CS 4824/ECE 4424: Logistic Regression

Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.

Logistic Regression

Idea:

- Naïve Bayes allows estimating $P(Y|X)$ by learning $P(Y)$ and $P(X|Y)$
- Why not learn $P(Y|X)$ directly?

Problem setting

X_1 = Temp, Y = Play T
$\mathbb{R}$ Y N

- Consider learning f: $X \rightarrow Y$
  - X is a vector of real-valued features $<X_1, X_2, \ldots, X_n>$ $X_i \in \mathbb{R}$
  - Y is boolean $Y \in \{0, 1\}$
  - Assume all X_i's are conditionally independent given Y C.I.
  - Model $P(X_i \mid Y = y_k)$ as Gaussian ~ $\mathcal{N}(\mu_{ik}, \sigma_i)$ not σ_{ik}
  - Model P(Y) as Bernoulli (π)

feature X_i class label y_k

- Given that, what's the parametric form of $\boxed{P(Y \mid X)}$? $\propto P(Y)\,P(X \mid Y)$

Parametric form of P(Y|X)

- $$\boxed{P(Y=1\mid X)} = \frac{P(Y=1)P(X\mid Y=1)}{P(Y=1)P(X\mid Y=1)+P(Y=0)P(X\mid Y=0)}$$

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

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

$\exp(\ln x) = x$

$\prod_i P(X_i / Y=0)$

$\boxed{\hat{P}(Y=1)=\pi}$

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

C.I.

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

Parametric form of P(Y | X)

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

$$\boxed{P(x \mid y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2\right)}}$$

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

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

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

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

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

Parametric form of P(Y|X)

- Therefore, $P(Y = 1 | X) = \dfrac{1}{1 + exp(w_0 + \sum_{i=1}^{n} w_i x_i)}$

  - where, $w_0 = \ln \dfrac{1 - \pi}{\pi} + \boxed{\sum_i \dfrac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}}$; and $w_i = \dfrac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}$ for $i = 1 \ldots n$

Very convenient!

- $\boxed{P(Y=1 \mid X=<X_1,\ldots,X_n>)} = \boxed{\dfrac{1}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}}$
- implies
  - $P(Y=0 \mid X=<X_1,\ldots,X_n>) = \dfrac{exp(w_0+\sum_{i=1}^{n} x_i w_i)}{1+exp(w_0+\sum_{i=1}^{n} x_i w_i)}$
- implies
  - $\dfrac{P(Y=0 \mid X)}{P(Y=1 \mid X)} = exp(w_0+\sum_{i=1}^{n} x_i w_i) \gtrless 1$

linear decision rule.

- or equivalently
  - $\ln \dfrac{P(Y=0 \mid X)}{P(Y=1 \mid X)} = \boxed{w_0+\sum_{i=1}^{n} x_i w_i} \gtrless 0$

Very convenient!

- $P(Y=1 \mid X=<X_1,\ldots,X_n>) = \boxed{\dfrac{1}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}}$
- implies
  - $P(Y=0 \mid X=<X_1,\ldots,X_n>) = \dfrac{exp(w_0+\sum_{i=1}^{n} w_i x_i)}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}$
- implies
  - $$\frac{P(Y=0 \mid X)}{P(Y=1 \mid X)} = exp(w_0 + \sum_{i=1}^{n} w_i x_i)$$

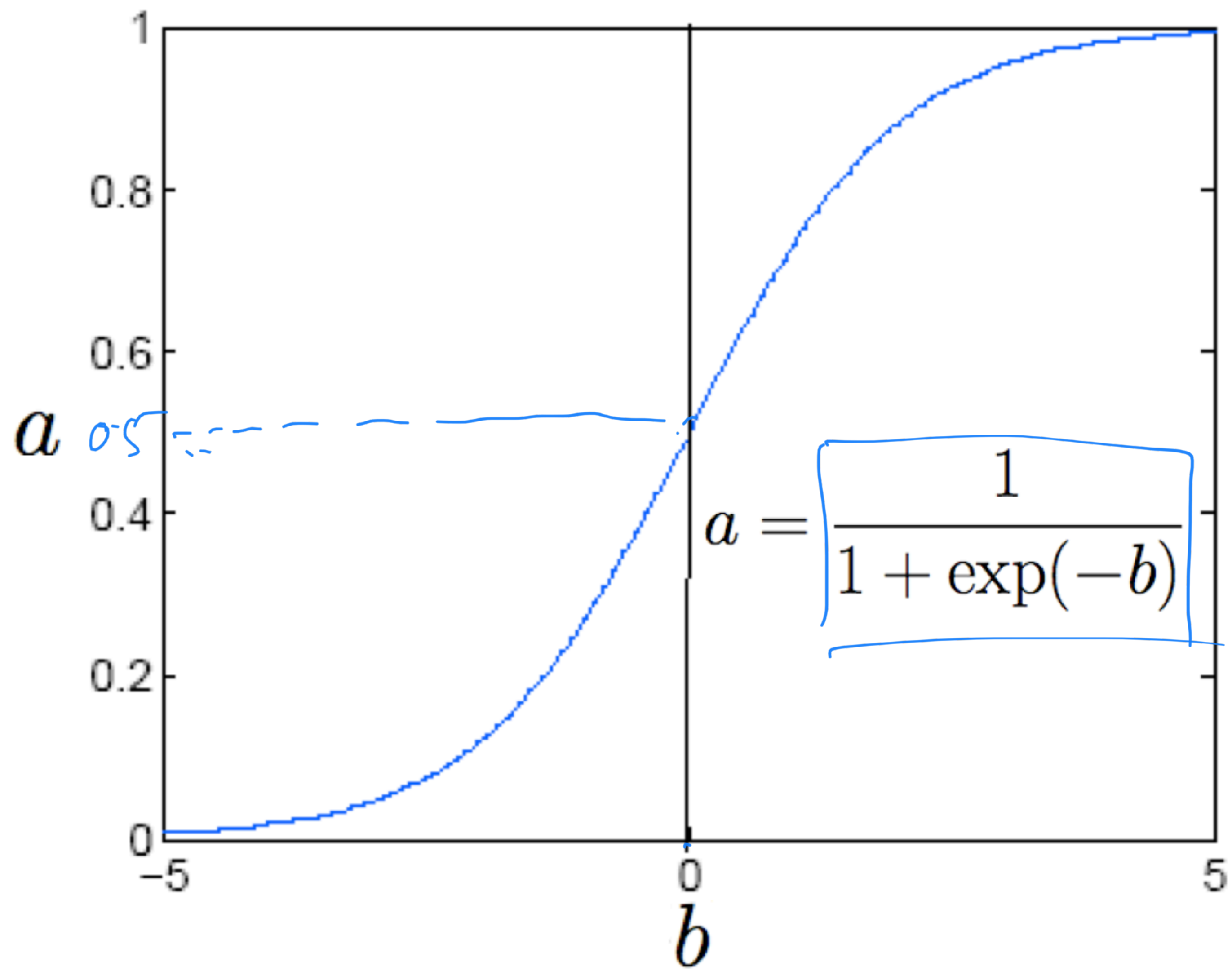
linear classification rule!

- or equivalently
  - $$\ln \frac{P(Y=0 \mid X)}{P(Y=1 \mid X)} = \boxed{w_0 + \sum_{i=1}^{n} w_i x_i}$$

dot product of weights and the features

$$a \, . \, b = \sum_{i=1}^{n} a_i b_i$$

Logistic function



$$P(Y=1 \mid X) = \frac{1}{1 + exp(w_0 + \sum_{i=1}^{n} w_i x_i)}$$

Logistic regression more generally

- Logistic regression when Y not boolean, but still discrete valued
- Now $Y \in \{y_1, \ldots y_R\}$ and so we need to learn *R-1* sets of weights

  - for $k < R$: $$P(Y = y_k | X) = \frac{exp(w_{k0} + \sum_{i=1}^{n} w_{ki}x_i)}{1 + \sum_{j=1}^{R-1} exp(w_{j0} + \sum_{i=1}^{n} w_{ji}x_i)}$$

  - for $k = R$: $$P(Y = y_R | X) = \frac{1}{1 + \sum_{j=1}^{R-1} exp(w_{j0} + \sum_{i=1}^{n} w_{ji}x_i)}$$

Training logistic regression: MCLE

- We have L training examples {<X¹, Y¹>,…, <Xᴸ, Yᴸ>}
- Maximum likelihood estimate (MLE) for parameters W

$w_0, w_i \ \forall i$

  - $$W_{MLE} = \arg\max_W P(<X^1, Y^1> \ldots <X^L, Y^L> \mid W)$$
  - $$= \arg\max_W \prod_l P(<X^l, Y^l> \mid W)$$
- Maximum conditional likelihood estimate (MCLE)

Data conditional likelihood $\prod_l P(Y^l \mid X^l, W)$

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

Training logistic regression: MCLE

- We have L training examples {<X^1, Y^1>,…, <X^L, Y^L>}
- Maximum likelihood estimate (MLE) for parameters W
  - $$W_{MLE} = \arg\max_W P(<X^1, Y^1> \ldots <X^L, Y^L> \mid W)$$
  - $$= \arg\max_W \prod_l P(<X^l, Y^l> \mid W)$$
- Maximum <u>conditional</u> likelihood estimate (MCLE)

Training logistic regression: MCLE

- We need to choose W = <w_0,…,w_n> to maximize the conditional likelihood of training data
    - where $P(Y=0\,|\,X,W) = \dfrac{1}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}$
    - and $P(Y=1\,|\,X,W) = \dfrac{exp(w_0+\sum_{i=1}^{n} w_i x_i)}{1+exp(w_0+\sum_{i=1}^{n} w_i x_i)}$
- Training data D = {<X^1, Y^1>,…, <X^L, Y^L>}
- Data likelihood is $\prod_l P(<X^l, Y^l> \,|\, W)$
- Data conditional likelihood is $\prod_l P(Y^l \,|\, X^l, W)$
- Therefore we need to estimate $\boxed{W_{MCLE} = \arg\max_W \prod_l P(Y^l \,|\, X^l, W)}$

Expressing conditional log likelihood

- $l(W) = \ln \prod_l P(Y^l | X^l, W) = \sum_l \ln P(Y^l | X^l, W)$
  - where $P(Y = 0 | X, W) = \dfrac{1}{1 + exp(w_0 + \sum_{i=1}^n w_i x_i)}$
  - and $P(Y = 1 | X, W) = \dfrac{exp(w_0 + \sum_{i=1}^n w_i x_i)}{1 + exp(w_0 + \sum_{i=1}^n w_i x_i)}$

- $$l(W) = \sum_l \boxed{Y^l \ln P(Y^l = 1 | X^l, W)} + \boxed{(1 - Y^l) \ln P(Y^l = 0 | X^l, W)}$$
  $$= \sum_l Y^l \ln \frac{P(Y^l = 1 | X^l, W)}{\boxed{P(Y^l = 0 | X^l, W)}} + \ln P(Y^l = 0 | X^l, W)$$
  $$= \sum_l Y^l (w_0 + \sum_i^n w_i X_i^l) - \ln(1 + exp(w_0 + \sum_i^n w_i X_i^l))$$

Maximizing conditional log likelihood

- $$l(W) = \ln \prod_{l} P(Y^l \mid X^l, W)$$

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

- **Good news**: *l(W)* is a concave function of *W*
- **Bad news**: no closed-form solution to maximize *l(W)*

What do we do? Optimization