CS446: Machine Learning **Fall 2015**

Problem Set 6

Dominik Durner *Handed In: November 20, 2015*

Question 1

1.a

Given a weight vector $w = (1,1,1,1,1,1,1,1,1)^T$ and a $\Theta = -4$. Then we have a linear threshold function such that $w^T x + \Theta \geq 0$ denotes the outcome of $f_{TH(4,9)} = 1$ and $w^T x + \Theta < 0$ denotes the outcome of $f_{TH(4,9)} = 0$. The described function counts the number of 1 values in a given vector x and denotes the result after adding Θ. Therefore, the count needs to be at least 4 to denote 1 as the outcome.

1.b

We distinguish between two classes that we have to learn ($f_{TH(m,n)} = 0/1$).

Therefore we get following Bayes classifier rule.

$$argmax_{y \in \{0,1\}} P(y) \prod_{i=1}^{9} P(x_i|y) =$$
$$argmax_y \{P(0)P(x_1|0) \cdots P(x_9|0), P(1)P(x_1|1) \cdots P(x_9|1)\}$$

Since the distribution over the Boolean cube $(0/1)^m$ is uniform, we know that each of the Y vector elements is with probability 50% on. Hence, we can calculate P(0) and P(1).

$$P(0) = \binom{9}{9}0.5^9 + \binom{9}{8}0.5^9 + \binom{9}{7}0.5^9 + \binom{9}{6}0.5^9 = \frac{65}{256}$$
$$P(1) = 1 - \frac{65}{256} = \frac{191}{256}$$

Since we don't know $P(x_i|y)$ we can apply Bayes rules to this.

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

To compute this value we need to find the results for $P(y|x_i)$. We can just think what happens for which class if x_1 is 1 or 0.

$$P(0|x_i = 0) = \binom{8}{8}0.5^8 + \binom{8}{7}0.5^8 + \binom{8}{6}0.5^8 + \binom{8}{5}0.5^8 = \frac{93}{256}$$
$$P(0|x_i = 1) = \binom{8}{8}0.5^8 + \binom{8}{7}0.5^8 + \binom{8}{6}0.5^8 = \frac{37}{256}$$
$$P(1|x_i = 0) = \frac{163}{256}$$
$$P(1|x_i = 1) = \frac{219}{256}$$

Hence, we can calculate $P(x_i|y)$.

$$P(x_i = 0|0) = \frac{0.5 * \frac{93}{256}}{\frac{65}{256}} = \frac{93}{130}$$
$$P(x_i = 1|0) = \frac{0.5 * \frac{37}{256}}{\frac{65}{256}} = \frac{37}{130}$$
$$P(x_i = 0|1) = \frac{0.5 * \frac{163}{256}}{\frac{191}{256}} = \frac{163}{382}$$
$$P(x_i = 1|1) = \frac{0.5 * \frac{163}{256}}{\frac{191}{256}} = \frac{219}{382}$$

Therefore, we need to find the maximize over the following two functions to predict a new label.

$$h(x) = argmax_y(\frac{65}{256}\prod_{i=1}^{9}\frac{37 + 56(1 - x_i)}{130} \text{ (for y = 0)},$$
$$\frac{191}{256}\prod_{i=1}^{9}\frac{163 + 56x_i}{382} \text{ (for y = 1)})$$

1.c

Let's assume the vector $x = (1, 1, 1, 0, 0, 0, 0, 0, 0)^T$. We know that this vector should be out put $f(x)_{TH(4,9)=0}$. If we run Naive Bayes we get the following results.

$$\frac{65}{256}\prod_{i=1}^{9}\frac{37 + 56(1 - x_i)}{130} = 0.000784 \text{ (for y = 0)},$$
$$\frac{191}{256}\prod_{i=1}^{9}\frac{163 + 56x_i}{382} = 0.000848 \text{ (for y = 1)}$$

Hence, we would assume that the vector is a positive result since 0.000848 is greater than 0.000784. This is a contradiction to the real result and therefore we showed that the final hypothesis is not always correct for this example.

1.d

No, not all constraints of Naive Bayes are satisfied. Our assumption is that $P(x_1|0)$ is independent of $P(x_2|0)$ and therefore $P(0|x_1)$ is independent of $P(0|x_2)$. This does not need to be right and therefore we would need to consider all other x_i features to consider the value of $P(0|x_1)$. So if we already see 4 values being 1 we know that the probability of the remaining $P(0|x_i)$ should be 0, regardless of the value x_i.

Question 2

2.a

$$Pr[X_i = x|Y = A] = \frac{e^{-\lambda_i^A}(\lambda_i^A)^x}{x!}$$
$$Pr[X_i = x|Y = B] = \frac{e^{-\lambda_i^B}(\lambda_i^B)^x}{x!}$$

For a given example (x_1, x_2, y) we get the following equations.

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

We can combine those two functions if we choose $y = 0$ iff its value is A and $y = 1$ iff its value is B.

$$Pr[x_1, x_2, y] = \left[\frac{e^{-\lambda_1^A-\lambda_2^A}(\lambda_1^A)^{x_1}(\lambda_2^A)^{x_2}}{x_1!x_2!} * \frac{3}{7}\right]^{1-y} \cdot \left[\frac{e^{-\lambda_1^B-\lambda_2^B}(\lambda_1^B)^{x_1}(\lambda_2^B)^{x_2}}{x_1!x_2!} * \frac{4}{7}\right]^{y}$$

Now wen can take the log of the whole equation and we get the following term (with combined constant C).

$$\begin{aligned} log(Pr[x_1, x_2, y]) = (1 - y)[(-\lambda_1^A - \lambda_2^A) + x_1 log(\lambda_1^A) + x_2 log(\lambda_2^A) + C]+ \\ y[(-\lambda_1^B - \lambda_2^B) + x_1 log(\lambda_1^B) + x_2 log(\lambda_2^B) + C'] \end{aligned}$$

With $\sum_{x_1,x_2,y} log(Pr[x_1, x_2, y])$ as the whole data set we get the following equation for λ_1^A.

$$\frac{\partial \sum_{x_1,x_2,y} log(Pr[x_1, x_2, y])}{\partial \lambda_1^A} = \sum (1 - y)(-1 + \frac{x_1}{\lambda_1^A})$$

Now we want to maximize this and therefore we set the derivative to 0. Since (1-y) gives us only examples where y=A we can rewrite it as following equation and similar if we partially derivate for the other λs.

$$\sum_A (-1 + \frac{x_1}{\lambda_1^A}) = 0$$
$$\sum_A (-1 + \frac{x_2}{\lambda_2^A}) = 0$$
$$\sum_B (-1 + \frac{x_1}{\lambda_1^B}) = 0$$
$$\sum_B (-1 + \frac{x_2}{\lambda_2^B}) = 0$$

With the given data set we get following equations.

$$3 = \frac{6}{\lambda_1^A} \Leftrightarrow \lambda_1^A = 2$$
$$3 = \frac{15}{\lambda_2^A} \Leftrightarrow \lambda_2^A = 5$$
$$4 = \frac{16}{\lambda_1^B} \Leftrightarrow \lambda_1^B = 4$$
$$4 = \frac{12}{\lambda_2^B} \Leftrightarrow \lambda_2^B = 3$$

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

2.b

$$Pr[X_1 = 2|Y = A] = \frac{e^{-2}(2)^2}{2!}$$
$$Pr[X_2 = 3|Y = A] = \frac{e^{-5}(5)^3}{3!}$$
$$Pr[X_1 = 2|Y = B] = \frac{e^{-4}(4)^2}{2!}$$
$$Pr[X_2 = 3|Y = B] = \frac{e^{-3}(3)^3}{3!}$$
$$\frac{Pr[X_1 = 2, X_2 = 3|Y = A]}{Pr[X_1 = 2, X_2 = 3|Y = B]} = \frac{\frac{e^{-2}(2)^2}{2!} * \frac{e^{-5}(5)^3}{3!}}{\frac{e^{-4}(4)^2}{2!} * \frac{e^{-3}(3)^3}{3!}} = 1.15741$$

2.c

Naive Bayes tries to $argmax_{y \in \{A,B\}} P(y)P(x_1|y)P(x_2|y)$ we get following equation that states Y = A iff

$$\frac{Pr[x_1|Y = A]Pr[x_2|Y = A]Pr[Y = A]}{Pr[x_1|Y = B]Pr[x_2|Y = B]Pr[Y = B]} > 1$$

By inserting the calculated values we get the following equation.

$$\frac{\frac{e^{-2}(2)^{x_1}}{x_1!} \frac{e^{-5}(5)^{x_2}}{x_2!} \frac{3}{7}}{\frac{e^{-4}(4)^{x_1}}{x_1!} \frac{e^{-3}(3)^{x_2}}{x_2!} \frac{4}{7}} > 1 \Leftrightarrow 2^{-2-x_1} 3^{1-x_2} 5^{x_2} > 1$$

2.d

$$2^{-2-2} 3^{1-3} 5^2 = \frac{125}{144}$$

Since this is smaller 1 the classifier would predict as label $Y = B$!

Question 3

3.a

We just consider the word count but we loose the context of the document since we don't represent sentences and how the words are combined.

3.b

$$
\begin{aligned}
Pr[D_i|y=0] &= \frac{n!}{a_i!b_i!c_i!}\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i} \\
Pr[D_i|y=1] &= \frac{n!}{a_i!b_i!c_i!}\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i} \\
\Rightarrow Pr[D_i|y_i] &= \frac{n!}{a_i!b_i!c_i!}[\alpha_1^{a_i}\beta_1^{b_i}\gamma_1^{c_i}]^{y_i}[\alpha_0^{a_i}\beta_0^{b_i}\gamma_0^{c_i}]^{1-y_i}
\end{aligned}
$$

Sine we know that $Pr[y_i = 1] = \Theta$ we get following equation.

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

Taken the log of the equation we get the following result.

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

3.c

We can rewrite the equation from b with $C = log(n!) - log(a_i!b_i!c_i!)$ such that

$$
\begin{aligned}
log(Pr[D_i, y_i]) = &\, y_i(log(\Theta) + C + a_i log(\alpha_1) + b_i log(\beta_1) + c_i log(\gamma_1)) \\
&+ (1-y_i)(log(1-\Theta) + C + a_i log(\alpha_0) + b_i log(\beta_0) + c_i log(\gamma_0))
\end{aligned}
$$

I can now use the Lagrange Multiplier to get the result for this equation. To make it easier I am splitting the calculation into $x_0 = (\alpha_0, \beta_0, \gamma_0)$ and $x_1 = (\alpha_1, \beta_1, \gamma_1)$.

$$
\begin{aligned}
f(x_0) &= \sum_i (1-y_i)(log(1-\Theta) + C + a_i log(\alpha_0) + b_i log(\beta_0) + c_i log(\gamma_0)) \\
f(x_1) &= \sum_i y_i(log(\Theta) + C + a_i log(\alpha_1) + b_i log(\beta_1) + c_i log(\gamma_1))
\end{aligned}
$$

Because those two functions can be calculated with the same approach I am just concentrating on $f(x_1)$ in the following.

$$\bigtriangledown L(x_1,\lambda) = \begin{pmatrix} \bigtriangledown f(x_1) + \lambda \bigtriangledown g(x_1) \\ g(x_1) \end{pmatrix} = \begin{pmatrix} \sum_i (\frac{1}{\alpha_1} a_i y_i) - \lambda \\ \sum_i (\frac{1}{\beta_1} b_i y_i) - \lambda \\ \sum_i (\frac{1}{\gamma_1} c_i y_i) - \lambda \\ 1 - \alpha_1 - \beta_1 - \gamma_1 \end{pmatrix} = \vec{0}$$

Hence we can get following equations with the help of the first three equations.

$$\sum_i y_i(a_i + b_i + c_i) = \lambda(\alpha_1 + \beta_1 + \gamma_1)$$

From our constraint we know that $(\alpha_1 + \beta_1 + \gamma_1) = 1$. Furthermore we know that $n = |D_i| = a_i + b_i + c_i$. Hence, we get following λ.

$$\lambda = n \sum_i y_i$$

Inserting λ into our initial equations gives us the following ones for $\alpha_1, \beta_1, \gamma_1$.

$$\alpha_1 = \frac{\sum_i a_i y_i}{n \sum_i y_i}$$
$$\beta_1 = \frac{\sum_i b_i y_i}{n \sum_i y_i}$$
$$\gamma_1 = \frac{\sum_i c_i y_i}{n \sum_i y_i}$$

As already described earlier we can to the same approach to get the results for $\alpha_0, \beta_0, \gamma_0$. I will just name the result here since there is no hidden work left.

$$\alpha_0 = \frac{\sum_i a_i (1 - y_i)}{n \sum_i (1 - y_i)}$$
$$\beta_0 = \frac{\sum_i b_i (1 - y_i)}{n \sum_i (1 - y_i)}$$
$$\gamma_0 = \frac{\sum_i c_i (1 - y_i)}{n \sum_i (1 - y_i)}$$

Question 4

We know that with probability p we throw a 6. Since we can only observe 6's in a two consecutive 6 rolls, we are going to look for the possibility p^2. We can build following equation with n observed values and k 6's.

$$argmax_p Pr[D|p]$$
$$Pr[D|p] = \binom{n}{k} p^{2k}(1-p^2)^{n-k}$$

To easier calculate the derivatives we again log the equation.

$$log(Pr[D|p]) = log(\binom{n}{k}) + 2k * log(p) + (n-k) * log(1-p^2)$$
$$\frac{\partial log(Pr[D|p])}{\partial p} = \frac{2k}{p} + (n-k) * \frac{-2p}{1-p^2}$$

To find the maximum we need to set the equation to 0 and solve it according to p.

$$0 = \frac{2k}{p} + (n-k) * \frac{-2p}{1-p^2} \Leftrightarrow p^2 = \frac{k}{n}$$
$$\Rightarrow p = \sqrt{\frac{k}{n}}$$

With the given example we get $p = \sqrt{\frac{4}{10}} = 0.6325$