CS446: Machine Learning	**Fall 2015**

Problem Set 6

Jifu Zhao	*Handed In: 11/18/2015*

1. **Answer to problem 1**

(a) Let's write $f_{TH(4,9)}$ as the a function in the form of $f(x) = sign(\vec{w}\vec{x} + \theta)$.

Choose $\vec{w} = [1, 1, 1, 1, 1, 1, 1, 1, 1]$ and $\theta = -4$, in this way,

$$f(x) = sign(\vec{w}\vec{x} - 4)$$

So, when $\vec{w}\vec{x} - 4 \geqslant 0$, $f(x) = 1$, otherwise, $f(x) = 0$.

It is clear that: $y = \vec{w}\vec{x} - 4$ is a linear function, so, $f_{TH(4,9)}$ **has a linear decision surface over the 9 dimensional Boolean cube**.

(b) According to the Naive Bayes algorithm,

$$y = argmax_{y=0or1} P(y) \prod_{i=1}^{9} P(x_i|y)$$

1, Let's first calculate $P(y)$. When $y = 0$,

$$P(y=0) = \binom{9}{0}(\frac{1}{2})^9 + \binom{9}{1}(\frac{1}{2})^9 + \binom{9}{2}(\frac{1}{2})^9 + \binom{9}{3}(\frac{1}{2})^9 = \frac{65}{256}$$

When $y = 1$,

$$P(y=1) = \binom{9}{4}(\frac{1}{2})^9 + \binom{9}{5}(\frac{1}{2})^9 + \binom{9}{6}(\frac{1}{2})^9 + \binom{9}{7}(\frac{1}{2})^9 + \binom{9}{8}(\frac{1}{2})^9 + \binom{9}{9}(\frac{1}{2})^9 = \frac{191}{256}$$

2, Now, let's calculate $P(x_i|y)$,

$$P(x_i|y) = \frac{P(x_i)P(y|x_i)}{P(y)}$$

$P(y=0) = \frac{65}{256}$, $P(y=1) = \frac{191}{256}$, and $P(x_i = 1) = P(x_i = 0) = \frac{1}{2}$

Finally, we need to calculate $P(y|x_i)$, there are 4 situations.

When $x_i = 0$ **and** $y = 0$:

$$P(y=0|x_i=0) = \binom{8}{5}(\frac{1}{2})^8 + \binom{8}{6}(\frac{1}{2})^8 + \binom{8}{7}(\frac{1}{2})^8 + \binom{8}{8}(\frac{1}{2})^8 = \frac{93}{256}$$

When $x_i = 0$ **and** $y = 1$:

$$P(y=1|x_i=0) = \binom{8}{4}(\frac{1}{2})^8 + \binom{8}{5}(\frac{1}{2})^8 + \binom{8}{6}(\frac{1}{2})^8 + \binom{8}{7}(\frac{1}{2})^8 + \binom{8}{8}(\frac{1}{2})^8 = \frac{163}{256}$$

When $x_i = 1$ **and** $y = 0$:

$$P(y=0|x_i=1) = \binom{8}{6}(\frac{1}{2})^8 + \binom{8}{7}(\frac{1}{2})^8 + \binom{8}{8}(\frac{1}{2})^8 = \frac{37}{256}$$

When $x_i = 1$ **and** $y = 1$:

$$P(y=1|x_i=1) = \binom{8}{3}(\frac{1}{2})^8 + \binom{8}{4}(\frac{1}{2})^8 + \binom{8}{5}(\frac{1}{2})^8 + \binom{8}{6}(\frac{1}{2})^8 + \binom{8}{7}(\frac{1}{2})^8 + \binom{8}{8}(\frac{1}{2})^8 = \frac{219}{256}$$

So, **When** $x_i = 0$ **and** $y = 0$:

$$P(x_i = 0|y=0) = \frac{93}{130}$$

When $x_i = 0$ **and** $y = 1$:

$$P(x_i = 0|y=1) = \frac{163}{382}$$

When $x_i = 1$ **and** $y = 0$:

$$P(x_i = 1|y=0) = \frac{37}{130}$$

When $x_i = 1$ **and** $y = 1$:

$$P(x_i = 1|y=1) = \frac{219}{382}$$

3, Now, according to 1 and 2, we can conclude that:

when $y = 0$**,**

$$h(y=0) = P(y=0)\prod_{i=1}^{9} P(x_i|y=0) = \frac{65}{256}\prod_{i=1}^{9}\{\frac{93}{130}\, if\; x_i = 0,\; or \frac{37}{130}\, if\; x_i = 1\}$$

when $y = 1$**,**

$$h(y=1) = P(y=1)\prod_{i=1}^{9} P(x_i|y=0) = \frac{191}{256}\prod_{i=1}^{9}\{\frac{163}{382}\, if\; x_i = 0,\; or \frac{219}{382}\, if\; x_i = 1\}$$

So, after calculating the $h(y = 0)$ **and** $h(y = 1)$**, we just need to compare their value and choose the larger one.**

(c) To show that the final hypothesis in (b) does not represent this function, we just need to find a contradiction.

Let's calculate $\vec{x} = [1, 1, 1, 0, 0, 0, 0, 0, 0]$ according to (b), we can get:

$$h(y = 0) = 7.8466 * 10^{-4}$$
$$h(y = 1) = 8.4855 * 10^{-4}$$

According to the analysis in (b), we should think that $y = 1$. However, in fact, the true value for $\vec{x} = [1, 1, 1, 0, 0, 0, 0, 0, 0]$ is $y = 0$.

So, the final hypothesis in (b) does not represent this function.

(d) In Naive Bayes algorithm, we assume that feature values are independent given the target value. But, in this question, this assumption is not satisfied.

In the previous analysis, we assume that $P(x_i|y)$ are the same for $i = 1\ to\ 9$. However, this is not right.

For example, let's analysis the situation when $y = 1$. If we know that x_1, x_2, x_3, x_4 are all 1, then, the value of x_5, x_6, x_7, x_8, x_9 will not have any effect on the final result. So, now,

$$\mathbf{P(x_i = 0\ \text{or}\ 1\ |\ y = 1) = \frac{1}{2},\ \ \text{for}\ i = 5,\ 6,\ 7,\ 8,\ 9}$$

In another way, we mean that: the value of y is only dependent on some features, not dependent on all features. But in Naive Bayes algorithm, we always assume that the final hypothesis is dependent on all features.

So, it is not appropriate to make the assumption that feature values are independent given the target value.

2. **Answer to problem 2**

(a) From Table 1, we can calculate that:

$$Pr(Y = A) = \frac{3}{7}, \quad Pr(Y = B) = \frac{4}{7}$$

Now, let's calculate λ_1^A, λ_2^A, λ_1^B, λ_2^B, our goal is to calculate the value of

$$P(X_1, X_2, Y \mid \lambda)$$

where λ denote λ_1^A, λ_2^A, λ_1^B, λ_2^B

$$P(X_1, X_2, Y \mid \lambda) = P(X_1, X_2|Y, \lambda)P(Y|\lambda)$$

Since that:

$$P(Y = A|\lambda) = 3/7, \quad P(Y = B|\lambda) = 4/7$$

and

$$P(X_1, X_2|Y = A, \lambda) = Pr[X_1|Y = A]Pr[X_2|Y = A] = \frac{e^{-\lambda_1^A}(\lambda_1^A)^{x_1}}{x_1!} \frac{e^{-\lambda_2^A}(\lambda_2^A)^{x_2}}{x_2!}$$

$$P(X_1, X_2|Y = B, \lambda) = Pr[X_1|Y = B]Pr[X_2|Y = B] = \frac{e^{-\lambda_1^B}(\lambda_1^B)^{x_1}}{x_1!} \frac{e^{-\lambda_2^B}(\lambda_2^B)^{x_2}}{x_2!}$$

So, if we set that: $y = 0, if \quad Y = A$ and $y = 1, if \quad Y = B$

$$P(X_1, X_2, Y \mid \lambda) = [\frac{e^{-\lambda_1^A}(\lambda_1^A)^{x_1}}{x_1!} * \frac{e^{-\lambda_2^A}(\lambda_2^A)^{x_2}}{x_2!} * \frac{3}{7}]^{(1-y)} * [\frac{e^{-\lambda_1^B}(\lambda_1^B)^{x_1}}{x_1!} * \frac{e^{-\lambda_2^B}(\lambda_2^B)^{x_2}}{x_2!} * \frac{4}{7}]^{y}$$

So,

$$log(P(X_1, X_2, Y \mid \lambda)) = (1-y)[(-\lambda_1^A - \lambda_2^A) + x_1 log(\lambda_1^A) + x_2 log(\lambda_2^A) + Const_1]$$
$$+y[(-\lambda_1^B - \lambda_2^B) + x_1 log(\lambda_1^B) + x_2 log(\lambda_2^B) + Const_2]$$

So, the likelihood for all dataset will be:

$$L = \sum log(P(X_1, X_2, Y \mid \lambda))$$

So, for λ_1^A, let:

$$\frac{dL}{d\lambda_1^A} = 0$$

We can get that: $\sum_{Y=A} -1 + \frac{x_1}{\lambda_1^A} = 0$ So, finally we get that:

$$\boldsymbol{\lambda_1^A = 2}$$

Similarly, we can get that:

$$\boldsymbol{\lambda_2^A = 5}$$
$$\boldsymbol{\lambda_1^B = 4}$$
$$\boldsymbol{\lambda_2^B = 3}$$

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

Table 1: Parameters for Poisson naïve Bayes

(b) According to

$$Pr(X_1, X_2|Y) = Pr[X_1|Y] * Pr[X_2|Y]$$

we have:

$$Pr(X_1 = 2, X_2 = 3|Y = A) = \frac{e^{-2}2^2}{2!} * \frac{e^{-5}5^3}{3!} = 0.038$$

$$Pr(X_1 = 2, X_2 = 3|Y = B) = \frac{e^{-4}4^2}{2!} * \frac{e^{-3}3^3}{3!} = 0.033$$

So,

$$\frac{\Pr(X_1=2, X_2=3 \mid Y=A)}{\Pr(X_1=2, X_2=3 \mid Y=B)} = 1.157$$

(c) In order to derive the expression for the Poisson Naive Bayes predictor for Y, we can just easily compare the value of $P(Y = A|X_1, X_2)$ and $P(Y = B|X_1, X_2)$. Since $P(Y|X_1, X_2) = P(X_1|Y) * P(X_2|Y) * P(Y)$, so, we just need to compute

$$y = sign(\frac{P(X_1|Y) * P(X_2|Y) * P(Y = A)}{P(X_1|Y) * P(X_2|Y) * P(Y = B)})$$

if $\frac{P(X_1|Y)*P(X_2|Y)*P(Y=A)}{P(X_1|Y)*P(X_2|Y)*P(Y=B)} > 1$, we choose $y = A$

if $\frac{P(X_1|Y)*P(X_2|Y)*P(Y=A)}{P(X_1|Y)*P(X_2|Y)*P(Y=B)} < 1$, we choose $y = B$

So,

$$y = sign(\frac{P(Y = A) * e^{-\lambda_1^A}(\lambda_1^A)^{x_1} e^{-\lambda_2^A}(\lambda_2^A)^{x_2}/(x_1!x_2!)}{P(Y = B) * e^{-\lambda_1^B}(\lambda_1^B)^{x_1} e^{-\lambda_2^B}(\lambda_2^B)^{x_2}/(x_1!x_2!)})$$

(d) According to Table 1 and the expression in (c), we can get that:

$$\mathbf{y = sign(\frac{3 * e^{-2}2^{x_1}e^{-5}5^{x_2}/(x_1!x_2!)}{4 * e^{-4}4^{x_1}e^{-3}3^{x_2}/(x_1!x_2!)}) = sign(0.75 * 0.5^{x_1} * 1.67^{x_2})}$$

When $X_1 = 2, \quad X_2 = 3$, through the above expression, we can get that:

$$y = sign(0.87) = 0$$

So, we should choose $\mathbf{Y = B}$

3. **Answer to problem 3**

(a) In order to successfully represent the documents, we also need the **prior probability**, such as $\mathbf{Pr(y=1)}$ and $\mathbf{Pr(y=0)}$

(b) According to the definition, we can get that:

$$Pr(D_i, y=1) = Pr(D_i|y=1) * Pr(y=1) = \theta * \frac{n!}{a_i!b_i!c_i!}\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}$$

$$Pr(D_i, y=0) = Pr(D_i|y=0) * Pr(y=0) = (1-\theta) * \frac{n!}{a_i!b_i!c_i!}\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}$$

So, the final expression would be:

$$Pr(D_i, y) = [\theta * \frac{n!}{a_i!b_i!c_i!}\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}]^y * [(1-\theta) * \frac{n!}{a_i!b_i!c_i!}\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}]^{1-y}$$

After simplification, we can get that:

$$Pr(D_i, y_i) = \frac{n!}{a_i!b_i!c_i!} * [\theta\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}]^y * [(1-\theta)\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}]^{1-y}$$

So, the

$$\begin{gathered} logPr(D_i, y_i) = log(n!) - log(a_i!b_i!c_i!) \\ +y[log(\theta) + a_i log(\alpha_1) + b_i log(\beta_1) + c_i log(\gamma_1)] \\ +(1-y)[log(1-\theta) + a_i log(\alpha_0) + b_i log(\beta_0) + c_i log(\gamma_0)] \end{gathered}$$

(c) First, let's calculate α_1. Let $L = \sum_i logPr(D_i, y_i)$

Notice that we have $\alpha_1 + \beta_1 + \gamma_1 = 1$ and $a_i + b_i + c_i = n$, in order to maximize L, we can use Lgrange Multipliders method.

Condition: $f = \alpha_1 + \beta_1 + \gamma_1 - 1 = 0$ and $g = a_i + b_i + c_i - n$

So: $Lag = \sum_i logPr(D_i, y_i) - \eta_1 f - \eta_2 g$

We need that:

$$\frac{dLag}{d\alpha_1} = 0$$
$$\frac{dLag}{d\beta_1} = 0$$
$$\frac{dLag}{d\gamma_1} = 0$$

$$\frac{dLag}{d\eta_1} = 0$$
$$\frac{dLag}{d\eta_2} = 0$$

So, we can get:

$$\sum_i y_i a_i / \alpha_1 = \eta_1$$
$$\sum_i y_i b_i / \beta_1 = \eta_1$$
$$\sum_i y_i c_i / \gamma_1 = \eta_1$$
$$\alpha_1 + \beta_1 + \gamma_1 = 1$$
$$a_i + b_i + c_i = n$$

Solve these 5 equations, we can get that:

$$n \sum_i y_i = \eta_1(\alpha_1 + \beta_1 + \gamma_1) = \eta_1$$

So, we can get that:

$$\alpha_\mathbf{1} = \frac{\sum_\mathbf{i} \mathbf{y_i a_i}}{\mathbf{n} \sum_\mathbf{i} \mathbf{y_i}}$$
$$\beta_\mathbf{1} = \frac{\sum_\mathbf{i} \mathbf{y_i b_i}}{\mathbf{n} \sum_\mathbf{i} \mathbf{y_i}}$$
$$\gamma_\mathbf{1} = \frac{\sum_\mathbf{i} \mathbf{y_i c_i}}{\mathbf{n} \sum_\mathbf{i} \mathbf{y_i}}$$

In the similar way, we can calculate the value of $\alpha_0,\ \beta_0,\ and\ \gamma_0$. We just need to replace y_i in the above analysis to $1 - y_i$, then, we will got the value of $\alpha_0,\ \beta_0,\ and\ \gamma_0$.

So, the final result is:

$$\alpha_\mathbf{0} = \frac{\sum_\mathbf{i} (\mathbf{1} - \mathbf{y_i}) \mathbf{a_i}}{\mathbf{n} \sum_\mathbf{i} (\mathbf{1} - \mathbf{y_i})}$$
$$\beta_\mathbf{0} = \frac{\sum_\mathbf{i} (\mathbf{1} - \mathbf{y_i}) \mathbf{b_i}}{\mathbf{n} \sum_\mathbf{i} (\mathbf{1} - \mathbf{y_i})}$$
$$\gamma_\mathbf{0} = \frac{\sum_\mathbf{i} (\mathbf{1} - \mathbf{y_i}) \mathbf{c_i}}{\mathbf{n} \sum_\mathbf{i} (\mathbf{1} - \mathbf{y_i})}$$

4. **Answer to problem 4**

For every shown number, there are two situations:

1). Number 6:
It must be the second one, so the probability will be:

$$P(x_i = 6) = p^2$$

2). Number 1, 2, 3, 4, 5:
It can come from the first roll, or it can come from the second roll, so the probability will be:

$$P(x_i = 1\, or\, 2\, or\, 3\, or\, 4\, or\, 5) = (1-p)\frac{1}{5} + p(1-p)\frac{1}{5} = \frac{1}{5}(1+p)(1-p)$$

3). Finally, for this dataset, the **likelihood** would be:

$$P(\vec{x}) = [\frac{1}{5}(1+p)(1-p)]^6 * [p^2]^4 = (\frac{1}{5})^6(1+p)^6(1-p)^6p^8$$

To make

$$\frac{dP(\vec{x})}{dp} = 0$$

we can get that:

$$-12(1-p^2)^6p^9 + 8(1-p^2)^6p^7 = 0$$
$$p^7(1-p^2)^5(8-20p^2) = 0$$

So, we can get that: $\mathbf{p = 0}, \ \mathbf{1}, \ \mathbf{-1}, \ \sqrt{\mathbf{\frac{2}{5}}}, \ -\sqrt{\mathbf{\frac{2}{5}}}$

Since that p should be between 0 and 1, so

$$\mathbf{p} = \sqrt{\mathbf{\frac{2}{5}}}$$