**1.** **a:** formula for the conditional likelihood

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

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

**b.** classification rule:

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

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

**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: $W \leftarrow \underset{W}{\arg\max} \prod P(Y^i \mid X^i, W)$.

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: Since there is no closed form solution to maximizing $l(w)$. choose parameters $w = \langle w_0, \ldots, w_n \rangle$ to maximize Conditional likelihood of training data. called MCLE

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 $Y$.

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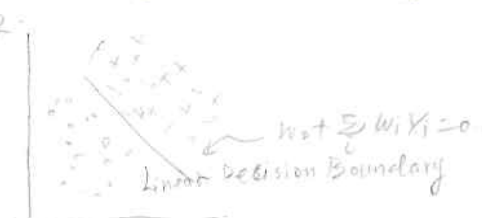as directly discriminating the value of the target value $Y$ for any given instance $X$.

b)



NB

LR

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

$w = \langle w_0 \ldots w_n \rangle$ is the Parameters learned.

II: decide the
position of the
contour

$\Sigma$: 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.
$l(w) = \log \prod_{j=1}^{n} P(Y^{(j)} = y | X^{(j)} = X ; w)$.
~~since~~ Since there is no closed-form Solution to maximizing $l(w)$.
we use gradient ascent.

(C) $P(Y=1|X) = \dfrac{P(Y=1)\,P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$

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

$= \dfrac{1}{1 + \exp\left(\ln \dfrac{P(Y=0)\,P(x|Y=0)}{P(Y=1)\,P(x|Y=1)}\right)}$

Suppose:
$\boxed{\begin{array}{l} P(Y=1)=\pi \\ P(Y=0)=1-\pi \end{array}}$

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

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

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 $\theta_{i1}$: Parameter of feature $i$ ~~under class~~ when $Y=1$.

$\theta_{i0}$: $\qquad\qquad Y=0$

Now consider just

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

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

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

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

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

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

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 ~~provied~~ 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.

②

(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 $O(\log p)$