**CS446: Machine Learning** **Fall 2015**

# Problem Set 6

Shivam Upadhyay *Handed In: November 19, 2015*

1. Naïve Bayes and Learning Threshold Functions

(a) Given $f_{TH(4,9)}(x) = 1$ if and only if 4 or more of $x$'s components are 1. Thus, if we consider the 9 dimensional weight vector $W = [111111111]^T$. To realise the condition $f_{TH(4,9)}(x) = 1$ we can set the threshold value $\theta = -4$ so that the equation $w^T\dot{x} + \theta \geq 0$ for $y = 1$ and $w^T\dot{x} + \theta < 0$ for $y = 0$. This is clearly the equation for linear surface.

(b) We know that maximum a posterior hypothesis is given by,

$$h_{MAP} = \underset{y\in\{0,1\}}{\arg\max}\, P(y) \prod_{i=0}^{n} P(x_i|y) \qquad (1)$$

where P(y) ⇒ P(Y=y), that is all small alpha-bates represent values for that particular random variable, and this mathematical notation will be used throughout for convenience and consistency.

Moreover, we are given a uniform distribution thus for the input 9 dimensional input vector X, we can sample all the possible combination of values $x_i$ and thus can compute the probabilities used in (1).

As there are two possible outcomes $y \in \{0, 1\}$, we know from the class that hypothesis H here is the optimal prediction. Equivalently we can rewrite the (1) as:

$$\underset{y\in\{0,1\}}{\arg\max}\, P(y) \prod_{i=0}^{9} P(x_i|y)$$

$$= \underset{y}{\arg\max}(P(0)P(x_1|0)P(x_2|0)\ldots P(x_9|0),$$

$$P(1)P(x_1|1)P(x_2|1)\ldots P(x_9|1))$$

Now, P(0) is the probability that the output for a particular value of vector X is 0, for this to happen we are give a threshold function $f_{TH(4,9)}(x) = 1$ which means that out of 9 features if 6 or more are zero than the output is 0. Solving for the same:

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

Also,

$$P(1) = 1 - P(0) = \frac{191}{256}$$

Now we have to compute $\forall x_i \in \{0, 1\}$ and $y \in \{0, 1\}$ value of $P(x_i|y)$. First let's compute the probabilities where $y = 0$

By Bayes theorem,

$$P(x_i|0) = \frac{P(x_i)P(0|x_i)}{P(0)} \qquad (2)$$

$$P(x_i|1) = \frac{P(x_i)P(1|x_i)}{P(1)} \qquad (3)$$

$P(0|x_i)$ as $x_i$ can take either 0 or 1.
if $x_i = 1$ then for $y = 0$ remaining 8 values in vector X should have 6 or more values as 0

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

if $x_i = 0$ then for $y = 0$ remaining 8 values in vector X should have 5 or more values as 0

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

similarly,

$$P(1|x_i = 1) = \binom{8}{3}0.5^8+\binom{8}{4}0.5^8+\binom{8}{5}0.5^8+\binom{8}{6}0.5^8+\binom{8}{7}0.5^8+\binom{8}{8}0.5^8 = \frac{219}{256}$$

$$P(1|x_i = 0) = \binom{8}{4}0.5^8 + \binom{8}{5}0.5^8\binom{8}{6}0.5^8 + \binom{8}{7}0.5^8 + \binom{8}{8}0.5^8 = \frac{163}{256}$$

Plugging values in (2) and (3) we get,

$$P(x_i = 0|0) = \frac{93}{130} \qquad (4) \qquad \text{Note: } P(x_i) = 1/2$$

$$P(x_i = 1|0) = \frac{37}{130} \qquad (5)$$

$$P(x_i = 0|1) = \frac{163}{382} \qquad (6)$$

$$P(x_i = 1|1) = \frac{219}{382} \qquad (7)$$

Putting this all together, our hypothesis is one that, given x, picks the $y$ value that gives the larger of the two products:

$$\frac{65}{256}\prod_{i=1}^{9}\{\frac{37}{130} \text{ if } x_i = 1, \text{ or } \frac{93}{130} \text{ if } x_i = 0\} \text{ if y} = 0 \qquad (8)$$

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

Combining (8) and (9) we get:

$$\boxed{h(x) = \arg\max_{y\in\{0,1\}} \tfrac{65+126y}{256} \prod_{i=1}^{9} \tfrac{163+56x_i}{382} y + \tfrac{93-56x_i}{130}(1-y)}$$

(c) We can prove this by evaluating the above hypothesis for a value of vector X i.e. by the proof of contradiction. If the label given by hypothesis and the true label doesn't match for the same input then derived hypothesis doesn't represent the function.
Let's take the input vector [1 1 1 0 0 0 0 0 0], which for the given function $f_{TH(4,9)}(x) = 0$, let's put the same input in the hypothesis derived in part(b):
For $y = 0$,

$$\frac{65}{256} * \left(\frac{37}{130}\right)^3 * \left(\frac{93}{130}\right)^6 = 0.0007846682571721$$

For $y = 1$,

$$\frac{191}{256} * \left(\frac{219}{382}\right)^3 * \left(\frac{163}{382}\right)^6 = 0.0008485571685$$

The value for $y = 1$ is larger, so the hypothesis predicts $y = 1$, which is different than the original function. Therefore, our hypothesis does not represent the actual function.

(d) **No the assumptions are not satisfied** by the given threshold function $f_{TH(4,9)}(x)$ because the label given by this function depends on the number of features being on in a given input vector, i.e. if number of features on in a given input vector exceeds threshold than, label will always be 1 irrespective of the values of remaining features. Naïve Bayes assumes attribute values are conditionally independent and learning requires various parameters to be estimated based on their frequencies over the training data. The set of these estimates corresponds to learned hypothesis.You can't assign a probability of a label just given one feature by itself. For example, in this formulation, we assume that $P(x_1|0)$ is independent of $P(x_2|0)$. By Bayes equation we can say that,$P(0|x_1)$ is independent of $P(0|x_2)$. But we can't really get the probability of a label given just one feature's information. A more fair calculation of $P(0|x_1)$ would include all of the other features' values. For example, if there were already 4 other features with values of 1, $P(1|x_1)$ should always be 1, regardless of the value of $x_1$.

2. Multivariate Poisson naïve Bayes

(a) Given:

$$P(X_i = x|Y = A) = \frac{e^{-\lambda_i^A}(\lambda_i^A)^x}{x!}$$

$$P(X_i = x|Y = B) = \frac{e^{-\lambda_i^B}(\lambda_i^B)^x}{x!}$$

Where $X \in \{x_1, x_2\}$ and $Y \in \{A, B\}$ We know by product rule that:

$$P(X = x_i, Y = y_i) = P(Y = y_i|X = x_i)P(X = x_i)$$

Therefore,

$$P(x_1, x_2, y = A) = \frac{e^{-\lambda_i^A(\lambda_i^A)^{x_1}}}{x_1!} * \frac{e^{-\lambda_i^A(\lambda_i^A)^{x_2}}}{x_2!} * P(Y = A) \qquad (1)$$

$$P(x_1, x_2, y = B) = \frac{e^{-\lambda_i^A(\lambda_i^A)^{x_1}}}{x_1!} * \frac{e^{-\lambda_i^B(\lambda_i^B)^{x_2}}}{x_2!} * P(Y = B) \qquad (2)$$

$$P(Y = A) = \frac{3}{7} \qquad (3)$$

$$P(Y = B) = \frac{4}{7} \qquad (4)$$

Combining (1),(2),(3), and (4), [Note: For mathematical convenience we can assume that y=0 for y=A and y=1 for Y=B]

$$P(x_1, x_2, y) = [\frac{e^{-\lambda_1^A - \lambda_2^A}(\lambda_1^A)^{x_1}(\lambda_2^A)^{x_2}}{x_1!x_2!} * \frac{3}{7}]^{1-y} * [\frac{e^{-\lambda_1^B - \lambda_2^B}(\lambda_1^B)^{x_1}(\lambda_2^B)^{x_2}}{x_1!x_2!} * \frac{4}{7}]^{y} \qquad (5)$$

For computing the log likelihood, taking the log of equation (5) we get :

$$logP(x_1, x_2, y) = (1 - y)[(-\lambda_1^A - \lambda_2^A) + x_1 log(\lambda_1^A) + x_2 log(\lambda_2^A) + C]$$
$$+y[(-\lambda_1^B - \lambda_2^B) + x_1 log(\lambda_1^B) + x_2 log(\lambda_2^B) + C']$$

where $C = log(\frac{3}{7*(x_1!x_2!)}$ and $C' = log(\frac{4}{7*(x_1!x_2!)}$

So,given training data maximising the log likelihood requires taking the partial derivative w.r.t that parameter, we also know that summation of probability across the entire data set is given by(for every parameter value):

$$\sum_{x_1,x_2,y} logP(x_1, x_2, y)$$

For $\lambda_1^A$:

$$\frac{d\sum_{x_1,x_2,y} logP(x_1, x_2, y)}{d\lambda_1^A} = \sum(1 - y)[-\lambda_1^A + \frac{x_1}{\lambda_1^A}] = 0$$

As a note, the sum multiplying by $(1 - y)$ is just going to be the sum of all examples where y=0, or y=A. Likewise, when multiplying by $y$ it's just the sum of all examples where y=B.

$$\sum_A -\lambda_1^A + \frac{x_1}{\lambda_1^A} = 0$$

Going through the actual data:

$$3\lambda_1^A = \frac{6}{\lambda_1^A}$$

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

Similarly, for $\lambda_1^B$, $\sum_B -\lambda_1^B + \frac{x_1}{\lambda_1^B} = 0$. And so $4\lambda_1^B = \frac{16}{\lambda_1^B}$, and then $\lambda_1^B = 2$. We can use the same steps to obtain $\lambda_2^A = \sqrt{5}$ and $\lambda_2^B = \sqrt{3}$.

| $\Pr(Y=A) = \ {}^{3}\!/\!{}_{7}$ | $\Pr(Y=B) = \ {}^{4}\!/\!{}_{7}$ |
|---|---|
| $\lambda_1^A = \ \sqrt{2}$ | $\lambda_1^B = 2$ |
| $\lambda_2^A = \ \sqrt{5}$ | $\lambda_2^B = \sqrt{3}$ |

Table 1: Parameters for Poisson naïve Bayes

(b)

$$P(X_1 = 2|Y = A) = \frac{e^{-\sqrt{2}}(\sqrt{2})^2}{2!} = 0.2431167$$

$$P(X_2 = 3|Y = A) = \frac{e^{-\sqrt{5}}(\sqrt{5})^3}{3!} = 0.1991552$$

$$P(X_1 = 2|Y = B) = \frac{e^{-2}(2)^2}{2!} = 0.2431167 = 0.2706706$$

$$P(X_2 = 3|Y = B) = \frac{e^{-\sqrt{3}}(\sqrt{3})^3}{3!} = 0.1532183$$

$$\frac{P(X_1 = 2, X_2 = 3|Y = A)}{P(X_1 = 2, X_2 = 3|Y = B)} = \frac{0.2431167 * 0.1991552}{0.2706706 * 0.1532183} = 1.167$$

(c)

$$h(x_1, x_2) = sgn(\left\lfloor \frac{P(X_1 = x_1|Y = A)P(X_2 = x_2|Y = A)}{P(X_1 = x_1|Y = B)P(X_2 = x_2|Y = B)} \right\rfloor)$$

Where a result of 1 means A, and a result of 0 means B. To avoid messiness, I'm going to omit the sgn and floor functions when simplifying... but know they're still there!

$$= \frac{e^{-\lambda_1^A}(\lambda_1^A)^{x_1} e^{-\lambda_2^A}(\lambda_2^A)^{x_2}}{e^{-\lambda_1^B}(\lambda_1^B)^{x_1} e^{-\lambda_2^B}(\lambda_2^B)^{x_2}}$$

Substituting our values,

$$= \frac{e^{-\sqrt{2}-\sqrt{5}}(\sqrt{2})^{x_1}(\sqrt{5})^{x_2}}{e^{-2-\sqrt{3}}(2)^{x_1}(\sqrt{3})^{x_2}}$$

$$= e^{2+\sqrt{3}-\sqrt{2}-\sqrt{5}}(\frac{\sqrt{2}}{2})^{x_1}(\sqrt{\frac{5}{3}})^{x_2}$$

Which is approximately,

$$h(x_1, x_2) = sgn(\lfloor e^{0.0817693}(0.707107)^{x_1}(1.290994)^{x_2} \rfloor)$$

(d) Again, where 1 means A, and 0 means B.

Given the point $X_1 = 2$, $X_2 = 3$, we get:

$$h(x_1, x_2) = sgn(\lfloor e^{0.0817693}(0.707107)^2(1.290994)^3 \rfloor) = sgn(\lfloor 1.167 \rfloor) = 1$$

So the classifier will predict **A** for $X_1 = 2$, $X_2 = 3$.

3. Naïve Bayes over Multinomial Distribution

   (a) To be answered

   (b) Given:

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

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

$$Pr(y_i = 1) = \theta$$
$$Pr(y_i = 0) = 1 - \theta$$

Combining them together we get:

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

Then,

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

$$log[P(D_i, y_i)] = log(n!) - log(a_i!b_i!c_i!) + y_i[log(\theta) + a_i log(\alpha_1) + b_i log(\beta_1) + c_i log(\gamma_1)]$$
$$+(1-y_i)[log(1-\theta) + a_i log(\alpha_0) + b_i log(\beta_0) + c_i log(\gamma_0)]$$

   (c) For solving the maximisation problem for parameters $\alpha_1$, $\beta_1$, $\gamma_1$, $\alpha_0$, $\beta_0$, and $\gamma_0$, which we can find out by maximising the equation $f(\alpha_1, \beta_1, \gamma_1) = \sum_{i=1}^{m} log[P(D_i, y_i)]$, and our constraint, $g(\alpha_1, \beta_1, \gamma_1) = \alpha_1 + \beta_1 + \gamma_1$.
Lagranges, $\nabla f(\alpha_1, \beta_1, \gamma_1) = \lambda\nabla g(\alpha_1, \beta_1, \gamma_1$ Finding the partial differential w.r.t $\alpha_1$, $\beta_1$, $\gamma_1$ from the b part we know that:

$$log[P(D_i, y_i)] = log(n!) - log(a_i!b_i!c_i!) + y_i[log(\theta) + a_i log(\alpha_1) + b_i log(\beta_1) + c_i log(\gamma_1)]$$
$$+(1-y_i)[log(1-\theta) + a_i log(\alpha_0) + b_i log(\beta_0) + c_i log(\gamma_0)]$$

Taking the derivative w.r.t to $\alpha_1$ i.e. $\frac{\partial log[P(D_i, y_i=1)]}{\partial\alpha_1} = \lambda\frac{\partial(\alpha_1+\beta_1+\gamma_1)]}{\partial\alpha_1}$

$$\Rightarrow \sum_i^m \frac{y_i a_i}{\alpha_1} = \lambda \qquad (1)$$

Similarly, we can solve for other parameters, and we get equations :

$$\sum_i^m \frac{y_i b_i}{\beta_1} = \lambda \qquad (2)$$

$$\sum_i^m \frac{y_i c_i}{\gamma_1} = \lambda \qquad (3)$$

and the constraint,

$$\alpha_1 + \beta_1 + \gamma_1 = 1 \qquad (4)$$

Combining (1) (2) (3) we get:

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

We know that $\alpha_1 + \beta_1 + \gamma_1 = 1$ (4), and that $a_i + b_i + c_i = n$ given in the problem,therefore we get:

$$\lambda = n \sum_i^m y_i$$

Putting this back into our original $\alpha$ equation:

$$\sum_i^m \frac{y_i a_i}{\alpha_1} = n \sum_i y_i$$

$$\frac{\sum_i^m a_i}{\alpha_1} = n$$

$$\alpha_1 = \frac{\sum_i^m y_i a_i}{n \sum_i y_i}$$

The numerator counts how many total times the word $a$ showed up in all of the good documents, and the denominator counts the total number of words in all good documents. This makes $\alpha_1$ the probability that the word $a$ is in a good document.

We can do a similar procedure for $\alpha_0$ which will yield $\frac{\sum_i a_i(1-y_i)}{n \sum_i 1-y_i}$, which also makes sense for the same reason ($1 - y_i$ is always the opposite of $y_i$). Using symmetry, we get the remaining results:

$$\beta_1 = \frac{\sum_i^m y_i b_i}{n \sum_i y_i}$$

$$\gamma_1 = \frac{\sum_i^m y_i c_i}{n \sum_i y_i}$$

$$\beta_0 = \frac{\sum_i^m b_i(1 - y_i)}{n \sum_i 1 - y_i}$$

$$\gamma_0 = \frac{\sum_i^m c_i(1 - y_i)}{n \sum_i 1 - y_i}$$

4. Dice Roll
   Given: A sequence `3463661622`
   probability p of occurrence 6 on the dice
   other numbers are equally likely to appear ⇒ $Pr(1) = Pr(2) = Pr(3) = Pr(4) = Pr(5) = \frac{1-p}{5}$
   There are two cases:

1 Since, on first appearance of 6 dice is rolled again and user is shown the number on the second role so probability of 6 appearing in the sequence is given by: $Pr(6) = p^2$ ,as for 6 to appear two consecutive roles should have 6 on the face up and probability of 6 is p so for consecutive 6's it will be p*p.

2 Numbers 1 → 5 can appear in the sequence in two ways:

   i. They appear on the first throw, which is given by $\frac{1-p}{5}$
   or

   ii. they appear followed by 6 on the first role, which is given by $\frac{p*(1-p)}{5}$

   $\Rightarrow Pr(1or2or3or4or5) = \frac{1-p}{5} + \frac{p*(1-p)}{5} = \frac{1-p^2}{5}$

Thus, for the given sequence likelihood is given by:

$$P(\vec{x}) = \left(\frac{1-p^2}{5}\right)^6 * \left(p^2\right)^4$$

$$\frac{d(P(\vec{x}))}{dp} = 0$$
$$\Rightarrow -12(1-p^2)^5p^9 + 8(1-p^2)^6p^7 = 0$$
$$\Rightarrow p^7(1-p^2)^5(8-20p^2) = 0$$

Solving we will get $p = 0, 1, -1, \sqrt{\frac{2}{5}}, -\sqrt{\frac{2}{5}}$
p can not be negative or 0 or 1

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