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.
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
  - Assume all $X_i$'s are conditionally independent given $Y$
  - Model $P(X_i | Y = y_k)$ as Gaussian $\sim N(\mu_{ik}, \sigma_i)$
  - Model $P(Y)$ as Bernoulli ($\pi$)

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

\[ P(Y = 1 \mid X) = \frac{P(Y = 1)P(X \mid Y = 1)}{P(Y = 1)P(X \mid Y = 1) + P(Y = 0)P(X \mid Y = 0)} \]
Parametric form of $P(Y \mid X)$

\[
\sum_i \ln \frac{P(X_i \mid Y = 0)}{P(X_i \mid Y = 1)}
\]

\[
P(x \mid y_k) = \frac{1}{\sigma_k \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x - \mu_{ik}}{\sigma_{ik}} \right)^2}
\]
Parametric form of $P(Y | X)$

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

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

\[ \frac{P(Y = 0 \mid X)}{P(Y = 1 \mid X)} = \]

° or equivalently

\[ \ln \frac{P(Y = 0 \mid X)}{P(Y = 1 \mid X)} = \]
Very convenient!

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

- implies

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

linear classification rule!

- or equivalently

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

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

- 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 \{<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 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 \( P(Y = 0 \mid X, W) = \frac{1}{1 + exp(w_0 + \sum_{i=1}^{n} w_i x_i)} \)

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

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

- Data likelihood is \( \prod_{l} P( <X^l, Y^l> \mid W) \)

- Data conditional likelihood is \( \prod_{l} P(Y^l \mid X^l, W) \)

- Therefore we need to estimate \( W_{MCLE} = \arg \max_{W} \prod_{l} P(Y^l \mid X^l, W) \)