CS446: Machine Learning **Fall 2014**

Problem Set 6

Nikhil Unni *Handed In: November 18, 2014*

1. Naïve Bayes and Learning Threshold Functions

   (a) If we make our weight vector $w = [1, 1, 1, 1, 1, 1, 1]^T$, and $\theta = -3$ such that $y = 1$ if $w^T x + \theta \geq 0$, and $y = 0$ if $w^T x + \theta < 0$, then if any 3 or more components are 1, then $y = 1$

   (b) Let x be the input over the 7 dimension cube. Then, as we learned from class $h(x) = \arg\max_{y \in \{0,1\}} P(y) \prod_{i=0}^{n} P(x_i|y)$. Since we know we're sampling from a uniform distribution, and we have a lot of examples, we can estimate these probabilities.

   Also noting that $P(y)$ actually means $P(Y = y)$, and I'll use this syntax throughout, for other probabilities, too.

$$\arg\max_{y \in \{0,1\}} P(y) \prod_{i=0}^{7} P(x_i|y)$$

$$= \arg\max_{y}(P(0)P(x_1|0)P(x_2|0) \ldots P(x_7|0),$$
$$P(1)P(x_1|1)P(x_2|1) \ldots P(x_7|1))$$

   Breaking this up, $P(0)$ is just the probability that out of the 7 features, at least 5 of them are 0. This is just:

$$P(0) = \binom{7}{5}0.5^7 + \binom{7}{6}0.5^7 + \binom{7}{7}0.5^7 = \frac{29}{128}$$

   This makes $P(1) = \frac{99}{128}$.

   Next:

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

   This is for all x_i since they're all the same, since they're sampled from the uniform distribution. The probability $P(0|x_i)$ is split up into two probabilities, depending on the value of x_i. If $x_i = 1$, then it's the probability that at least 5 of the remaining 6 features are 0. Similarly, $P(0|x_i = 0)$ is the probability that at least 4 of the remaining 6 features are 0.

$$P(0|x_i = 1) = \binom{6}{5}0.5^6 + \binom{6}{6}0.5^6 = \frac{7}{64}$$

$$P(0|x_i = 0) = \binom{6}{4}0.5^6 + \binom{6}{5}0.5^6 + \binom{6}{6}0.5^6 = \frac{22}{64}$$

Because of this, we also know $P(1|x_i = 1) = \frac{57}{64}$ and $P(1|x_i = 0) = \frac{42}{64}$.
So we have:

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

$$= \frac{1/2 * \{7/64 \text{ or } 22/64\}}{29/128}$$

$$= \frac{7}{29} \text{ if } x_i = 1 \text{ or } \frac{22}{29} \text{ if } x_i = 0$$

Similarly for $P(x_i|1)$:

$$P(x_i|1) = \frac{1/2 * \{57/64 \text{ or } 42/64\}}{99/128}$$

$$= \frac{57}{99} \text{ if } x_i = 1 \text{ or } \frac{42}{99} \text{ if } x_i = 0$$

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

$$\frac{29}{128}\prod_{i=1}^{7}\{\frac{7}{29} \text{ if } x_i = 1, \text{ or } \frac{22}{29} \text{ if } x_i = 0\} \text{ if y} = 0$$

$$\frac{99}{128}\prod_{i=1}^{7}\{\frac{57}{99} \text{ if } x_i = 1, \text{ or } \frac{42}{99} \text{ if } x_i = 0\} \text{ if y} = 1$$

Or, the more cutesy representation:

$$h(x) = \arg\max_{y\in\{0,1\}} \frac{29 + 70y}{128}\prod_{i=1}^{7}\frac{42 + 15x_i}{99}y - \frac{22 - 15x_i}{29}(y - 1)$$

(c) We can show this by a simple proof by contradiction. If we assume the final hypothesis **does** represent our function, then our hypothesis and the function should output the same output given the same input, for all inputs. If we had an input $x = [1, 1, 0, 0, 0, 0, 0]^T$, by our original function, the output should be 0, since there are only 2 features that have a value of 1. But using this in our hypothesis, the value for $y = 0$ is:

$$\frac{29}{128} * \frac{7}{29}^2 * \frac{22}{29}^5 = .003316741$$

The value for $y = 1$ is:

$$\frac{99}{128} * \frac{57^2}{99} * \frac{42}{99}^5 = .003523506$$

The value for $y = 1$ is slightly 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, they're not. 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 $P(x_2|0)$. This means 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 3 other features with values of 1, $P(0|x_1)$ should always be 0, regardless of the value of x_1.

In a more hand-wavy explanation, because the value is just dependent on a certain threshold number of features being on, it'll always be 1 regardless of the remaining features. This means that the remaining features should have no affect on the final output, but with the Naïve Bayes assumptions, they **will** have an affect.

2. Naïve Bayes over Multinomial Distribution

(a)

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

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

Together,

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

Then, if we let $\eta = P(y_i = 1)$,

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

$$log[P(D_i, y_i)] = log(n!) - log(a_i!b_i!c_i!) + y_i[log(\eta) + a_i log(\alpha_1) + b_i log(\beta_1) + c_i log(\eta_1)]$$
$$+(1-y_i)[log(1-\eta) + a_i log(\alpha_0) + b_i log(\beta_0) + c_i log(\eta_0)]$$

(b) For α_1:

Using Lagrange multipliers, we have the function we want to maximize, $f(\alpha_1, \beta_1, \gamma_1) = \sum_{i=1}^{m} log[P(D_i, y_i)]$, and our constraint, $g(\alpha_1, \beta_1, \gamma_1) = \alpha_1 + \beta_1 + \gamma_1$. So we get 4 equations:

$$(1)\sum_i \frac{y_i a_i}{\alpha_1} = \lambda$$

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

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

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

Just as a note, all summations will be from 1 to m, even when I don't explicitly give the limits.

Combining the top 3 equations, and moving terms around, we get:

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

We know that $\alpha_1 + \beta_1 + \gamma_1 = 1$ from our constraint, and that $a_i + b_i + c_i = n$ because they're all the counts of the only feature types, so they have to sum to the number of total features. This'll give us:

$$\lambda = n \sum_i y_i$$

Putting this back into our original α equation:

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

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

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

Just as a sanity check, this result does make sense. 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 features in all good documents. This makes α_1 the probability that the word a is in a good document.

We can do a similar procedure for α_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 (keeping in mind $1 - y_i$ is always the opposite of y_i). Using symmetry, we get the remaining results:

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

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

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

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

3. Multivariate Poisson Naïve Bayes

(a) I'm just going to follow the same steps as problem 2 to find the MLE of the parameters.

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

So for a given example, (x_1, x_2, y), where y is either 0, if its value is A, or 1 if its value is B (I'll use this notation throughout), the probability of an example is:

$$P(x_1, x_2, y = 0) = \frac{e^{-\lambda_i^A}(\lambda_i^A)^{x_1}}{x_1!} * \frac{e^{-\lambda_i^A}(\lambda_i^A)^{x_2}}{x_2!} * P(Y = A)$$

Or,

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

Putting it all together...

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

Taking the log of this, and combining constants:

$$logP(x_1, x_2, y) = (1 - y)[(-\lambda_1^A - \lambda_2^A) + x_1log(\lambda_1^A) + x_2log(\lambda_2^A) + C]$$
$$+y[(-\lambda_1^B - \lambda_2^B) + x_1log(\lambda_1^B) + x_2log(\lambda_2^B) + C']$$

So the probability of the entire data is:

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

For λ_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. With this in mind, we can drop the $A \to 0$, $B \to 1$ notation.

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

Going through the actual data:

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

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

Similarly, for λ_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,

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

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 $X_1 = 2$, $X_2 = 3$.

4. Coin Toss To get a T, we could've either gotten a T the first time, or gotten an H then a T. And the only way to get an H is to get two H's in a row.

$$P(T) = (1-p) + p(1-p) = 1 - p^2$$

$$P(H) = p^2$$

Since we got 6 T's and 4 H's, and we're using a Bernoulli model:

$$P(data) = (1-p^2)^6(p^2)^4$$

$$\begin{aligned}\frac{dP(data)}{dp} &= 6(1-p^2)^5(-2p)(p^2)^4 + 4(p^2)^3(2p)(1-p^2)^6 \\ &= p^7(1-p^2)^5(20p^2-8) = 0\end{aligned}$$

So $p = 0, -1, 1, \sqrt{\frac{2}{5}}, -\sqrt{\frac{2}{5}}$

Because p can't be negative or 0 or 1 (since we got both H's and T's), $p = \sqrt{\frac{2}{5}}$ (most likely).