# CS 4780/5780 Homework 4

Due: Sunday 10/18/18 11:55pm on Gradescope

## Problem 1: Intuition for Naive Bayess

### a)

**Solution:** From Bayes Theorem, we have:

$$P(y = R|x = H) = \frac{P(x = H|y = R)P(y = R)}{P(x = H)}$$

We can easily calculate $P(x = H|y = R)$ and $P(x = H)$:

$$P(x = H|y = R) = \frac{3}{5} = 0.6$$
$$P(y = R) = \frac{1}{2} = 0.5$$

To calculate $P(x = H)$ we can use the law of total probability to get:

$$\begin{aligned} P(x = H) &= P(y = R)P(x = H|y = R) + P(\neg(y = R))P(x = H|\neg(y = R)) \\ &= P(y = R)P(x = H|y = R) + P(y = B)P(x = H|y = B) \\ &= \left(\frac{1}{2}\right)\left(\frac{3}{5}\right) + \left(\frac{1}{2}\right)\left(\frac{7}{10}\right) \end{aligned}$$

Now we can calculate $P(y = R|x = H)$ to get:

$$\frac{6}{13} = 0.462$$

### b)

**Solution:** We are trying to find:

$$P(y = R|\mathbf{x} = [H, H, T, H])$$

Using Bayes' Theorem, we can have:

$$\begin{aligned} P(y = R|\mathbf{x} = [H, H, T, H]) &= \frac{P(\mathbf{x} = [H, H, T, H]|y = R)P(y = R)}{P(\mathbf{x} = [H, H, T, H])} \\ &= \frac{P(\mathbf{x} = [H, H, T, H]|y = R)P(y = R)}{P(\mathbf{x} = [H, H, T, H]|y = R)P(y = R) + P(\mathbf{x} = [H, H, T, H]|y = B)P(y = B)} \\ &= \frac{(\frac{3}{5} \times \frac{3}{10} \times \frac{1}{2} \times \frac{4}{5}) \times (\frac{1}{2})}{(\frac{3}{5} \times \frac{3}{10} \times \frac{1}{2} \times \frac{4}{5}) \times (\frac{1}{2}) + (\frac{7}{10} \times \frac{1}{5} \times \frac{9}{10} \times \frac{2}{5}) \times (\frac{1}{2})} \\ &= \frac{10}{17} \end{aligned}$$

### c)

**Solution:**
i. Estimation with +1 additive smoothing with 2 categories is as follows:

$$[\hat{\theta}_{jc}]_\alpha = \frac{\sum_{i=1}^{18} I(y_i = c)I(x_{ia} = j) + 1}{\sum_{i=1}^{18} I(y_i = c) + 2}$$

Now recalculating the probabilities we get:

$$P(penny = H|hat = Red) = \frac{7}{10} = 0.700$$
$$P(nickel = H|hat = Red) = \frac{8}{10} = 0.800$$
$$P(dime = H|hat = Red) = \frac{5}{10} = 0.500$$
$$P(quarter = H|hat = Red)\frac{2}{10} = 0.200$$
$$P(penny = H|hat = Blue) = \frac{2}{12} = 0.167$$
$$P(nickel = H|hat = Blue) = \frac{4}{12} = 0.333$$
$$P(dime = H|hat = Blue) = \frac{10}{12} = 0.833$$
$$P(quarter = H|hat = Blue)\frac{5}{12} = 0.417$$

Therefore,

| hat | $P(penny = H\|hat)$ | $P(nickel = H\|hat)$ | $P(dime = H\|hat)$ | $P(quarter = H\|hat)$ |
|---|---|---|---|---|
| Red | 0.700 | 0.800 | 0.500 | 0.200 |
| Blue | 0.167 | 0.333 | 0.833 | 0.417 |

ii. Since we are not given the probabilities of the coins outright, we have to estimate them. That is we can estimate them using MLE. Additionally, From Bayes' Theorem, we know that :

$$\begin{aligned}
P(y = R|\mathbf{x} = [H, H, T, H]) &= \frac{P(\mathbf{x} = [H, H, T, H]|y = R)P(y = R)}{P(\mathbf{x} = [H, H, T, H])} \\
&= \frac{P(\mathbf{x} = [H, H, T, H]|y = R)P(y = R)}{P(\mathbf{x} = [H, H, T, H])} \\
&= \frac{\prod_{\alpha=1}^{k}[\hat{\theta}_{jR}]_\alpha P(y = R)}{\prod_{\alpha=1}^{k}[\hat{\theta}_{jR}]_\alpha P(y = R) + \prod_{\alpha=1}^{k}[\hat{\theta}_{jB}]_\alpha P(y = B)} \\
&= \frac{\prod_{\alpha=1}^{k}[\hat{\theta}_{jR}]_\alpha}{\prod_{\alpha=1}^{k}[\hat{\theta}_{jR}]_\alpha + \prod_{\alpha=1}^{k}[\hat{\theta}_{jB}]_\alpha}, \text{where } j \in [H, H, T, H]
\end{aligned}$$

Now to estimate the probability of a feature $\alpha$ has value $j$ given the label is $c$, $[\hat{\theta}_{jc}]_\alpha$:

$$[\hat{\theta}_{jc}]_\alpha = \frac{\sum_{i=1}^{18} I(y_i = c)I(x_{ia} = j)}{\sum_{i=1}^{18} I(y_i = c)}$$
$$[\hat{\theta}_{HR}]_{penny} = \frac{3}{4} = 0.75$$
$$[\hat{\theta}_{HR}]_{nickel} = \frac{7}{8} = 0.875$$
$$[\hat{\theta}_{TR}]_{dime} = \frac{1}{2} = 0.5$$
$$[\hat{\theta}_{HR}]_{quarter} = \frac{1}{8} = 0.125$$
$$[\hat{\theta}_{HB}]_{penny} = \frac{1}{10} = 0.1$$
$$[\hat{\theta}_{HB}]_{nickel} = \frac{3}{10} = 0.3$$
$$[\hat{\theta}_{TB}]_{dime} = \frac{1}{10} = 0.1$$
$$[\hat{\theta}_{HB}]_{quarter} = \frac{4}{10} = 0.4$$

Finally, taking the products of the probabilities we get:

$$\frac{4375}{4503} = 0.972$$

## d)

**Solution:** Feature space $\mathcal{X} \in \{H, T\}^4$. Label space $\mathcal{Y} \in \{Red, Blue\}$. The Naive Bayess assumption is valid because any two coin flips are independent of each other given a label.

# Problem 2: Linearity of Gaussian Naive Bayess

## a)

**Solution:**
First, note that the numerator follows immediately from Bayes' rule, we have just substituted the actual given value $y = 1$ for $y$ in the first equation in this second. For the denominator, we expand $p(\mathbf{x})$ using first the sum rule and the product rule. By the sum rule, $p(\mathbf{x}) = p(\mathbf{x}, y = 1)+p(\mathbf{x}, y = 0)$. Applying the product rule to both terms on the right hand side, we get:

$$p(\mathbf{x}) = p(\mathbf{x}|y = 1)p(y = 1) + p(\mathbf{x}|y = 0)p(y = 0)$$

Next, we apply the Naive Bayess' assumption to $p(\mathbf{x}|y = 1)$ and $p(\mathbf{x}|y = 0)$ to get:

$$p(\mathbf{x}) = \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 1)p(y = 1) + \prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 0)p(y = 0)$$

Plugging this in for the denominator in Bayes' rule, we achieve the desired result.

## b)

**Solution:**
Observe that, in general, $\frac{a}{a+b}$ can equivalently be written as $\frac{1}{1+\frac{b}{a}}$. Furthermore, the equation for $p(y = 1|\mathbf{x})$ derived in the previous part has exactly the form $\frac{a}{a+b}$! Therefore, we can rewrite it as:

$$p(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=0)p(y=0)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=1)p(y=1)}}$$

Next, since $\exp(\log(\mathbf{x})) = \mathbf{x}$,

$$p(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left(\log\left(\frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=0)p(y=0)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=1)p(y=1)}\right)\right)}$$

Finally, pulling a negative sign out of the log lets us flip the fraction inside:

$$p(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-\log\frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=1)p(y=1)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=0)p(y=0)}\right)}$$

## c)

**Solution:**
To show this, we simply plug in the following definitions to the equation we derived in part b:

$$\begin{aligned}
p(y = 1) &= \rho \\
p([\mathbf{x}]_\alpha \mid y = 1) &= \frac{1}{\sqrt{2\pi[\sigma]_\alpha}} \exp\left(\frac{-([\mathbf{x}]_\alpha - [\mu_1]_\alpha)^2}{2[\sigma]_\alpha}\right) \\
p([\mathbf{x}]_\alpha \mid y = 0) &= \frac{1}{\sqrt{2\pi[\sigma]_\alpha}} \exp\left(\frac{-([\mathbf{x}]_\alpha - [\mu_0]_\alpha)^2}{2[\sigma]_\alpha}\right)
\end{aligned}$$

Expanding $-\log\frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=1)p(y=1)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y=0)p(y=0)}$ we get:

$$-\log p(y = 1) - \log\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 1) + \log p(y = 0) + \log\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha|y = 0)$$

Observing that $\log\prod_i \mathbf{x}_i = \sum_i \log\mathbf{x}_i$ and rearranging terms, this is equal to:

$$\log\frac{p(y = 0)}{p(y = 1)} + \sum_{\alpha=1}^{d} \log\frac{p([\mathbf{x}]_\alpha|y = 0)}{p([\mathbf{x}]_\alpha|y = 1)}$$

Plugging in the definition of $p(y=1)$, the first term in this is equal to $\log \frac{1-\rho}{\rho}$.
For the second term, we plug in the Gaussian distributions for $p([\mathbf{x}]_\alpha \mid y=1)$ and $p([\mathbf{x}]_\alpha \mid y=0)$, and then do a bit of algebra to get:

$$\sum_{\alpha=1}^{d} \frac{([\mu_0]_\alpha - [\mu_1]_\alpha)[\mathbf{x}]_\alpha}{[\sigma]_\alpha} + \frac{[\mu_1]_\alpha^2 - [\mu_0]_\alpha^2}{2[\sigma]_\alpha}$$

Putting everything together we get:

$$\log \frac{1-\rho}{\rho} + \sum_{\alpha=1}^{d} \frac{([\mu_0]_\alpha - [\mu_1]_\alpha)[\mathbf{x}]_\alpha}{[\sigma]_\alpha} + \frac{[\mu_1]_\alpha^2 - [\mu_0]_\alpha^2}{2[\sigma]_\alpha}$$

And finally we just start renaming terms. Let's first define:

$$b = \log \frac{1-\rho}{\rho} + \sum_{\alpha=1}^{d} \frac{[\mu_1]_\alpha^2 - [\mu_0]_\alpha^2}{2[\sigma]_\alpha}$$

Next, create a vector $\mathbf{w}$ so that:

$$[\mathbf{w}]_\alpha = \frac{[\mu_0]_\alpha - [\mu_1]_\alpha}{[\sigma]_\alpha}$$

Then the sum (the second term) is simply equal to $\mathbf{w}^\top \mathbf{x}$. Therefore,

$$-\log \frac{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha | y=1) p(y=1)}{\prod_{\alpha=1}^{d} p([\mathbf{x}]_\alpha | y=0) p(y=0)} = \mathbf{w}^\top \mathbf{x} + b$$

Plugging this in to the decision rule $p(y=1|\mathbf{x})$ we derived in part b, we finally see that:

$$p(y=1|\mathbf{x}) = \frac{1}{1+\exp\left(\mathbf{w}^\top \mathbf{x} + b\right)}$$

Notice: This is not only a linear decision boundary, but should look very similar indeed to the linear decision rule you've seen from logistic regression.