THEORY

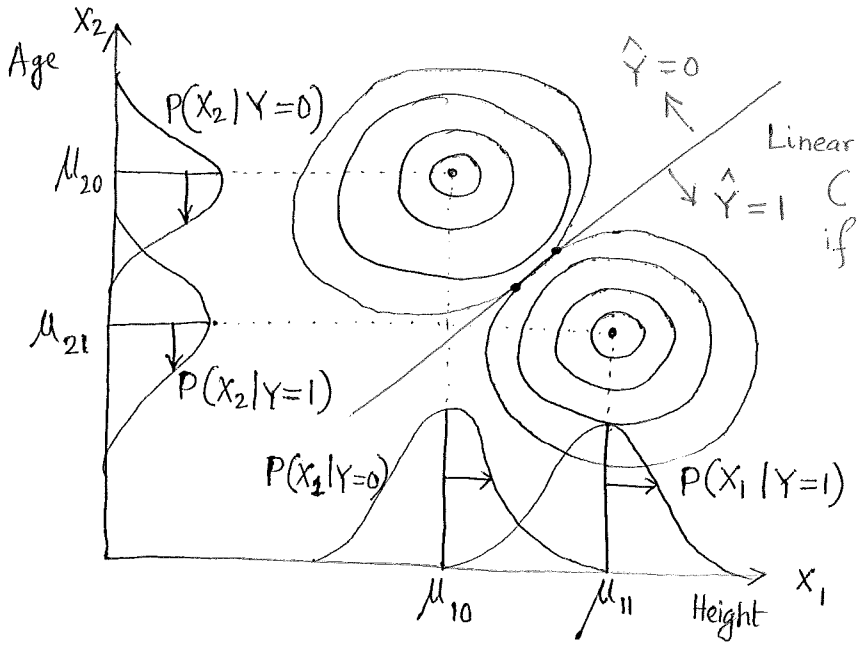
OF

LOGISTIC REGRESSION

This page intentionally left blank

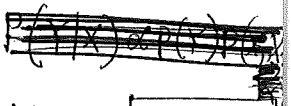
Gaussian Naive Bayes

- Consider boolean $Y \in \{0,1\}$ e.g. Y = Plays Basketball ?
- Consider continious X_i . e.g. X_1 = Height, X_2 = Age.
- Consider $P(Y=0) = P(Y=1) = 0.5$



Non-spherical Gaussians.

Linear decision surface (Not always!!) if $\sigma_{ik} = \sigma_{ij}$ $\forall k, j$



$P(Y|X) \propto P(Y) P(X_1, X_2|Y)$

$\uparrow$

$\prod P(X_i|Y=k)$

$\sim \mathcal{N}(\mu_{ik}, \sigma_i)$

$$\hat{Y} \leftarrow \underset{y \in \{0,1\}}{\operatorname{argmax}} \; P(Y=y) \prod_i P(X_i|Y=y)$$

Parametric Form of P(Y|X)

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

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$P(B) = \sum_j P(B|A_j)\,P(A_j) \quad \text{if } A \in \{0,1\}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(A)\,P(B|A) + P(A')\,P(B|A')}$$

Dividing both Nr and Dr by Nr gives :

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

$$\text{or,} \quad P(Y=1|X) = \frac{1}{1 + \exp\left(\ln \frac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}\right)} \qquad \left| \; X = \exp(\ln X) \right.$$

$$= \frac{1}{1 + \exp\left(\sum_i \left(\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}\right) + \ln \frac{P(Y=0)}{P(Y=1)}\right)} \qquad \left| \; \ln ab = \ln a + \ln b \right.$$

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

C. Indep.

$$\sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)} = \sum_i \ln \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)}$$

$$P(X|y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

$$P(X_i|Y=y_k) \sim \mathcal{N}(\mu_{ik}, \sigma_i)$$

$$\frac{e^a}{e^b} = e^{a-b}$$

$$= \sum_i \ln \exp\left(\frac{(X_i-\mu_{i1})^2 - (X_i-\mu_{i0})^2}{2\sigma_i^2}\right)$$

$$= \sum_i \left(\frac{(X_i-\mu_{i1})^2 - (X_i-\mu_{i0})^2}{2\sigma_i^2}\right) = \sum_i \left(\frac{(X_i^2 - 2X_i\mu_{i1} + \mu_{i1}^2) - (X_i^2 - 2X_i\mu_{i0} + \mu_{i0}^2)}{2\sigma_i^2}\right)$$

$$= \sum_i \left(\frac{2X_i(\mu_{i0}-\mu_{i1}) + \mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}\right) = \boxed{\sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}\right)}$$

Therefore, $$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln \frac{1-\pi}{\pi} + \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}\right)\right)}$$

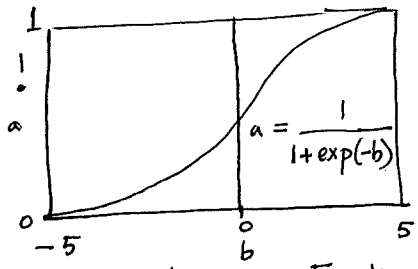
$$= \boxed{\frac{1}{1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}}$$

$$\begin{cases} w_0 = \ln \frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} \\ w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} \quad \text{for } i = 1 \dots n \end{cases}$$

Linear Classifier - Very convenient!

$$P(Y=1|X) = \frac{1}{1+\exp\left(w_0 + \sum_i w_i X_i\right)}$$

$$P(Y=0|X) = 1 - P(Y=1|X) = \frac{\exp\left(w_0 + \sum_i w_i X_i\right)}{1 + \exp\left(w_0 + \sum_i w_i X_i\right)}$$



Logistic Function

$$\Rightarrow \frac{P(Y=0|X)}{P(Y=1|X)} = \exp\left(w_0 + \sum_i w_i X_i\right) \gtrless 1$$

$$\Rightarrow \ln \frac{P(Y=0|X)}{P(Y=1|X)} = w_0 + \sum_i w_i X_i \gtrless 0 \quad \text{(linear classification rule!)}$$

Conditional Log Likelihood

$$l(W) = \sum_l \left(Y^l \ln P(Y^l=1|X^l,W)\right) + \left((1-Y^l) \ln P(Y^l=0|X^l,W)\right)$$

$$= \sum_l Y^l \ln \frac{P(Y^l=1|X^l,W)}{P(Y^l=0|X^l,W)} + \ln P(Y^l=0|X^l,W)$$

$$= \sum_l Y^l \left(w_0 + \sum_{i=1}^{n} w_i X_i^l\right) - \ln\left(1 + \exp\left(w_0 + \sum_{i=1}^{n} w_i X_i^l\right)\right)$$

$$l(W) = \ln \prod_l P(Y^l|X^l,W) = \sum_l \ln P(Y^l|X^l,W)$$

Assume:

$$P(Y=0|X) = \frac{1}{1+\exp\left(w_0 + \sum_i w_i X_i\right)}$$

$$P(Y=1|X) = \frac{\exp\left(w_0 + \sum_i w_i X_i\right)}{1+\exp\left(w_0 + \sum_i w_i X_i\right)}$$

Logistic Regression More Generally

- When Y is not boolean, but still discrete-valued
- Now $y \in \{y_1, \ldots y_R\}$: learn $R-1$ sets of weights

for $k < R$

$$P(Y=y_k|X) = \frac{\exp(\omega_{k\phi} + \sum_{i=1}^{n} \omega_{ki} X_i)}{1 + \sum_{j=1}^{R-1} \exp(\omega_{j\phi} + \sum_{i=1}^{n} \omega_{ji} X_i)}$$

for $k = R$

$$P(Y=y_R|X) = \frac{1}{1 + \sum_{j=1}^{R-1} \exp(\omega_{j\phi} + \sum_{i=1}^{n} \omega_{ji} X_i)}$$

Training Logistic Regression: MCLE

- we have L training examples $\{\langle X^1, Y^1\rangle, \ldots \langle X^L, Y^L\rangle\}$
- maximum likelihood estimate for parameters W

$$W_{MLE} = \arg\max_W P(\langle X^1, Y^1\rangle \ldots \langle X^L, Y^L\rangle | W)$$

$$= \arg\max_W \prod_l P(\langle X^l, Y^l | W)$$

- maximum conditional likelihood estimate

$$W_{MCLE} = \arg\max_W \prod_l P(Y^l | X^l, W)$$

Steps for training logistic regression:

- Choose parameters $W = \langle \omega_0, \omega_1, \ldots \omega_n \rangle$ to maximize cond. likelihood of traing data where $P(Y=0|X) = \frac{1}{1+\exp(\omega_0 + \sum_i \omega_i X_i)}$; $P(Y=1|X) = \frac{\exp(\omega_0 + \sum_i \omega_i X_i)}{1+\exp(\omega_0 + \sum_i \omega_i X_i)}$
- Training data $D = \{\langle X^1, Y^1\rangle \ldots \ldots \langle X^L, Y^L\rangle\}$
- Data conditional likelihood $= \prod_l P(Y^l | X^l, W)$

$$W_{MCLE} = \arg\max_W \prod_l P(Y^l | W, X^l)$$

Maximizing Conditional Log likelihood

$$P(Y=\phi \mid X, W) = \frac{1}{1+\exp(w_0 + \sum_i w_i X_i)}$$

$$P(Y=1 \mid X, W) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1+\exp(w_0 + \sum_i w_i X_i)}$$

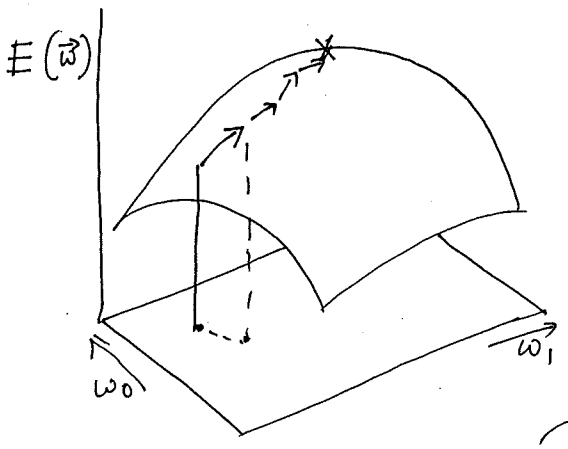
$$l(W) = \ln \prod_l P(Y^l \mid X^l, W)$$

$$= \sum_l Y^l (w_0 + \sum_{i=1}^{n} w_i X_i^l) - \ln\left(1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i^l)\right)$$

Good News: $l(W)$ is a concave function of W

Bad News: no closed-form solution to maximize $l(W)$

Gradient Ascent:



Gradient $\nabla E(\vec{w}) = \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots \frac{\partial E}{\partial w_n} \right]$

Training Rule: $\vec{w}^{(i+1)} \leftarrow \vec{w}^{(i)} + \eta \nabla E(\vec{w})$

↑ step size

i.e. $\Delta w_i = \eta \frac{\partial E}{\partial w_i}$

Batch gradient : use value of the objective function $E_D(\vec{w})$ over the entire training data D

Do Repeat

1) Calcalate $\nabla E_D(\vec{w}) = \left[\frac{\partial E_D(\vec{w})}{\partial w_0}, \dots \frac{\partial E_D(\vec{w})}{\partial w_n} \right]$

2) Update parameters $\vec{w} \leftarrow \vec{w} + \eta \nabla E_D(\vec{w})$

Stochastic gradient use value of the objective function $E_d(\vec{w})$ ove single example $d \in D$

Do Repeat

1) Choose (with replacement) a random training example $d \in D$

2) Calculate the gradient just for d

$$\nabla E_d(\vec{w}) = \left[\frac{\partial E_d(\vec{w})}{\partial w_0}, \dots \frac{\partial E_d(\vec{w})}{\partial w_n} \right]$$

3) Update parameters $\vec{w} \leftarrow \vec{w} + \eta \nabla E_d(\vec{w})$

N.B:

- Stochastic approximates batch arbitrarily closely as $\eta \to 0$
- Stochastic is much faster when D is very large
- An intermediate approach can be to choose d as a subset of D

Maximize Conditional Log Likelihood : Gradient Ascent

$$l(W) = \ln \prod_l P(Y^l \mid X^l, W)$$

$$= \sum_l Y^l \left(w_0 + \sum_i w_i X_i^l \right) - \ln \left(1 + \exp\left(w_0 + \sum_i w_i X_i^l \right) \right)$$

$$\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l \left(Y^l - \hat{P}(Y^l = 1 \mid X^l, W) \right)$$

weighted by feature value

misassigned prob. mass given current sets of weights

Gradient ascent algorithm : iterate until change $< \epsilon$

$\forall\, i$, repeat

$$w_i \leftarrow w_i + \eta \sum_l x_i^l \left(Y^l - \hat{P}(Y^l = 1 \mid x^l, W) \right)$$

assume $X_\phi = 1$ for w_0

We used M(C)LE. Can we use MAP?

- One common approach is to define priors on W
- Helps avoid very large weights and overfitting
- MAP estimate

$$W \leftarrow \arg\max_W \ln P(W) \prod_l P(Y^l \mid X^l, W)$$

- Let's assume Gaussian prior $W \sim \mathcal{N}(\phi, \sigma)$

Gradient ascent algorithm: iterate until change $< \epsilon$

$$w_i \leftarrow w_i - \boxed{\eta \lambda w_i} + \eta \sum_l x_i^l \left(Y^l - \hat{P}(Y^l = 1 \mid X^l, W) \right)$$

regularization term

(this is pushing the weights to be zero)