CS446: Machine Learning **Spring 2017**

Problem Set 6

Kexin Hui *Handed In: April 20, 2017*

1. Answer to problem 1 - Naïve Bayes and Learning Threshold Functions

   a. $f_{TH(4,9)} = 1$ if and only if 4 or more x's components are 1. That is to say, $f_{TH(4,9)} = 1$ if and only if $\sum_{i=1}^{9} x_i \geq 4$. Therefore, we can construct the linear decision surface to be $f(x) = \vec{w}^T x + \theta$, where $\vec{w} = \{1, 1, 1, 1, 1, 1, 1, 1, 1\}$ and $\theta = -4$.

   b. Prior probability:

$$\begin{aligned} p(f_{TH(4,9)} = 0) &= \frac{\binom{9}{0}\binom{9}{1}\binom{9}{2}\binom{9}{3}}{2^9} = \frac{65}{256} \\ p(f_{TH(4,9)} = 1) &= 1 - p(f_{TH(4,9)} = 0) = \frac{191}{256} \end{aligned} \tag{1}$$

   According to Naïve Bayes algorithm,

$$\begin{aligned} x_{MAP} &= argmax\ p(f_{TH}|x) = argmax\ p(x|f_{TH})p(f_{TH})/p(x) \\ &= argmax\ p(x|f_{TH})p(f_{TH}) \\ &= argmax\ \prod_i p(x_i|f_{TH})p(f_{TH}) \end{aligned} \tag{2}$$

   We need to know $p(x_i|f_{TH} = 0)p(f_{TH} = 0)$ and $p(x_i|f_{TH} = 1)p(f_{TH} = 1)$. Since data is uniform distributed, we can calculate that for $i = 1, 2, ..., 9$,

$$\begin{aligned} p(x_i = 1|f_{TH} = 0) &= \frac{p(x_i = 1, f_{TH(4,9)} = 0)}{p(f_{TH(4,9)} = 0)} \\ &= \frac{\left[0 + \binom{8}{0} + \binom{8}{1} + \binom{8}{2}\right]/2^9}{65/256} = \frac{37}{130} \\ p(x_i = 0|f_{TH} = 0) &= 1 - p(x_i = 1|f_{TH} = 0) = \frac{93}{130} \\ p(x_i = 1|f_{TH} = 1) &= \frac{p(x_i = 1, f_{TH(4,9)} = 1)}{p(f_{TH(4,9)} = 1)} \\ &= \frac{\left[\binom{8}{3} + \binom{8}{4} + \binom{8}{5} + \binom{8}{6} + \binom{8}{7} + \binom{8}{8}\right]/2^9}{191/256} = \frac{219}{382} \\ p(x_i = 0|f_{TH} = 1) &= 1 - p(x_i = 1|f_{TH} = 1) = \frac{163}{382} \end{aligned} \tag{3}$$

Let $p_i = \frac{219}{382}$ and $q_i = \frac{37}{130}$.In order to predict 1 here, we have

$$\begin{aligned} & log\frac{p(f_{TH}=1)}{p(f_{TH}=0)} + \sum_i log\frac{1-p_i}{1-q_i} + \sum_i (log\frac{p_i}{1-p_i} - log\frac{q_i}{1-q_i})x_i > 0 \\ & log\frac{191}{65} + \sum_{i=1,\ldots,9} log\frac{163/382}{93/130} + \sum_{i=1,\ldots,9} (log\frac{219}{163} - log\frac{37}{93})x_i > 0 \\ & -1.551618 + \sum_i 0.528538 x_i > 0 \\ & \sum_{i=1}^{9} x_i > 2.9357 \end{aligned} \tag{4}$$

c. It is quite straight-forward that the hypothesis generated by naïve Bayes algorithm does not match our specification at first. If we have three of nine x's components are 1, we still satisfy Equation (4) which is our hypothesis. But in this case, the hypothesis cannot represent our linear decision surface.

d. No. Naïve Bayes assumes that the features are conditionally independent. However, we can prove that

$$\begin{aligned} p(x_1 = 1, x_2 = 1 | f_{TH} = 0) &= \frac{\left[1 + \binom{8}{1}\right]/2^9}{65/256} = \frac{9}{130} \\ p(x_1 = 1 | f_{TH} = 0) &= p(x_2 = 1 | f_{TH} = 0) = \frac{37}{130} \end{aligned} \tag{5}$$

In this case, we see that

$$p(x_1 = 1, x_2 = 1 | f_{TH} = 0) \neq p(x_1 = 1 | f_{TH} = 0) \times p(x_2 = 1 | f_{TH} = 0) \tag{6}$$

Therefore, we can conclude that, conditional independence assumption based on naïve Bayes algorithm does not hold.

2. Answer to problem 2 - Multivariate Poisson Naïve Bayes

a.

$\Pr(Y=A) = \frac{3}{7}$	$\Pr(Y=B) = \frac{4}{7}$
$\lambda_1^A = 2$	$\lambda_1^B = 4$
$\lambda_2^A = 5$	$\lambda_2^B = 3$

Table 1: Parameters for Poisson naïve Bayes

$p(Y = A)$ and $p(Y = B)$ can be obtained directly from the dataset table since there are three A events and four B events.

Using the log likelihood representation, we have

$$\begin{aligned} log\prod_i \Pr(X_i = x|Y = A) &= \sum_i log\frac{e^{-\lambda^A}(\lambda^A)^x}{x!} \\ &= \sum_i \left[-\lambda^A + xlog(\lambda^A) - logx!\right] \\ &= -n\lambda^A + \sum_i xlog(\lambda^A) - \sum_i log(x!) \end{aligned} \tag{7}$$

Take the derivative of the log likelihood, we can calculate that

$$\begin{aligned} \frac{\partial log(\Pr(x|Y = A))}{\partial\lambda} &= n + \frac{1}{\lambda^A}\sum_i x = 0 \\ \lambda^A &= \frac{\sum_i x}{n} \end{aligned} \tag{8}$$

Therefore, we get Poisson parameters for all cases

$$\begin{aligned} \lambda_1^A &= \frac{0+4+2}{3} = 2 \\ \lambda_2^A &= \frac{3+8+4}{3} = 5 \\ \lambda_1^B &= \frac{6+3+2+5}{4} = 4 \\ \lambda_2^B &= \frac{2+5+1+4}{4} = 3 \end{aligned} \tag{9}$$

b. Given the naïve Bayes assumption

$$\begin{aligned} \frac{\Pr(X_1 = 2, X_2 = 3|Y = A)}{\Pr(X_1 = 2, X_2 = 3|Y = B)} &= \frac{\Pr(X_1 = 2|Y = A)\Pr(X_2 = 3|Y = A)}{\Pr(X_1 = 2|Y = B)\Pr(X_2 = 3|Y = B)} \\ &= \frac{\frac{e^{-\lambda_1^A}(\lambda_1^A)^2}{2!}\frac{e^{-\lambda_2^A}(\lambda_2^A)^3}{3!}}{\frac{e^{-\lambda_1^B}(\lambda_1^B)^2}{2!}\frac{e^{-\lambda_2^B}(\lambda_2^B)^3}{3!}} \\ &= \frac{e^{-\lambda_1^A-\lambda_2^A}(\lambda_1^A)^2(\lambda_2^A)^3}{e^{-\lambda_1^B-\lambda_2^B}(\lambda_1^B)^2(\lambda_2^B)^3} \\ &= e^{-2-5+4+3} \times (\frac{2}{4})^2(\frac{5}{3})^3 \\ &= \frac{125}{108} \end{aligned} \tag{10}$$

c. Based on naïve Bayes algorithm,

$$\begin{aligned} Y &= argmax_y \Pr(Y = y|X_1 = x_1, X_2 = x_2) \\ &= argmax_y \Pr(X_1 = x_1, X_2 = x_2|Y = y)\Pr(Y = y)/\Pr(X_1 = x_1, X_2 = x_2) \\ &= argmax_y \Pr(X_1 = x_1, X_2 = x_2|Y = y)\Pr(Y = y) \end{aligned} \tag{11}$$

In order to predict $Y = A$, we want to have

$$\begin{aligned}
\frac{\Pr(X_1 = x_1, X_2 = x_2|Y = A)\Pr(Y = A)}{\Pr(X_1 = x_1, X_2 = x_2|Y = B)\Pr(Y = B)} &\geq 1 \\
\frac{\frac{e^{-\lambda_1^A}(\lambda_1^A)^{x_1}}{x_1!} \times \frac{e^{-\lambda_2^A}(\lambda_2^A)^{x_2}}{x_2!} \times \Pr(Y = A)}{\frac{e^{-\lambda_1^B}(\lambda_1^B)^{x_1}}{x_1!} \times \frac{e^{-\lambda_2^B}(\lambda_2^B)^{x_2}}{x_2!} \times \Pr(Y = B)} &\geq 1 \\
e^{(-\lambda_1^A - \lambda_2^A + \lambda_1^B + \lambda_2^B)} \times \left(\frac{\lambda_1^A}{\lambda_1^B}\right)^{x_1} \times \left(\frac{\lambda_2^A}{\lambda_2^B}\right)^{x_2} \times \frac{\Pr(Y = A)}{\Pr(Y = B)} &\geq 1 \\
log\frac{\Pr(Y = A)}{\Pr(Y = B)} + (-\lambda_1^A - \lambda_2^A + \lambda_1^B + \lambda_2^B) + x_1 log\frac{\lambda_1^A}{\lambda_1^B} + x_2 log\frac{\lambda_2^A}{\lambda_2^B} &\geq 0
\end{aligned} \tag{12}$$

d. From Part (c), we get that in order to predict Y=A, we need to satisfy

$$\begin{aligned}
log\frac{3}{4} + (-2 - 5 + 4 + 3) + x_1 log\frac{2}{4} + x_2 log\frac{5}{3} &\geq 0 \\
-0.1249 - 0.3010x_1 + 0.2218x_2 &\geq 0
\end{aligned} \tag{13}$$

For $X_1 = 2$ and $X_2 = 3$, we have

$$-0.1249 - 0.3010 \times 2 + 0.2218 \times 3 = -0.0615 \leq 0 \tag{14}$$

Therefore, the classifier will predict $Y = B$ for this example.

3. Answer to problem 3 - Naïve Bayes over Multinomial Distribution

a. We are only caring about the frequency of each word's appearances in the documents, yet we ignore the positions of the words in the representation above. The sequence of the words also matters sometimes. Therefore, we are losing information about the positions of the words.

b.

$$\begin{aligned}
log\Pr(D_i, y_i) &= log\prod_i \Pr(D_i|y_i)\Pr(y_i) \\
&= log\prod_i \left[\left(\theta\frac{n!}{a_i!b_i!c_i!}\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}\right)^{y_i}\left((1-\theta)\frac{n!}{a_i!b_i!c_i!}\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}\right)^{1-y_i}\right] \\
&= \sum_i y_i\left(log\theta + log\frac{n!}{a_i!b_i!c_i!} + a_i log\alpha_1 + b_i log\beta_1 + c_i log\gamma_1\right) \\
&+ (1-y_i)\left(log(1-\theta) + log\frac{n!}{a_i!b_i!c_i!} + a_i log\alpha_0 + b_i log\beta_0 + c_i log\gamma_0\right)
\end{aligned} \tag{15}$$

c. Given that $\alpha_1 + \beta_1 + \gamma_1 = 1$, we have $\gamma_1 = 1 - \alpha_1 - \beta_1 \Rightarrow \frac{\partial\gamma_1}{\partial\alpha_1} = -1$ and $\frac{\partial\gamma_1}{\partial\beta_1} = -1$. Therefore, substitute γ_1 in the equation above and take the derivative in order to

do MLE, we have

$$\begin{aligned}\frac{\partial log\Pr(D_i, y_i)}{\partial \alpha_1} &= \sum_i y_i(\frac{a_i}{\alpha_1} + \frac{c_i}{\gamma_i} \times (-1)) \\ &= \sum_i y_i(\frac{a_i}{\alpha_1} - \frac{c_i}{1-\alpha_1-\beta_1}) = 0\end{aligned} \tag{16}$$

Similarly, we can deduce $\sum_i y_i(\frac{b_i}{\beta_1} - \frac{c_i}{1-\alpha_1-\beta_1}) = 0$.
By observation we notice that only $y_i = 1$ contributes to the result.That is,

$$\begin{cases}\sum_i y_i(\frac{a_i}{\alpha_1} - \frac{c_i}{1-\alpha_1-\beta_1}) = 0 \\ \sum_i y_i(\frac{b_i}{\beta_1} - \frac{c_i}{1-\alpha_1-\beta_1}) = 0\end{cases}$$
$$\begin{cases}\sum_{i,y_i=1} \frac{a_i}{\alpha_1} - \frac{c_i}{1-\alpha_1-\beta_1} = 0 \\ \sum_{i,y_i=1} \frac{b_i}{\beta_1} - \frac{c_i}{1-\alpha_1-\beta_1} = 0\end{cases} \tag{17}$$
$$\begin{cases}\sum_{i,y_i=1} [a_i \times (1-\alpha_1-\beta_1) - c_i \times \alpha_1] = 0 \\ \sum_{i,y_i=1} [b_i \times (1-\alpha_1-\beta_1) - c_i \times \beta_1] = 0\end{cases}$$
$$\begin{cases}\sum_{i,y_i=1} [-(a_i+c_i)\alpha_1 - a_i\beta_1 + a_i] = 0 \\ \sum_{i,y_i=1} [-(b_i+c_i)\beta_1 - b_i\alpha_1 + b_i] = 0\end{cases}$$

Given that $a_i + b_i + c_i = n$, we can calculate that

$$\begin{aligned}\alpha_1 &= \frac{\sum_{i,y_i=1} a_i}{n} \\ \beta_1 &= \frac{\sum_{i,y_i=1} b_i}{n} \\ \gamma_1 &= \frac{\sum_{i,y_i=1} c_i}{n}\end{aligned} \tag{18}$$

By the same token, we have

$$\begin{aligned}\alpha_0 &= \frac{\sum_{i,y_i=0} a_i}{n} \\ \beta_0 &= \frac{\sum_{i,y_i=0} b_i}{n} \\ \gamma_0 &= \frac{\sum_{i,y_i=0} c_i}{n}\end{aligned} \tag{19}$$

4. Answer to problem 4 - Dice Roll
   According to the assumption that the probability that a dice roll coming up as 6 is p, in order to observe a 6 in the sequence, we need to consecutively roll two 6. That is, the probability of observing a 6 in the sequence is p^2. Thus, the probability of observing other numbers 1-5 in the sequence is $1 - p^2$.
   The probability of the generated sequence 3463661622 is $\Pr = (p^2)^4(1-p^2)^6$ based on

Bernoulli distribution. To maximize the parameter p, we take the derivative of the log likelihood. That is,

$$\begin{aligned}\frac{\partial log \Pr}{\partial p} &= \frac{\partial \left(8logp + 6log(1-p^2)\right)}{\partial p} \\ &= \frac{8}{p} + \frac{6}{1-p^2} \times (-2p) \\ &= \frac{8(1-p^2) + 6p \times (-2p)}{p(1-p^2)} \\ &= \frac{8-20p^2}{p(1-p^2)} = 0\end{aligned} \tag{20}$$

Therefore, the most likely value of p for this given sequence is $p = \sqrt{\frac{8}{20}} = \sqrt{\frac{2}{5}}$.