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 $P(X_i \mid Y = y_k)$ as Gaussian $\sim \mathcal{N}(\mu_{ik}, \sigma_i)$
- Model $P(Y)$ as Bernoulli ($\pi$)
- Given that, what’s the parametric form of $P(Y \mid X)$?
Parametric form of $P(Y \mid X)$

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

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

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

$$= \frac{1}{1 + \exp\left(\ln\frac{1 - \Pr(x)}{\Pr(x)} + \sum_{i=1}^{n} \ln\frac{P(x_{i} \mid y = 0)}{P(x_{i} \mid y = 1)}\right)}$$

$\exp(\ln x) = x$

$\Pr(x = 1) = \pi$

cond. ind.
Parametric form of $P(Y|X)$

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

$$\sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

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

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

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

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

$$= \sum_i \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} x_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$$
Parametric form of $P(Y \mid X)$

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

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

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

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

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

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

Very convenient!
Very convenient!

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

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

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

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

linear classification rule!

or equivalently

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

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

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

\[ \sum_{i=1}^{n} w_i x_i \]
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

\[
\begin{align*}
\text{for } k < R: & \quad P(Y = y_k | X) = \frac{\exp(w_{k0} + \sum_{i=1}^{n} w_{ki}x_i)}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji}x_i)} \\
\text{for } k = R: & \quad P(Y = y_R | X) = \frac{1}{1 + \sum_{j=1}^{R-1} \exp(w_{j0} + \sum_{i=1}^{n} w_{ji}x_i)}
\end{align*}
\]
Training logistic regression: MCLE

- We have L training examples \{<X_1, Y_1>,..., <X_L, Y_L>\}
- Maximum likelihood estimate (MLE) for parameters \(W\)
  \[
  W_{MLE} = \arg \max_W P( <X_1, Y_1>, ..., <X_L, Y_L> | W)
  \]
  
  \[
  = \arg \max_W \prod_l P( <X_l, Y_l> | W)
  \]
- Maximum conditional likelihood estimate (MCLE)
  \[
  W_{MLE} = \arg \max_W \prod_l P(Y_l | X_l, W)
  \]
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)
Training logistic regression: MCLE

- We need to choose \( W = \langle w_0, \ldots, w_n \rangle \) 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 \( D = \{ \langle X^1, Y^1 \rangle, \ldots, \langle X^L, Y^L \rangle \} \)

- Data likelihood is \( \prod_l P( \langle X^l, Y^l \rangle \mid W) \)

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

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