Problem 1

1. a: formula for the conditional likelihood

$$NB: P(Y=1 \mid X_1, \dots X_p) = \frac{P(Y=1)\prod_{i=1}^{p} P(X_i \mid Y=1)}{\sum_{j=1}^{2} P(Y=y_j)\prod_{i=1}^{p} P(X_i \mid Y=y_j)}$$

$$LR: P(Y=1 \mid X_1, \dots X_p) = \frac{\exp(w_0 + \sum_{i=1}^{p} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{p} w_i X_i)}$$

b. classification rule:

NB: classification rule for $X^{new} = \langle X_1, \dots, X_n \rangle$ is: $Y^{new} \leftarrow \arg\max_{y_k} P(Y=y_k) \prod_i P(X_i^{new} \mid Y=y_k)$

LR: $w_0 + \sum_i w_i X_i \underset{0}{\overset{1}{\gtrless}} 0$

c. parameters we have to estimate

For NB: NB for discrete-valued inputs: have to estimate two sets of parameters:

first: $\boxed{\theta_{ijk} \equiv P(X_i = x_{ij} \mid Y=y_k)}$ for each input features X_i, each of its possible values x_{ij}, and each of the possible values y_k of Y.

second: we have to estimate the parameters that define the prior probability over Y:

$\boxed{\pi_k \equiv P(Y=y_k)}$.

NB for continuous Inputs: we must estimate the mean and standard deviation of each of these Gaussians
for example Gaussian NB

$\boxed{\mu_{ik} \text{ and } \sigma_{ik}^2}$ for each feature X_i and each possible value y_k of Y

For LR: $W \leftarrow \arg\max_W \prod P(Y^l \mid X^l, W)$.

where $\boxed{W = \langle w_0, w_1, \dots w_n \rangle}$ is the vector of parameters to be estimated.

(d) methods:

For NB: We can use either MLE or MAP.

For LR: Choose parameters $W = \langle W_0, \ldots, W_n \rangle$ to maximize conditional likelihood of training data. called MCLE

Since there is no closed form solution to maximizing $l(W)$

we use ~~the~~ gradient ascent.

Details are in the following

2. a) NB is classed a generative classifier, because we can view the distribution $P(X|Y)$ as describing how to generate random instances X conditioned on the target attribute Y.

LR is referred to as a discriminative classifier because we can view the distribution $P(Y|X)$ as directly discriminating the value of the target value Y for any given instance X.

b)

μ: decide the position of the contour

$w_0 + \sum_i w_i X_i = 0$. Linear Decision Boundary

$W = \langle w_0 \ldots w_n \rangle$ is the parameters learned.

Σ: decide the shape of the contour.

(d) of 1

d. methods to do MLE.

For NB:

$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$

$= \arg\max_{\theta} \prod_{i=1}^{n} P(Y_i|\theta)$

Then take derivative and set it to 0, then we get $\hat{\theta}_{MLE}$.

$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta/D)$

$= \arg\max_{\theta} P(D/\theta) \cdot P(\theta)$

Then take derivative and set it to 0, then we get $\hat{\theta}_{MAP}$

For LR: we solved for LR parameters with MCLE

$l(w) = \log \prod_{j=1}^{n} P(Y^{(j)} = y \mid X^{(j)} = X; w)$

~~Since~~ Since there is no closed-form solution to maximizing $l(w)$.

We use gradient ascent.

(c) $p(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)}$

$= \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$

$= \frac{1}{1+\exp\left(\ln\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}\right)}$

Suppose: $P(Y=1)=\pi$, $P(Y=0)=1-\pi$

$= \frac{1}{1+\exp\left(\ln\frac{P(Y=0)}{P(Y=1)}+\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$

$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$

write $P(X_i=x|Y=1)=\theta_{i1}^{x}(1-\theta_{i1})^{1-x}$

$P(X_i=x|Y=0)=\theta_{i0}^{x}(1-\theta_{i0})^{1-x}$

where θ_{i1}: parameter of feature i ~~under class~~ when $Y=1$.

θ_{i0}: " $Y=0$

Now consider just

$\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} = \sum_i \ln\left(\frac{\theta_{i0}}{\theta_{i1}}\right)^{X_i}\cdot\left(\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)^{1-X_i}$

$= \sum_i\left[\left(\ln\frac{\theta_{i0}}{\theta_{i1}}-\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)X_i+\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right]$

Let $w_i = \ln\frac{\theta_{i0}}{\theta_{i1}}-\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}$

then $P(Y=1|X) = \frac{1}{1+\exp\left[\ln\frac{1-\pi}{\pi}+\sum_i\left(w_iX_i+\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right]}$

Let $w_0 = \ln\frac{1-\pi}{\pi}+\sum_i \ln\frac{1-\theta_{i0}}{1-\theta_{i1}}$

then $P(Y=1|X) = \frac{1}{1+\exp\left(w_0+\sum_{i=1}^{p} w_iX_i\right)}$

So we can write $P(Y=1|X)$ in a form that matches the Logistic class distribution

(d) From (c), we show that the value of the weights w_i of LR can be ~~ex~~ provied in terms of the parameters estimated by the NB classifier. They have the same form. And when we optimize the conditional likelihood, ~~they~~ we get the same classifier.

(e) The Naive Bayes assumption (that assumes the attributes $X_1 \ldots X_d$ are all conditionally independent of one another, given Y) makes NB less generic than LR.

(f) LR and NB are identical in the limit as the number of training examples approaches infinity, provided the Naive Bayes assumptions hold

3 (a) LR will generally have a lower asymptotic error rate

(b) $n = \Omega(P)$

(c) $n = O(\log P)$

(d) when the number of training data is less than $\Omega(P)$ but greater than $O(\log P)$

Problem 2

1\. (a) 1. Maximize the conditional log-likelihood ~~#~~ is easier.

2\. maximize the log of a function and the function itself is the same

(b) Because log function is a monotonically increasing function

$$2.(a)\ l(w) = \log\left[\prod_{j=1}^{n}\left(\sum_{k=1}^{C} I(y_j=k)\frac{e^{w_k^T x_j}}{\sum_{c'=1}^{C} e^{w_{c'}^T x_i}}\right)\right]$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{C}\left[I(y_j=k)\, w_k^T x_j - \log \sum_{c'=1}^{C} e^{w_{c'}^T x_j}\right]$$

$$(b).\ \nabla_{w_c} l(w) = \sum_{j=1}^{n}\left(\overbrace{\nabla_{w_c} w_c^T x_j}^{I(y_j=c)} - \nabla_{w_c} \ln \sum_{c'=1}^{C} e^{w_{c'}^T x_j}\right)$$

$$= \sum_{j=1}^{n}\left[I(y_j=c)\, x_j - \frac{e^{w_c^T x_j}\cdot x_j}{\sum_{c''}^{C} e^{w_{c''}^T x_j}}\right]$$

$$= \sum_{j=1}^{n} x_j\left[I(y_j=c) - P(y_j=c\,|\,x^j, w)\right]$$

(c).

When $c'=c$,

$$\frac{\partial^2 l(w)}{\partial w_c^2} = \left\{\sum_{j=1}^{n} x_j\left[I(y_j=c) - P(y_j=c\,|\,x_j, w)\right]\right\}'$$

$$= \sum_{j=1}^{n} x_j\left[I(y_j=c) - \frac{e^{w_c^T x_j}}{\sum_{c'=1}^{C} e^{w_{c'}^T x_j}}\right]'$$

$$= \sum_{j=1}^{n} x_j (x_j)^T\left[\left(\frac{e^{w_c^T\cdot x_j}}{\sum_{c'=1}^{C} e^{w_{c'}^T x_j}}\right)^2 - \left(\frac{e^{w_c^T\cdot x_j}}{\sum_{c'=1}^{C} e^{w_{c'}^T x_j}}\right)\right]$$

$$= \sum_{j=1}^{n} P(y_j=c\,|\,x_j, w)\left[P(y_j=c\,|\,x_j, w) - 1\right] x_j\cdot(w_j)^T$$

when $c' \neq c$

$$\frac{\partial^2 (w)}{\partial w_c \partial w_{c'}} = \left\{ \sum_{j=1}^{n} x_j \left[I(y_j = c) - P(y_j = c \mid x_j, w) \right] \right\}'$$

$$= \sum_{j=1}^{n} x_j \left[I(y_j = c) - \frac{e^{w_c^T x_j}}{\sum_{c'=1} e^{w_{c'}^T x_j}} \right]'$$

$$= \sum_{j=1}^{n} x_j (x_j)^T \cdot \frac{e^{w_c^T x_j} \cdot e^{w_{c'}^T x_j}}{\left(\sum_{c=1}^{C} e^{w_c^T x_j} \right)^2}$$

$$= \sum_{j=1}^{n} P(y_j = c \mid x_j, w) \cdot P(y_j = c' \mid x_j, w) \cdot x_j (x_j)^T$$