Discussion on Logistic Regression and Naïve Bayes

Jingrui He
09/27/2007
Review of Logistic Regression

- Discriminative classifier
- Function form for $P(Y|X)$
  - $P(Y = 1|X, w) = \frac{\exp(w_0 + \sum_i w_i X_i)}{1 + \exp(w_0 + \sum_i w_i X_i)}$
- Can NOT obtain a sample of the data, because $P(X)$ is not available
Parameter Estimation

- Gradient ascent
  \[ w_{0}^{t+1} \leftarrow w_{0}^{t} + \eta \sum_{j} \left[ Y^{j} - \hat{P}(Y^{j} = 1|X^{j}, w^{t}) \right] \]
  \[ w_{i}^{t+1} \leftarrow w_{i}^{t} + \eta \sum_{j} X_{i}^{j} \left[ Y^{j} - \hat{P}(Y^{j} = 1|X^{j}, w^{t}) \right] \]

- Upon convergence
  \[ \frac{\partial l(w)}{\partial w_{0}} = \sum_{j} \left[ Y^{j} - P(Y^{j} = 1|X^{j}, w) \right] = 0 \]
  \[ \frac{\partial l(w)}{\partial w_{i}} = \sum_{j} X_{i}^{j} \left[ Y^{j} - P(Y^{j} = 1|X^{j}, w) \right] = 0 \]
Linear Separable

- What’s the value of \( w \)?
  - \textit{INFINITY!}

- Why?
  - Maximum likelihood

\[
\begin{align*}
P(Y = 1 | X, w) &= \frac{\exp\left(w_0 + \sum_i w_i X_i\right)}{1 + \exp\left(w_0 + \sum_i w_i X_i\right)}
\end{align*}
\]
More Training Examples

- No change in $w$
- Why?

$$w_{t+1}^0 \leftarrow w_0^t + \eta \sum_j \left[ Y_j - \hat{P}(Y_j = 1|X^j, w^t) \right]$$

$$w_{t+1}^i \leftarrow w_i^t + \eta \sum_j X_i^j \left[ Y_j - \hat{P}(Y_j = 1|X^j, w^t) \right]$$
Non-Linear Separable
More Training Examples
Still More Training Examples
Why?

- Originally, upon convergence
  \[
  \frac{\partial l(w)}{\partial w_0} = \sum_j \left[ Y^j - P(Y^j = 1 | X^j, w) \right] = 0
  \]

- With 3 more points
  \[
  \frac{\partial l(w)}{\partial w_0} > 0
  \]

- To let the derivative be 0 again
  - Increase \( P(Y^j = 1 | X^j, w) \)
Multiple Classes

- $R-1$ sets of weights
  - $P(Y = j|X, w_j) \propto \exp\left(w_{j0} + \sum_i w_{ji}X_i\right)$, $j = 1, \ldots, R - 1$
  - $P(Y = R|X, w_j) = \frac{1}{1 + \sum_{j=1}^{R-1} \exp\left(w_{j0} + \sum_i w_{ji}X_i\right)}$

- Classification
  - Comparing $\exp\left(w_{j0} + \sum_i w_{ji}X_i\right)$ and 1
  - Comparing $w_{j0} + \sum_i w_{ji}X_i$ and 0
4 Classes in 2d Space
LR vs. NB

- Loss functions
  - LR: maximum conditional data likelihood
    \[ \sum_j \ln \left( P(Y^j | X^j, w) \right) \]
  - NB: maximum data likelihood
    \[ \sum_j \ln \left( P(X^j, Y^j | w) \right) \]
- Different solutions!
LR vs. NB

- In NB, assume class independent variance

\[
P(Y = 1 | X, w) = \frac{1}{1 + \exp\left(w_0 + \sum_i w_i x_i \right)}
\]

\[
\ln \frac{1-\theta}{\theta} + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}
\]

\[
\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}
\]
LR vs. NB