ y=1)\ p(y=1)+p(x_1=0,x_2=0|\ y=2)\ p(y=2)}$$

$$= \frac{0.3*0.6}{0.3*0.6\ +\ 0.4*0.4} = \frac{9}{17}$$

- $p(y=1|x_1=0,x_2=1) =$

$$\frac{p(x_1=0,x_2=1|\ y=1)\ p(y=1)}{p(x_1=0,x_2=1)}$$

$$\frac{p(x_1=0,x_2=1|\ y=1)\ p(y=1)}{p(x_1=0,x_2=1|\ y=1)\ p(y=1)+p(x_1=0,x_2=1|\ y=2)\ p(y=2)}$$

$$= \frac{0.1*0.6}{0.1*0.6\ +\ 0.3*0.4} = \frac{1}{3}$$

- $p(y=1|x_1=1,x_2=0) =$

$$\frac{p(x_1=1,x_2=0|\ y=1)\ p(y=1)}{p(x_1=1,x_2=0)}$$

$$\frac{p(x_1=1,x_2=0|\ y=1)\ p(y=1)}{p(x_1=1,x_2=0|\ y=1)\ p(y=1)+p(x_1=1,x_2=0|\ y=2)\ p(y=2)}$$

$$= \frac{0.4*0.6}{0.4*0.6\ +\ 0.2*0.4} = \frac{3}{4}$$

- $p(y=1|x_1=1,x_2=1) =$

$$\frac{p(x_1=1,x_2=1|\ y=1)\ p(y=1)}{p(x_1=1,x_2=1)}$$

$$\frac{p(x_1=1,x_2=1|\ y=1)\ p(y=1)}{p(x_1=1,x_2=1|\ y=1)\ p(y=1)+p(x_1=1,x_2=1|\ y=2)\ p(y=2)}$$

$$= \frac{0.2*0.6}{0.2*0.6\ +\ 0.1*0.4} = \frac{3}{4}$$

$\therefore \mathbf{p(y=1|x)}:$

$\mathbf{p(y=1|x_1=0,x_2=0)=\frac{9}{17}}$ $\mathbf{p(y=1|x_1=0,x_2=1)=\frac{1}{3}}$

$\mathbf{p(y=1|x_1=1,x_2=0)=\frac{3}{4}}$ $\mathbf{p(y=1|x_1=1,x_2=1)=\frac{3}{4}}$

(c) For $x_1 = 0$ *and* $x_2 = 1$, the posterior probabilities are

$p(y = 1|x_1 = 0, x_2 = 1) = \frac{1}{3}$
$p(y = 2|x_1 = 0, x_2 = 1) = 1 - p(y = 1|x_1 = 0, x_2 = 1) = 1 - \frac{1}{3} = \frac{2}{3}$
Here,
$(y = 2|x_1 = 0, x_2 = 1) > p(y = 1|x_1 = 0, x_2 = 1)$

$\implies$ **The classifier will assign label y = 2 to $\mathbf{x_1 = 0}$ and $\mathbf{x_2 = 1}$**

$\therefore$ **The risk of the decision = $\mathbf{p(mistake) = p(y = 1|x_1 = 0, x_2 = 1) = \frac{1}{3}}$**

(d) There are 4 possible values of x

$$x \in \{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\}$$

- $p(y = 1|x_1 = 0, x_2 = 0) = \frac{9}{17}$
  $p(y = 2|x_1 = 0, x_2 = 0) = 1 - \frac{9}{17} = \frac{8}{17}$
  $p(y = 1|x_1 = 0, x_2 = 0) \;>\; p(y = 2|x_1 = 0, x_2 = 0)$
  $\implies$ Assign y = 1 to $x_1 = 0, x_2 = 0$

- $p(y = 1|x_1 = 0, x_2 = 1) = \frac{1}{3}$
  $p(y = 2|x_1 = 0, x_2 = 1) = 1 - \frac{1}{3} = \frac{2}{3}$
  $p(y = 2|x_1 = 0, x_2 = 1) \;>\; p(y = 1|x_1 = 0, x_2 = 1)$
  $\implies$ Assign y = 2 to $x_1 = 0, x_2 = 1$

- $p(y = 1|x_1 = 1, x_2 = 0) = \frac{3}{4}$
  $p(y = 2|x_1 = 1, x_2 = 0) = 1 - \frac{3}{4} = \frac{1}{4}$
  $p(y = 1|x_1 = 1, x_2 = 0) \;>\; p(y = 2|x_1 = 1, x_2 = 0)$
  $\implies$ Assign y = 1 to $x_1 = 1, x_2 = 0$

- $p(y = 1|x_1 = 1, x_2 = 1) = \frac{3}{4}$
  $p(y = 2|x_1 = 1, x_2 = 1) = 1 - \frac{3}{4} = \frac{1}{4}$
  $p(y = 1|x_1 = 1, x_2 = 1) \;>\; p(y = 2|x_1 = 1, x_2 = 1)$
  $\implies$ Assign y = 1 to $x_1 = 1, x_2 = 1$

$$\therefore \textbf{The decision regions are: } \begin{cases} \mathbf{y = 1}, & for\ \mathbf{x_1 = 0},\ \mathbf{x_2 = 0} \\ \mathbf{y = 2}, & for\ \mathbf{x_1 = 0},\ \mathbf{x_2 = 1} \\ \mathbf{y = 1}, & for\ \mathbf{x_1 = 1},\ \mathbf{x_2 = 0} \\ \mathbf{y = 1}, & for\ \mathbf{x_1 = 1},\ \mathbf{x_2 = 1} \end{cases}$$

5. **Naive Bayes Classifier**

   Input Feature $X = (x_1, x_2, x_3, x_4, x_5)$

Class label y
Naive Bayes Classifier $x \longrightarrow y^*$

$$h_{NB}(x) \;=\; y^* \;=\; argmax_{y\in 0,1}\; p(y|x)$$
$$= argmax_{y\in 0,1}\; p(x_i|y)p(y)$$

Based on the Naive Bayes assumption, the joint probability of the 5 features is the product of individual probability of each element of x, i.e. $p(x_i|y)$

$$p(x|y) = p(x_1, x_2, x_3, x_4, x_5|y) \;=\; \prod_{i=1}^{5} p(x_i|y)$$

(a) The # of parameters for NB classifier
$\Leftrightarrow$ # of parameters for all possible $p(x|y)p(y)$
$\Leftrightarrow$ # of parameters for all possible $p(x_i|y)$, i = 1,2..5 and $p(y)$

As x_i is a binary variable, for a specific i,
if we know$p(x_i = 1|y = 1) \;\implies$ we can find $p(x_i = 0|y = 1)$

- For a corresponding class label (y), we need 1 parameter for each x_i. There are five elements in the feature vector x.
  So, the # of parameters to know $p(x_i|y = 1) = 5, \quad i = 1, 2, ..5$

  Similarly, for y = 0,
  the # of parameter to know $p(x_i|y = 0) = 5 \;\; i = 1, 2, ..5$

- For class prior p(y), the # of parameters = 1

# of parameters = 5 + 5 + 1 = 11
Here, K = 2, D = 5
# of parameters = K * D + (K - 1) = 2 * 5 + 1 = 11

$\therefore$ **# of independent parameters in the Naive Bayes Classifier = 11**

(b) The 11 parameters that need to be estimated are as below:
$p(x_1 = sunny|\; y = Yes)$, $p(x_1 = sunny|\; y = No)$
$p(x_2 = warm|\; y = Yes)$, $p(x_2 = warm|\; y = No)$
$p(x_3 = high|\; y = Yes)$, $p(x_3 = high|\; y = No)$
$p(x_4 = strong|\; y = Yes)$, $p(x_4 = strong|\; y = No)$
$p(x_5 = warm|\; y = Yes)$, $p(x_5 = warm|\; y = No)$
And, $p(y = Yes)$

If y = 1 for Yes,

$p(x_1 = sunny|\; y = 1) = \frac{\#\text{ points } x_1 = sunny,\; y=1}{\#\text{ points } y=1}$

Therefore,

$\mathbf{p(x_1 = sunny|\ y = 1) = \frac{4}{4} = 1}$ $\mathbf{p(x_1 = sunny|\ y = 0) = \frac{0}{3} = 0}$

$\mathbf{p(x_2 = warm|\ y = 1) = \frac{3}{4}}$ $\mathbf{p(x_2 = warm|\ y = 0) = \frac{0}{3} = 0}$

$\mathbf{p(x_3 = high|\ y = 1) = \frac{2}{4} = \frac{1}{2}}$ $\mathbf{p(x_3 = high|\ y = 0) = \frac{2}{3}}$

$\mathbf{p(x_4 = strong|\ y = 1) = \frac{2}{4} = \frac{1}{2}}$ $\mathbf{p(x_4 = strong|\ y = 0) = \frac{1}{3}}$

$\mathbf{p(x_5 = warm|\ y = 1) = \frac{3}{4}}$ $\mathbf{p(x_5 = warm|\ y = 0) = \frac{2}{3}}$

$\mathbf{p(y = 1) = \frac{4}{7}}$

(c) A new input vector is = (sunny, cold, high, strong, cool)

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x)}$$

$$p(x|y = 1) = p(x_1, x_2, x_3, x_4, x_5|y = 1) = \prod_{i=1}^{5} p(x_i|y = 1)$$

$p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)$

- $p(x_1 = sunny|\ y = 1) = 1$
- $p(x_2 = cold|\ y = 1) = 1 - p(x_2 = warm|\ y = 1) = 1 - \frac{3}{4} = \frac{1}{4}$
- $p(x_3 = high|\ y = 1) = \frac{1}{2}$
- $p(x_4 = strong|\ y = 1) = \frac{1}{2}$
- $p(x_5 = cool|\ y = 1) = 1 - p(x_5 = warm|\ y = 1) = 1 - \frac{3}{4} = \frac{1}{4}$

$$\therefore p(x|y = 1) = \prod_{i=1}^{5} p(x_i|y = 1) = 1 * \frac{1}{4} * \frac{1}{2} * \frac{1}{2} * \frac{1}{4} = \frac{1}{64}$$

- $p(x_1 = sunny|\ y = 0) = 0$

$$\therefore p(x|y = 0) = \prod_{i=1}^{5} p(x_i|y = 0) = 0$$

$p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0) = \frac{1}{64} * \frac{4}{7} = \frac{1}{112}$

And, $p(y = 1) = \frac{4}{7}$

$p(y = 1|x) = \frac{p(x|y=1)p(y=1)}{p(x)} = \frac{\frac{1}{64} * \frac{4}{7}}{\frac{1}{112}} = 1$

If y = 0 for No, then
$p(y = 0|x) = 1 - 1 = 0$

Here, $p(y = 1|x) > p(y = 0|x)$.
Hence, the classifier will assign class class label y = 1.

$\therefore$ **The Naive Bayes classifier will assign the class label y = Yes.**

———- THE END ———