# CS4780/5780 Homework 4 Solutions

September 28, 2018

## 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 $P(x = H)$:

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

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:

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

## 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.

## Problem 3: Gradient for Logistic Regression

1.

$$\begin{aligned}
\sigma(-s) &= \frac{1}{1+e^s} \\
&= \frac{e^{-s}}{e^{-s}(1+e^s)} \\
&= \frac{e^{-s}}{e^{-s}+1} \\
&= \frac{e^{-s}+1-1}{e^{-s}+1} \\
&= \frac{e^{-s}+1}{e^{-s}+1} - \frac{1}{e^{-s}+1} \\
&= 1 - \frac{1}{e^{-s}+1} \\
&= 1-\sigma(s)
\end{aligned}$$

2. (a)

$$\begin{aligned}
\sigma'(s) &= \frac{d}{ds}(\frac{1}{1+e^{-s}}) \\
&= \frac{d}{ds}(1+e^{-s}) \cdot (-(1+e^{-s})^{-2}) \\
&= (-e^{-s}) \cdot (-(1+e^{-s})^{-2}) \\
&= \frac{e^{-s}}{(1+e^{-s})^2} \\
&= \frac{1}{1+e^{-s}} \cdot \frac{e^{-s}}{1+e^{-s}} \\
&= \frac{1}{1+e^{-s}} \cdot \frac{e^{-s}+1-1}{1+e^{-s}} \\
&= \frac{1}{1+e^{-s}} \cdot (\frac{e^{-s}+1}{1+e^{-s}} - \frac{1}{1+e^{-s}}) \\
&= \frac{1}{1+e^{-s}} \cdot (1 - \frac{1}{1+e^{-s}}) \\
&= \sigma(s)(1-\sigma(s))
\end{aligned}$$

(b) Before we find the gradient, let's first write down the log likelihood function

$$\log P(\mathbf{y}|X,\mathbf{w}) = \log \prod_{i=1}^{n} \sigma(y_i(w^T\mathbf{x}_i))) = \sum_{i=1}^{n} \log \sigma(y_i(w^T\mathbf{x}_i)))$$

where in the last equality, we use the property of the logarithm function. To find the gradient, we will first find the k-th entry of the gradient. By definition, the k-th entry of the gradient is

$$\begin{aligned}\frac{\partial}{\partial w_k}\log P(\mathbf{y}|X,\mathbf{w}) &= \sum_{i=1}^{n} \frac{\partial}{\partial w_k}\log(\sigma(y_i(w^T\mathbf{x}_i))) \\ &= \sum_{i=1}^{n} \frac{\sigma(y_i(w^T\mathbf{x}_i))(1-\sigma(y_i(w^T\mathbf{x}_i)))}{\sigma(y_i(w^T\mathbf{x}_i))} y_i x_{ik} \\ &= \sum_{i=1}^{n}(1-\sigma(y_i(w^T\mathbf{x}_i)))y_i x_{ik}\end{aligned}$$

where in the 2nd step, we apply the Chain rule. Now, using the partial derivative, we know that

$$\begin{aligned}\nabla_w P(y|X,w) &= \sum_{i=1}^{n} \begin{bmatrix} \frac{\partial log(\sigma(y_i(w^T\mathbf{x}_i)))}{\partial w_1} \\ \vdots \\ \frac{\partial log(\sigma(y_i(w^T\mathbf{x}_i)))}{\partial w_d} \end{bmatrix} \\ &= \sum_{i=1}^{n} \begin{bmatrix} (1-\sigma(y_i(w^T\mathbf{x}_i)))y_i x_{i1} \\ \vdots \\ (1-\sigma(y_i(w^T\mathbf{x}_i)))y_i x_{id} \end{bmatrix} \\ &= \sum_{i=1}^{n}(1-\sigma(y_i(w^T x_i + b)))y_i\mathbf{x}_i\end{aligned}$$

## Problem 4: Optimization with Gradient Descent

(1) $f(w_5) = 80$

| $i$ | $w_i$ |
|---|---|
| 0 | 13 |
| 1 | 9 |
| 2 | 13 |
| 3 | 9 |
| 4 | 13 |
| 5 | 9 |

(2)

| $i$ | $w_i$ |
|---|---|
| 0 | 13 |
| 1 | 11 |

## Problem 5: Linear Regression

Consider we have the following 1-d training set:

| $x$ | $y$ |
|---|---|
| -2 | 7 |
| -1 | 4 |
| 0 | 3 |
| 1 | 4 |
| 2 | 7 |

and our goal is to find a regression model that could regress $x$ to our target value $y$. To do this, we are going to use a linear regression model. Namely, we are going to model our data by assuming the relationship

$$\begin{aligned} y &= w_1 x + w_0 + \epsilon \\ &= \mathbf{w}^T \phi(x) + \epsilon \end{aligned}$$

where $\phi(x) = [1, x]^T$. We call $\phi$ a feature mapping of $x$ and this feature mapping allows us to absorb the bias $w_0$ into the vector $\mathbf{w}$.

1. With this feature mapping, we can write down the design matrix as

   $$X = [\phi(x_1)...\phi(x_n)]$$

   Using the formula given in class, compute the closed form solution for $\mathbf{w}$. Even though you are not required to calculate the inverse by hand, we strongly encourage you to do so since we expect you to be able to calculate the inverse of a 2 × 2 and 3 × 3 matrix by hand.

   **Solution** From class,

   $$\mathbf{w} = (XX^\top)^{-1} X \mathbf{y}^\top$$

   . We are given,

   $$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 7 \\ 4 \\ 3 \\ 4 \\ 7 \end{bmatrix}$$

   Hence,

   $$\begin{aligned} XX^\top &= \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \\ (XX^\top)^{-1} &= \begin{bmatrix} 0.2 & 0 \\ 0 & 0.1 \end{bmatrix} \end{aligned}$$

Remember, given $AA^{-1} = \mathbf{I}$,

$$A^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Applying matrix multiplication, we find

$$\mathbf{w} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

2. Recall that the loss function for linear regression is

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \phi(x_i))^2$$

With the closed formed solution obtained in (a)(1), calculate the training loss.

**Solution** Directly applying the loss function,

$$\begin{aligned} \ell(\mathbf{w}) &= \sum_{i=1}^{5} (y_i - \mathbf{w}^T \phi(x_i))^2 \\ &= (7 - \begin{bmatrix} 5 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix})^2 + (4 - \begin{bmatrix} 5 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix})^2 + ... + (7 - \begin{bmatrix} 5 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix})^2 \\ &= 14. \end{aligned}$$