1.2

To solve for $\hat{p}$ our goal is to maximize the log-likelihood of observing the data.$argmax_p \sum_{i=1}^{m} logp(x_i|y)$. We know that $P(y=1)=p$. This gives us the problem $argmax_p \sum_{i=1}^{m}(\frac{1+y_i}{2})lg(pPr(x_i|y_i=1)) + (\frac{1-y_i}{2})lg((1-p)Pr(x_i|y_i=-1))$. Taking the derivative with respect to p:

$$\frac{d}{dp}\sum_{i=1}^{m}(\frac{1+y_i}{2})lg(pPr(x_i|y_i=1)) + (\frac{1-y_i}{2})lg((1-p)Pr(x_i|y_i=-1))$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2})\frac{1}{(pPr(x_i|y_i=1))}(Pr(x_i|y_i=1))+(\frac{1-y_i}{2})\frac{1}{((1-p)Pr(x_i|y_i=-1))}(-Pr(x_i|y_i=-1))$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2})\frac{1}{p} + (\frac{1-y_i}{2})\frac{1}{(1-p)}$$

Set this equal to 0:

$$\sum_{i=1}^{m}(\frac{1+y_i}{2})\frac{1}{p} + (\frac{1-y_i}{2})\frac{1}{(1-p)} = 0$$

$$(1-p)\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)\sum_{i=1}^{m}(\frac{1-y_i}{2}) = 0$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)\sum_{i=1}^{m}(\frac{1-y_i}{2}) = 0$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)\sum_{i=1}^{m}(\frac{1+y_i}{2}) + (\frac{1-y_i}{2}) = 0$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)\sum_{i=1}^{m}(\frac{1}{2}) + (\frac{1}{2}) = 0$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)\sum_{i=1}^{m}1 = 0$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2}) - (p)(m) = 0$$

$$\frac{\sum_{i=1}^{m}(\frac{1+y_i}{2})}{m} = p$$

We know that$\sum_{i=1}^{m}(\frac{1+y_i}{2})$ equals the number of occasions on which $y_i=1$ because if $y_i=1$, $\frac{1+y_i}{2}=1$ and if $y_i=0$, $\frac{1+y_i}{2}=0$. Thus, the $\hat{p}=\frac{\sum_{i=1}^{m}(\frac{1+y_i}{2})}{m}$ means that $\hat{p}$ is given by the number of occurrences of $y_i=1$ divided by the total number of data points, which intuitively makes sense because $p=Pr(y_i=1)$.

To solve for $\hat{\alpha_i}$, our goal is to maximize the log-likelihood of observing the data with respect to α_i.

$$\frac{d}{d\alpha_i}\sum_{i=1}^{m}(\frac{1+y_i}{2})lg(pPr(x_i|y_i=1))+(\frac{1-y_i}{2})lg((1-p)Pr(x_i|y_i=-1))$$

$$\frac{d}{d\alpha_i}\sum_{i=1}^{m}(\frac{1+y_i}{2})lg(p(\prod_{j=1}^{n}\alpha_j^{x_j}(1-\alpha_j)^{1-x_j})+(\frac{1-y_i}{2})lg((1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j}))$$

$$\frac{d}{d\alpha_i}\sum_{i=1}^{m}(\frac{1+y_i}{2})[lg(p)+\sum_{j=1}^{n}lg(\alpha_j^{x_j}(1-\alpha_j)^{1-x_j})]+(\frac{1-y_i}{2})lg((1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j}))$$

$$\frac{d}{d\alpha_i}\sum_{i=1}^{m}(\frac{1+y_i}{2})[lg(p)+\sum_{j=1}^{n}(lg(\alpha_j^{x_j})+lg(1-\alpha_j)^{1-x_j}))]+(\frac{1-y_i}{2})lg((1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j}))$$

$$\frac{d}{d\alpha_i}\sum_{i=1}^{m}(\frac{1+y_i}{2})[lg(p)+\sum_{j=1}^{n}(x_jlg(\alpha_j)+(1-x_j)lg(1-\alpha_j))]+(\frac{1-y_i}{2})lg((1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j}))$$

$$\frac{d}{d\alpha_i}\sum_{i=1}^{m}(\frac{1+y_i}{2})[lg(p)+\sum_{j=1}^{n}(x_jlg(\alpha_j)+(1-x_j)lg(1-\alpha_j))]+(\frac{1-y_i}{2})lg((1-p)(\prod_{j=1}^{n}\beta_j^{x_j}(1-\beta_j)^{1-x_j}))$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2})[\sum_{j=1}^{n}\frac{d}{d\alpha_i}(x_jlg(\alpha_j)+(1-x_j)lg(1-\alpha_j))]$$

$$\sum_{i=1}^{m}(\frac{1+y_i}{2})[\sum_{j=1}^{n}(x_j\frac{1}{\alpha_i})-(1-x_j)\frac{1}{1-\alpha_i})]$$

Setting this derivative equal to 0, we get:

$$\sum_{i=1}^{m}(\frac{1+y_i}{2})[(\sum_{j=1}^{n}x_j)(\frac{1}{\alpha_i})-(\sum_{j=1}^{n}1-x_j)(\frac{1}{1-\alpha_i})]=0$$

$$(\sum_{i=1}^{m}\frac{1+y_i}{2})(1-\alpha_i)(\sum_{j=1}^{n}x_j)=(\sum_{i=1}^{m}\frac{1+y_i}{2})(\sum_{j=1}^{n}1-\sum_{j=1}^{n}x_j)(\alpha_i)$$

$$(\sum_{i=1}^{m}\frac{1+y_i}{2})(\sum_{j=1}^{n}x_j)-(\sum_{i=1}^{m}\frac{1+y_i}{2})(\alpha_i)(\sum_{j=1}^{n}x_j)=(\sum_{i=1}^{m}\frac{1+y_i}{2})(\sum_{j=1}^{n}1)-(\sum_{i=1}^{m}\frac{1+y_i}{2})(\sum_{j=1}^{n}x_j)(\alpha_i)$$

$$(\sum_{i=1}^{m}\frac{1+y_i}{2}\sum_{j=1}^{n}x_j)=(\sum_{i=1}^{m}\frac{1+y_i}{2}\sum_{j=1}^{n}1)(\alpha_i)$$

$$\frac{(\sum_{i=1}^{m}\frac{1+y_i}{2}\sum_{j=1}^{n}x_j)}{(\sum_{i=1}^{m}\frac{1+y_i}{2}\sum_{j=1}^{n}1)}=(\hat{\alpha_i})$$

This intuitively means that $\hat{\alpha}_i$ is the count of observations in class $y = 1$ with attribute $x_i = 1$ divided by the total count of observations in class $y = 1$.

By symmetry to the above, the $\hat{\beta}_i = \frac{(\sum_{i=1}^{m} \frac{1-y_i}{2} \sum_{j=1}^{n} x_j)}{(\sum_{i=1}^{m} \frac{1-y_i}{2} \sum_{j=1}^{n} 1)}$. $\hat{beta_i}$ is the count of observations in class $y = -1$ with attribute $x_i = 1$ divided by the total count of observations in class $y = -1$.

1.3

$$h(\vec{x}) = argmax_{y \in \{\pm 1\}} \hat{Pr}(y|\vec{x})$$

This returns the y class value that yields the highest probability for $\hat{Pr}(y|\vec{x})$. Want to prove that this is equivalent to the function $h'(\vec{x}) = sign(\hat{Pr}(1|\vec{x}) - \hat{Pr}(-1|\vec{x})$.

If $h(\vec{x})$ returns $y = 1$, then this means that $\hat{Pr}(y = 1|\vec{x}) > \hat{Pr}(y = -1|\vec{x})$. Let $x = \hat{Pr}(y = 1|\vec{x}) - \hat{Pr}(y = -1|\vec{x})$ in which $x > 0$. Then, $h'(\vec{x}) = sign(x)$ which thus returns $+1$. Meanwhile, if $h(\vec{x})$ returns $y = -1$, then this means that $\hat{Pr}(y = -1|\vec{x}) > \hat{Pr}(y = +1|\vec{x})$. Let $x = \hat{Pr}(y = -1|\vec{x}) - \hat{Pr}(y = +1|\vec{x})$ in which $x < 0$. Then, $h'(\vec{x}) = sign(x)$ which thus returns -1. Thus we can see that $h(\vec{x})$ and $h'(\vec{x})$ always return the same result, and thus they are equivalent.

1.4 We want to show that $h(\vec{x}) = sign(\vec{w}^T \vec{x} + b)$. First by first using Bayes rule and then the total probability formula:

$$\hat{Pr}(y = a|\vec{x}) = \frac{Pr(\vec{x}|y = a)Pr(y = a)}{Pr(\vec{x})}$$

We want the a class value that maximizes the probability:

$$argmax_y \hat{Pr}(y = a|\vec{x}) = argmax_y \frac{Pr(\vec{x}|y = a)Pr(y = a)}{Pr(\vec{x})}$$

Since the denominator of this does not depend on y, this is equivalent to:

$$argmax_y Pr(\vec{x}|y = a)Pr(y = a)$$

From equation 2 we can rearrange to get:

$$log(Pr(\vec{x}|y = 1)) + log(Pr(y = 1)) - log(Pr(\vec{x}|y = -1)) - log(Pr(y = -1))$$

Because the activation of $log(\alpha_i)$ vs $log(\beta_i)$ depends on the value of $\vec{x}$, whether it's positive or 0, we can rearrange this to be:

$$(log(\alpha_i) - log(\beta_i))^T \vec{x} + (log(p) - log(1 - p))$$

Thus, we get that $\vec{w} = (log(\alpha_i) - log(\beta_i))$ and $b = (log(p) - log(1 - p))$.