CS446: Machine Learning **Spring 2017**

Problem Set 6 Solutions

Handed Out: April 12th, 2017 *Handed In: April 20th, 2017*

1. [**Naïve Bayes and Learning Threshold Functions - 25 points**]

   (a) [**5 points**] $f_{TH(4,9)}$ can be represented as a linear threshold function as follows:

   $$\sum_{i=1}^{9} x_i \geq 4 \tag{1}$$

   (b) [**10 points**] $f_{TH(4,9)}$ will predict negative iff there are 0, 1, 2, or 3 literals that are active. Thus,

   $$\begin{aligned} \Pr(f_{TH(4,9)} = 0) &= \frac{\binom{9}{0} + \binom{9}{1} + \binom{9}{2} + \binom{9}{3}}{2^9} \\ &= \frac{1 + 9 + 36 + 84}{512} = \frac{130}{512} \\ \text{Consequently, } \Pr(f_{TH(4,9)} = 1) &= 1 - \Pr(f_{TH(4,9)} = 0) = \frac{382}{512} \end{aligned}$$

   Since the distribution is uniform, we can write the following for all $i \in \{1, \ldots, 9\}$:

   $$\begin{aligned} \Pr(x_i = 1 | f_{TH(4,9)} = 1) &= \frac{\binom{8}{3} + \binom{8}{4} + \binom{8}{5} + \binom{8}{6} + \binom{8}{7} + \binom{8}{8}}{382} \\ &= \frac{56 + 70 + 56 + 28 + 8 + 1}{382} = \frac{219}{382} \\ \text{and } \Pr(x_i = 1 | f_{TH(4,9)} = 0) &= \frac{\binom{8}{0} + \binom{8}{1} + \binom{8}{2}}{130} = \frac{1 + 8 + 28}{130} = \frac{37}{130} \end{aligned}$$

   Let $p_i = \Pr(x_i = 1 \mid f_{TH(4,9)} = 1)$ and $q_i = P(x_i = 1 \mid f_{TH(4,9)} = 0)$. Then, the hypothesis produced by naïve Bayes can be written as

   $$\log \frac{\Pr(f_{TH(4,9)} = 1)}{\Pr(f_{TH(4,9)} = 0)} + \sum_{i=1}^{9} \log \frac{1 - p_i}{1 - q_i} + \sum_{i=1}^{9} \left(\log \frac{p_i}{1 - p_i} - \log \frac{q_i}{1 - q_i} \right) x_i \geq 0 \tag{2}$$

   By substituting the values we computed into (2), we get

   $$\begin{aligned} \log \frac{382}{130} + \sum_{i=1}^{9} \log \frac{163/382}{93/130} + \sum_{i=1}^{9} \left(\log \frac{219}{163} - \log \frac{37}{93} \right) x_i &\geq 0 \\ \therefore 0.468 + 9 \times (-0.22244) + \sum_{i=1}^{9} 0.5285 x_i &\geq 0 \\ \therefore \sum_{i=1}^{9} 0.5285 x_i &\geq 1.5516 \\ \therefore \text{The hypothesis generated by naïve Bayes is } \sum_{i=1}^{9} x_i &\geq 2.9359 \end{aligned} \tag{3}$$

(c) **[5 points]** If we compare the two equations (1) and (3), we can easily prove that the hypothesis produced by naïve Bayes cannot represent $f_{TH(4,9)}$ because (3) will make a mistake if any three literals are active. Therefore, naïve Bayes cannot learn $f_{TH(4,9)}$, although more general linear separators can.

(d) **[5 points]** No. The naïve Bayes assumption is conditional independence of its literals given the label, which means

$$\Pr(x_i \wedge x_j \mid v) = \Pr(x_i \mid v) \Pr(x_j \mid v) \tag{4}$$

To see why this doesn't hold in this case, compare the following two results:

$$\begin{aligned}
\Pr(x_1{=}1 \wedge x_2{=}1 \mid f_{TH(4,9)}{=}0) &= \frac{\binom{7}{0} + \binom{7}{1}}{130} = \frac{8}{130} \\
\Pr(x_1{=}1 \mid f_{TH(4,9)}{=}0) \times \Pr(x_2{=}1 \mid f_{TH(4,9)}{=}0) &= \left(\frac{\binom{8}{0} + \binom{8}{1} + \binom{8}{2}}{130}\right)^2 \\
&= \left(\frac{37}{130}\right)^2 \neq \frac{8}{130}
\end{aligned}$$

This means that the classifier derived using the naïve Bayes algorithm may not be the *optimal classifier*.

Indeed, (1) is the optimal classifier. In this case (but not in general), the optimal classifier is an *exact classifier* – one that never makes classification mistakes. As shown in (c), naïve Bayes cannot represent this function and does make mistakes.

Notice that there is a difference between the optimal classifier and an exact classifier. There are cases in which the optimal classifier is not exact. Any classifier you learn will make mistakes.

On the other hand, there could be cases in which the independence assumptions are violated, but nevertheless, the classifier learned using naïve Bayes is very good or even optimal.

2. **[Multivariate Poisson naïve Bayes - 30 points]** You have two non-negative integer-valued numbers X_1 and X_2 as input-features and your class label Y can take two values, A and B. Conditioned on Y, X_i for $i = 1, 2$ follows a Poisson distribution with parameter λ specific to the class label of Y. That is

$$\Pr[X_i = x|Y = A] = \frac{e^{-\lambda_i^A}(\lambda_i^A)^x}{x!} \quad \text{and} \quad \Pr[X_i = x|Y = B] = \frac{e^{-\lambda_i^B}(\lambda_i^B)^x}{x!} \text{ for } i = 1, 2$$

The given data in Table 1 is generated by a Poisson naïve Bayes model.

(a) **[10 points]** Under the naïve Bayes assumption, conditioned on $Y = A$, X_1 is generated independently of X_2; furthermore, the conditional distribution of X_1 given $Y = A$ follows a Poisson distribution with parameter λ_i^A (same goes for $Y = B$ and X_2.) We shall invoke this assumption in order to compute the MLE values of different λs. In particular, to compute λ_i^A, we need only look at the

X_1	X_2	Y
0	3	A
4	8	A
2	4	A
6	2	B
3	5	B
2	1	B
5	4	B

Table 1: Dataset for Poisson naïve Bayes

value of X_1 for those examples where $Y = A$. This underlines the basic approach towards calculating the MLE parameters under the naïve Bayes assumption.

Before anything else, we shall show, given observations drawn from Poisson distributions, how to compute the maximum likelihood value of corresponding λ. Let's say we observe values $z_1, z_2, ..., z_n$ from a Poisson distribution with parameter λ, then the log-likelihood of this data is

$$LL(\lambda) = \sum_{i=1}^{n} \log \Pr(z_i|\lambda) = \sum_{i=1}^{n} \log \frac{e^{-\lambda}(\lambda)^{z_i}}{z_i!} = \left(\sum_{i=1}^{n} z_i\right) \log \lambda - n\lambda - \left(\sum_{i=1}^{n} \log(z_i!)\right) .$$

Setting the derivative of $LL(\lambda)$ to zero, yields the maximum likelihood estimate of λ as

$$\lambda = \frac{\sum_i^n z_i}{n} . \tag{5}$$

Now, we use the (5), to estimate the MLE value of λ_1^A, λ_2^A, λ_1^B, and λ_2^B. Now, when $Y = A$, the value of X_i is 0, 4, and 2. So as per (5), the MLE value of λ_1^A is $\frac{0+4+2}{3} = 2$. Similarly the MLE values of λ_2^A, λ_1^B, and λ_2^B are 5, 4, and 3 respectively.

Also, Since the dataset contains 3 A and 4 B labels so the prior probabilities are $\Pr(Y = A) = \frac{3}{7}$ and $\Pr(Y = B) = \frac{4}{7}$. The parameter values are summarized in Table 2.

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

Table 2: Parameters for Poisson naïve Bayes

(b) **[10 points]** The required ratio for any $X_1 = x_1$ and $X_2 = x_2$ can be written as

$$\frac{\Pr(X_1 = x_1, X_2 = x_2|Y = A)}{\Pr(X_1 = x_1, X_2 = x_2|Y = B)} = \frac{\Pr(X_1 = x_1|Y = A)\Pr(X_2 = x_2|Y = A)}{\Pr(X_1 = x_1|Y = B)\Pr(X_2 = x_2|Y = B)} \text{ (NB assumption)}$$
$$= \frac{\frac{e^{-\lambda_1^A}(\lambda_1^A)^{x_1}}{x_1!}\frac{e^{-\lambda_2^A}(\lambda_2^A)^{x_2}}{x_2!}}{\frac{e^{-\lambda_1^B}(\lambda_1^B)^{x_1}}{x_1!}\frac{e^{-\lambda_2^B}(\lambda_2^B)^{x_2}}{x_2!}} = e^{\lambda_1^B+\lambda_2^B-\lambda_1^A-\lambda_2^A}\left(\frac{\lambda_1^A}{\lambda_1^B}\right)^{x_1}\left(\frac{\lambda_2^A}{\lambda_2^B}\right)^{x_2} \quad (6)$$

Plugging in the values of λs from Table 2, and $x_1 = 2$ and $x_2 = 3$ above, we get

$$\frac{\Pr(X_1 = 2, X_2 = 3|Y = A)}{\Pr(X_1 = 2, X_2 = 3|Y = B)} = e^{3+4-3-5}\left(\frac{2}{4}\right)^2\left(\frac{5}{3}\right)^3 = \frac{125}{108}$$

(c) **[5 points]** Given $X_1 = x_1$ and $X_2 = x_2$, we predict $Y = A$ *iff*

$$\frac{\Pr(Y = A|X_1 = x_1, X_2 = x_2)}{\Pr(Y = B|X_1 = x_1, X_2 = x_2)} \geq 1.$$

Now, let us consider this ratio:

$$\frac{\Pr(Y = A|X_1 = x_1, X_2 = x_2)}{\Pr(Y = B|X_1 = x_1, X_2 = x_2)} = \frac{\Pr(X_1 = x_1, X_2 = x_2|Y = A)\Pr(Y = A)}{\Pr(X_1 = x_1, X_2 = x_2|Y = B)\Pr(Y = B)}.$$

Using (6), we can write the above as

$$e^{\lambda_1^B+\lambda_2^B-\lambda_1^A-\lambda_2^A}\left(\frac{\lambda_1^A}{\lambda_1^B}\right)^{x_1}\left(\frac{\lambda_2^A}{\lambda_2^B}\right)^{x_2}\frac{\Pr(Y = A)}{\Pr(Y = B)}\ .$$

So we predict $Y = A$ *iff*

$$e^{\lambda_1^B+\lambda_2^B-\lambda_1^A-\lambda_2^A}\left(\frac{\lambda_1^A}{\lambda_1^B}\right)^{x_1}\left(\frac{\lambda_2^A}{\lambda_2^B}\right)^{x_2}\frac{\Pr(Y = A)}{\Pr(Y = B)} \geq 1$$
$$\Leftrightarrow \lambda_1^B + \lambda_2^B - \lambda_1^A - \lambda_2^A + \log\frac{\Pr(Y = A)}{\Pr(Y = B)} + x_1\log\frac{\lambda_1^A}{\lambda_1^B} + x_2\log\frac{\lambda_2^A}{\lambda_2^B} \geq 0 \text{ (taking log)}$$

(d) **[5 points]** Plugging the values from Table 2, we get

$$Y = A \Leftrightarrow \log\frac{3}{4} + (\log\frac{5}{3})x_2 - (\log 2)x_1 \geq 0\ .$$

For $x_1 = 2$ and $x_2 = 3$, $\log\frac{3}{4} + (\log\frac{5}{3})x_2 - (\log 2)x_1 = \log\frac{125}{144} < 0$. Thus we predict $Y = B$ for this example.

Notice that, even though the conditional probability of $X_1 = 2$ and $X_2 = 3$ is higher for $Y = A$ than $Y = B$ (in part (b)), the prediction gets reversed due to the prior probability of $Y = B$ being higher.

3. [**Naïve Bayes over Multinomial Distribution - 35 points**]

(a) [**2 points**] The aforementioned model loses the order of the words in a document and ignores the semantic meanings of the words.

(b) [**5 points**] Notice that you're asked to generate the **joint log-likelihood** of the data and the label i.e. $\log Pr(D_i, y_i)$ and not just the data conditioned on the label. The joint probability $Pr(D_i, y_i)$ can be written as $Pr(D_i|y_i)Pr(y_i)$. Assume that the prior probability of generating $Pr(Y_i = 1)$ is η; naturally $Pr(Y_i = 0)$ is $1 - \eta$. Now using the fact that $y_i \in \{0, 1\}$, we can conveniently write down the joint probability as

$$\begin{aligned} Pr(D_i, y_i) &= Pr(Y = y_i)Pr(D_i|Y = y_i) \\ &= (Pr(Y = 1)Pr(D_i|Y = 1))^{y_i}(Pr(Y = 0)Pr(D_i|Y = 0))^{1-y_i} \end{aligned} \tag{7}$$

Using the naïve Bayes model we're given, we can write $Pr(D_i, y_i)$ as

$$Pr(D_i, y_i) = \left(\eta \frac{n!}{a_i!b_i!c_i!}\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}\right)^{y_i} \left((1-\eta)\frac{n!}{a_i!b_i!c_i!}\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}\right)^{1-y_i} \tag{8}$$

Finally, we can express the log-likelihood of the document D_i, L_i, as

$$\begin{aligned} L_i = \log Pr(D_i, y_i) &= y_i \left[\log \eta + \log\left(\frac{n!}{a_i!b_i!c_i!}\right) + a_i \log \alpha_1 + b_i \log \beta_1 + c_i \log \gamma_1\right] \\ &+ (1 - y_i)\left[\log(1-\eta) + \log\left(\frac{n!}{a_i!b_i!c_i!}\right) + a_i \log \alpha_0 + b_i \log \beta_0 + c_i \log \gamma_0\right] \end{aligned} \tag{9}$$

(c) [**28 points**] We derive the parameter values by maximizing the joint likelihood $L = \sum_i L_i$ derived in Equation 9. However, notice that the parameters α_1, β_1, and γ_1 are not independent of each other (same goes for α_0, β_0, and γ_0) and so we cannot directly differentiate L to obtain MLE values. We shall have to substitute one of these values in terms of the others to obtain what we desire.

Let's substitute $\gamma_1 = 1 - \alpha_1 - \beta_1$. We will derive the values of α_1 and β_1 and those will automatically give the value of γ_1 as per the above relation. First, we aim at deriving the expression for α_1. We have that $\frac{\partial\, \gamma_1}{\partial \alpha_1} = -1$. Differentiating L as given in Eq. (7), we get

$$\begin{aligned} \frac{\partial L}{\partial \alpha_1} &= \sum_i y_i \left(\frac{a_i}{\alpha_1} + \frac{c_i}{\gamma_1}\frac{\partial\, \gamma_1}{\partial \alpha_1}\right) = \sum_i y_i \left(\frac{a_i}{\alpha_1} - \frac{c_i}{\gamma_1}\right) = \sum_i y_i \left(\frac{a_i\gamma_1 - c_i\alpha_1}{\alpha_1\gamma_1}\right) = 0 \\ \therefore \quad \gamma_1 \sum_i y_i a_i &= \alpha_1 \sum_i y_i c_i \\ \therefore \alpha_1 &= \gamma_1 \frac{\sum_i y_i a_i}{\sum_i y_i c_i} \end{aligned} \tag{10}$$

$$
\begin{aligned}
\text{Similarly,} \qquad \beta_1 &= \gamma_1 \frac{\sum_i y_i b_i}{\sum_i y_i c_i} \\
\text{But,} \qquad \alpha_1 + \beta_1 + \gamma_1 &= 1 \\
\therefore \gamma_1 \left[\frac{\sum_i y_i a_i}{\sum_i y_i c_i} + \frac{\sum_i y_i b_i}{\sum_i y_i c_i} + 1 \right] &= 1 \\
\therefore \gamma_1 \left[\frac{\sum_i y_i (a_i + b_i + c_i)}{\sum_i y_i c_i} \right] &= 1 \\
\text{Since, } a_i + b_i + c_i = n, \quad \gamma_1 &= \frac{\sum_i y_i c_i}{n \sum_i y_i} \qquad (11) \\
\text{Substituting in Eq. 10,} \quad \alpha_1 &= \frac{\sum_i y_i a_i}{n \sum_i y_i} \qquad (12) \\
\text{and } \beta_1 &= \frac{\sum_i y_i b_i}{n \sum_i y_i} \qquad (13)
\end{aligned}
$$

Doing a similar derivation as above with α_0, β_0, and γ_0, we can show that

$$
\begin{aligned}
\alpha_0 &= \frac{\sum_i (1 - y_i) a_i}{n \sum_i (1 - y_i)} \qquad (14) \\
\beta_0 &= \frac{\sum_i (1 - y_i) b_i}{n \sum_i (1 - y_i)} \qquad (15) \\
\text{and} \quad \gamma_0 &= \frac{\sum_i (1 - y_i) c_i}{n \sum_i (1 - y_i)} \qquad (16)
\end{aligned}
$$

Note that this is not the only way to solve this. We can also work out the expressions using KKT conditions to add the constraint in L. Let

$$
\begin{aligned}
L' &= \sum_i y_i \left[\log \eta + \log \left(\frac{n!}{a_i! b_i! c_i!} \right) + a_i \log \alpha_1 + b_i \log \beta_1 + c_i \log \gamma_1 \right] \\
&\quad + \sum_i (1 - y_i) \left[\log(1 - \eta) + \log \left(\frac{n!}{a_i! b_i! c_i!} \right) + a_i \log \alpha_0 + b_i \log \beta_0 + c_i \log \gamma_0 \right] \\
&\quad - \lambda_1 (\alpha_1 + \beta_1 + \gamma_1 - 1) - \lambda_0 (\alpha_0 + \beta_0 + \gamma_0 - 1) \\
\therefore \frac{\partial L'}{\partial \alpha_1} &= \sum_i y_i \left(\frac{a_i}{\alpha_1} \right) - \lambda_1 = \frac{\sum_i y_i a_i}{\alpha_1} - \lambda_1 = 0 \\
\therefore \alpha_1 &= \frac{1}{\lambda} \sum_i y_i a_i \\
\text{Similarly, } \beta_1 &= \frac{1}{\lambda} \sum_i y_i b_i \quad \text{and} \quad \gamma_1 = \frac{1}{\lambda} \sum_i y_i c_i \\
\alpha_1 + \beta_1 + \gamma_1 &= 1 = \frac{1}{\lambda} \left(\sum_i y_i (a_i + b_i + c_i) \right) \qquad \Longrightarrow \qquad \lambda = n \sum_i y_i
\end{aligned}
$$

Substituting λ gives the same expressions as Eq. (11) to (13).

4. [**Dice Roll - 10 points**]

According to the described scheme of generating a coin toss sequence, we see that the probability of `6` appearing in the observed sequence is p^2: the first toss comes out as an `6` (with probability p), and then the second toss also comes out as an `6`. Hence, the probability of any number within the range 1-5 appearing in the observed sequence is $1 - \Pr(\texttt{6 appearing}) = 1 - p^2$. For each one of the five, the probability is $(1 - p^2)/5$.

Assuming Bernoulli distribution, if the final sequence S of length n contains k `6`'s,

$$\begin{aligned} \Pr(S) &= (p^2)^k(1-p^2)^{n-k} \\ L = \log \Pr(S) &= 2k \log p + (n-k)\log(1-p^2) \\ \therefore \frac{\partial\, L}{\partial p} &= \frac{2k}{p} + \frac{(n-k)(-2p)}{1-p^2} = \frac{2k - 2kp^2 - 2np^2 + 2kp^2}{p(1-p^2)} = 0 \\ \therefore 2k = 2np^2 &\Rightarrow p^2 = \frac{k}{n}, \text{ or } p = \sqrt{\frac{k}{n}}. \end{aligned}$$

Note that if we use $\Pr(S) = (p^2)^k(\frac{1-p^2}{5})^{n-k}$ instead, we have

$$L = \log \Pr(S) = 2k \log p + (n-k)\log(1-p^2) - (n-k)\log 5,$$

where the last term is independent of p and does not affect our final solution. Both solutions are accepted. (What do these two different $\Pr(S)$'s represent, respectively?)

For the sequence `3463661622`, $k = 4$ and $n = 10$. Hence, the maximum likelihood estimate of $p = \sqrt{\frac{4}{10}} = 0.632$.