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

- Consider learning $f: X \rightarrow Y$
- $X$ is a vector of real-valued features $<X_1, X_2, \ldots, X_n>$
- $Y$ is boolean $Y \in \{0, 1\}$
- Assume all $X_i$'s are conditionally independent given $Y$
- Model $P(X_i | Y = y_k)$ as Gaussian $\sim \mathcal{N} \left( \mu_{ik}, \sigma_i \right)$
- Model $P(Y)$ as Bernoulli ($\pi$)

- Given that, what’s the parametric form of $P(Y | X)$?
Parametric form of $P(Y|X)$

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

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

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

$\exp(\ln x) = x$
Parametric form of $P(Y \mid X)$

\[
\sum_i \ln \frac{P(X_i \mid Y = 0)}{P(X_i \mid Y = 1)} = \sum_i \ln \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left( -\frac{(x_i - \mu_{i0})^2}{2\sigma_i^2} \right) \cdot \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left( -\frac{(x_i - \mu_{i1})^2}{2\sigma_i^2} \right) \cdot P(x \mid y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu_{ik}}{\sigma_{ik}} \right)^2}
\]

\[
e^a = e^{a-b} \quad \frac{e^a}{e^b} = e^{a-b}
\]

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

\[
= \sum_i \frac{(x_i - \mu_{i1})^2 - (x_i - \mu_{i0})^2}{2\sigma_i^2}
\]

\[
= \sum_i \left( x_i^2 - 2x_i \mu_{i1} + \mu_{i1}^2 - x_i^2 + 2x_i \mu_{i0} - \mu_{i0}^2 \right)
\]

\[
= \sum_i \left( 2x_i (\mu_{i0} - \mu_{i1}) + \mu_{i1}^2 - \mu_{i0}^2 \right)
\]

\[
= \sum_i \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} x_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}
\]
Therefore, \( P(Y = 1 \mid X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)} \)

where, \( w_0 = \ln \frac{1 - \pi}{\pi} + \sum_{i} \frac{\mu_{i1} - \mu_{i0}}{2\sigma_i^2} \); and \( w_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} \) for \( i = 1 \ldots n \)

Parametric form of \( P(Y \mid X) \)
Very convenient!

\[
P(Y = 1 \mid X = < X_1, \ldots, X_n > ) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)}
\]

° implies

° \( P(Y = 0 \mid X = < X_1, \ldots, X_n > ) = \)

° implies

\[
P(Y = 0 \mid X) = \frac{\exp(w_0 + \sum_{i=1}^{n} x_i w_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} x_i w_i)}
\]

° or equivalently

\[
\ln \frac{P(Y = 0 \mid X)}{P(Y = 1 \mid X)} = \left[ w_0 + \sum_{i=1}^{n} x_i w_i \right] \geq 0
\]
Very convenient!

\[
P(Y = 1 \mid X = < X_1, \ldots, X_n > ) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)}
\]

○ implies

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

○ implies

\[
\frac{P(Y = 0 \mid X)}{P(Y = 1 \mid X)} = \exp(w_0 + \sum_{i=1}^{n} w_i x_i)
\]

○ or equivalently

\[
\ln \frac{P(Y = 0 \mid X)}{P(Y = 1 \mid X)} = w_0 + \sum_{i=1}^{n} w_i x_i
\]

linear classification rule!

dot product of weights and the features

\[
a \cdot b = \sum_{i=1}^{n} a_i b_i
\]
Logistic function

\[
P(Y = 1 \mid X) = \frac{1}{1 + \exp(-b)}
\]

\[
a = \frac{1}{1 + \exp(-b)}
\]
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

  \begin{align*}
  \text{for } k < R: \quad \Pr(Y = y_k \mid 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)} \\
  \text{for } k = R: \quad \Pr(Y = y_R \mid X) &= \frac{1}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji}x_i)}
  \end{align*}
Training logistic regression: MCLE

- We have $L$ training examples $\{<X_1, Y_1>, \ldots, <X_L, Y_L>\}$
- Maximum likelihood estimate (MLE) for parameters $W$
  
  $W_{MLE} = \arg \max_W P( <X_1, Y_1> \ldots <X_L, Y_L> | W) = \arg \max_W \prod_l P(<X^l, Y^l> | W)$

- Maximum conditional likelihood estimate (MCLE)

\[
W_{MCLE} = \arg \max_W \prod_l P(Y^l | X^l, W)
\]
Training logistic regression: MCLE

- We have L training examples \{<X^1, Y^1>,..., <X^L, Y^L>\}
- Maximum likelihood estimate (MLE) for parameters W
  \[ W_{MLE} = \arg \max_W P( <X^1, Y^1> \ldots <X^L, Y^L> | W) \]
  \[ = \arg \max_W \prod_l P( <X^l, Y^l> | W) \]
- Maximum conditional likelihood estimate (MCLE)
Training logistic regression: MCLE

- We need to choose $W = <w_0,\ldots,w_n>$ to maximize the conditional likelihood of training data

  where $P(Y = 0 | X, W) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_ix_i)}$

  and $P(Y = 1 | X, W) = \frac{\exp(w_0 + \sum_{i=1}^n w_ix_i)}{1 + \exp(w_0 + \sum_{i=1}^n w_ix_i)}$

- Training data $D = \{<X^1, Y^1>, \ldots, <X^L, Y^L>\}$

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

\[ P(Y = 0 | X, W) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i x_i)} \]

and

\[ P(Y = 1 | X, W) = \frac{\exp(w_0 + \sum_{i=1}^n w_i x_i)}{1 + \exp(w_0 + \sum_{i=1}^n w_i x_i)} \]

\[ l(W) = \sum_l Y^l \ln P(Y^l = 1 | X^l, W) + (1 - Y^l) \ln P(Y^l = 0 | X^l, W) \]

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

\[ = \sum_l Y^l (w_0 + \sum_{i=1}^n w_i x_i^l) - \ln (1 + \exp(w_0 + \sum_{i=1}^n w_i x_i^l)) \]
Maximizing conditional log likelihood

\[ l(W) = \ln \prod_l P(Y^l | X^l, W) \]

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

- **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