Example Answer

3.1

a)

Naïve Bayes

$$P(Y = A|X_1 \dots X_{16}) = \frac{P(Y = A)P(X_1 \dots X_{16}|Y = A)}{P(Y = A)P(X_1 \dots X_{16}|Y = A) + P(Y = B)P(X_1 \dots X_{16}|Y = B)}$$

By conditional independence assumption of Naïve Bayes,

$$P(Y = A|X_1 \dots X_{16}) = \frac{P(Y = A)\prod_{i=1}^{16} P(X_i|Y = A)}{P(Y = A)\prod_{i=1}^{16} P(X_i|Y = A) + P(Y = B)\prod_{i=1}^{16} P(X_i|Y = B)}$$

Similarly for Y=B,

$$P(Y = B|X_1 \dots X_{16}) = \frac{P(Y = B)\prod_{i=1}^{16} P(X_i|Y = B)}{P(Y = A)\prod_{i=1}^{16} P(X_i|Y = A) + P(Y = B)\prod_{i=1}^{16} P(X_i|Y = B)}$$

Logistic Regression

$$P(Y = A|X_1 \dots X_{16}) = \frac{1}{1 + \exp\{w_0 + \sum_{i=1}^{16} w_i x_i\}}$$

$$P(Y = B|X_1 \dots X_{16}) = 1 - P(Y = A|X_1 \dots X_{16})$$

Note: there are some alternative answers for this question, for example

$$P(Y = A|X_1 \dots X_{16}) = \frac{1}{1 + \exp\{\mathrm{w}^T \mathrm{x}\}}, \text{where w} = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_{16} \end{bmatrix}, x = \begin{bmatrix} 1 \\ x_1 \\ \dots \\ x_{16} \end{bmatrix}$$

and

$$P(Y = A|X_1 \dots X_{16}) = \frac{1}{1 + \exp -\{w_0 + \sum_{i=1}^{16} w_i x_i\}}$$

All of them are correct, but one needs to be careful that different definition of logistic regression would end up with different answers in the next several questions.

(b)

Decision rule

Generally speaking, for both cases, if we have $(Y = A|X_1 \dots X_{16}) \geq P(Y = B|X_1 \dots X_{16})$ we will choose A; else, we will choose B. But as a homework question, we prefer more formal form as follows.

$$\arg \max_{y \in \{A,B\}} P(Y = y|X_1 \dots X_{16})$$

Or for logistic regression (if we define $P(Y = A|X_1 \dots X_{16}) = (1 + \exp\{w_0 + \sum_{i=1}^{16} w_i x_i\})^{-1}$)

Since we choose A if

$$\frac{P(Y=A|X_1 \dots X_{16})}{P(Y=B|X_1 \dots X_{16})} = \exp\left\{w_0 + \sum_{i=1}^{16} w_i x_i\right\}^{-1} \geq 1$$

We choose:

$$A, \qquad w_0 + \sum_{i=1}^{16} w_i x_i \leq 0$$

$$B, \qquad w_0 + \sum_{i=1}^{16} w_i x_i > 0$$

Please note that if one defines the logistic regression as $P(Y=A|X_1 \dots X_{16}) = (1 + \exp -\{w_0 + i=116wixi-1$, the result is just the opposite

(c)

Naïve Bayes

$$P(Y=y)_{y=\{A,B\}} = \frac{\#(Y=y)}{\#\text{samples}}$$

$$P(X_i = x_i|Y=y) = \frac{\#(X_i = x_i, Y = y_i)}{\#(Y=y)}$$

Logistic Regression

For purpose of simplicity, we will treat **Y=B as Y=1, and Y=A as Y=0**.

Hence the log likelihood of the data is: (where m is the number of training instances)

$$l(w) = \log\left(\prod_j^m \frac{exp^{\left(y^j\left(w_0+\sum_{i=1}^{16} w_i x_i^j\right)\right)}}{1 + exp^{\left(w_0+\sum_{i=1}^{16} w_i x_i^j\right)}}\right)$$

$$l(w) = \sum_j^m \left(y^j\left(w_0 + \sum_{i=1}^{16} w_i x_i^j\right) - log\left(1 + exp^{\left(w_0+\sum_{i=1}^{16} w_i x_i^j\right)}\right)\right)$$

Maximize the log-likelihood function using gradient descend. The gradients of the log likelihood function are:

$$\frac{\partial l(w)}{\partial w_0} = \sum_j^m \left(y^j - \frac{exp^{\left(w_0+\sum_{i=1}^{16} w_i x_i^j\right)}}{1 + exp^{\left(w_0+\sum_{i=1}^{16} w_i x_i^j\right)}}\right)$$

$$\frac{\partial l(w)}{\partial w_0} = \sum_j^m (y^j - P(Y=1|X = x^j))$$

$$\frac{\partial l(w)}{\partial w_k} = \sum_j^m \left(y^j x_k^j - \frac{x_k^j exp^{\left(w_0 + \sum_{i=1}^{16} w_i x_i^j\right)}}{1 + exp^{\left(w_0 + \sum_{i=1}^{16} w_i x_i^j\right)}} \right)$$

$$\frac{\partial l(w)}{\partial w_k} = \sum_j^m x_k^j (y^j - P(Y = 1|X = x^j))$$

The weight-updating rule is: (repeat until convergence, where α is the step size):

$$w_0^{(t+1)} = w_0^{(t)} + \alpha \sum_j^m \left(y^j - P(Y = 1|X = x^j, w^{(t)}) \right)$$

$$w_k^{(t+1)} = w_k^{(t)} + \alpha \sum_j^m x_k^j \left(y^j - P(Y = 1|X = x^j, w^{(t)}) \right)$$

Note that if defining the logistic regression as

$$P(Y = A|X_1 = x_1 \dots X_{16} = x_{16}) = \left(1 + \exp - \left\{ w_0 + \sum_{i=1}^{16} w_i x_i \right\} \right)^{-1}$$

Then we will treat **Y=A as Y=1, and Y=B as Y=0.**

3.2

Defining

$$P(A) = \pi_A$$
$$P(B) = \pi_B$$
$$P(X_i = 1 \mid Y = A) = \theta_{iA}$$
$$P(X_i = 1 \mid Y = B) = \theta_{iB}$$

We have

$$P(Y = A|X_1 \dots X_n) = \frac{P(X|Y = A)P(Y = A)}{P(X|Y = A)P(Y = A) + P(X|Y = B)P(Y = B)}$$

$$= \frac{1}{1 + \frac{P(X|Y = B)P(Y = B)}{P(X|Y = A)P(Y = A)}} = \frac{1}{1 + \frac{\pi_B \prod_i \theta_{iB}^{X_i} (1 - \theta_{iB})^{1 - X_i}}{\pi_A \prod_i \theta_{iA}^{X_i} (1 - \theta_{iA})^{1 - X_i}}}$$

$$= \frac{1}{1 + \exp \sum_i \left\{ X_i \log \frac{\theta_{iB}}{\theta_{iA}} + (1 - X_i) \log \frac{1 - \theta_{iB}}{1 - \theta_{iA}} + \log \frac{\pi_B}{\pi_A} \right\}}$$

$$= \frac{1}{1 + \exp \left\{ \log \frac{\pi_B}{\pi_A} + \sum_i \log \frac{(1 - \theta_{iB})}{(1 - \theta_{iA})} + \sum_i X_i \left(\log \frac{\theta_{iB}(1 - \theta_{iA})}{\theta_{iA}(1 - \theta_{iB})} \right) \right\}}$$

Define

$$w_i = \log \frac{\theta_{iB}(1-\theta_{iA})}{\theta_{iA}(1-\theta_{iB})}$$

$$w_0 = \log \frac{\pi_B}{\pi_A} + \sum_i \log \frac{(1-\theta_{iB})}{(1-\theta_{iA})}$$

we have

$$P(Y = A|X_1 \ldots X_n) = \frac{1}{1 + \exp\{w_0 + \sum_i w_i x_i\}}$$

which takes the form of logistic regression.

3.3

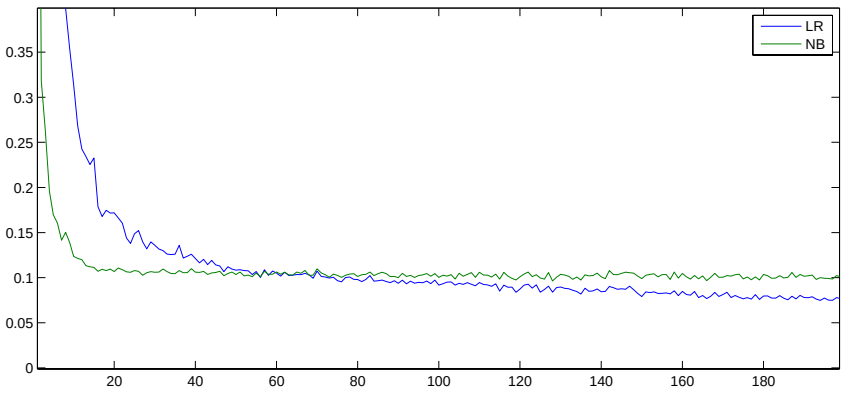
The key points are

a) If GNB modeling assumptions hold in the data (one assumption is conditional independence): they converge to identical classifiers; else they will typically learn different classifier functions.

b) No. Because logistic regression does not assume conditional independence. If the assumption does not hold in the data, the conditional log likelihood maximization algorithm for Logistic regression will adjust its parameters to maximize the fit to the conditional log likelihood of the data, even if the resulting parameters are inconsistent with the Naïve Bayes parameter estimates.

4.4

The result should look like this



If your classification accuracy is not around 90%, then there might be a problem in your algorithm implementation. If the error rate of logistic regression does not go below the error rate of naïve Bayes when increasing the data size, then there might be a problem in your step length (η may be too small or too large) or your stop condition. The logistic regression takes quite a long time to converge. But if you cannot have your result out in some hours, there might also be a problem in your step length and/or your stop condition.

4.5

The key points are:

a) Naïve Bayes
b) Logistic Regression
c) Prefer Naïve Bayes when having less data; and if we have more data, logistic regression is preferred. The reason is that (1) Naïve Bayes converges faster $o(n)$ to its optimal estimate; where logistic regression takes more data to converge. (2) Logistic regression does not assume "conditional independence" and thus will have higher classification accuracy if we have enough data.