Linear Classification (Part III: Logistic Regression)

nanopoulos@ismll.de

Outline

- Probabilistic discriminative models
- Logistic regression
- Maximum likelihood solution
- Iterative reweighted least squares
Discriminant vs. Probabilistic Discriminative

Use discriminant functions directly without probabilities:
Fisher’s LDA, Perceptron

Infer conditional class probabilities:
Compute the conditional probability of each class.
Determine the parameters directly using Maximum Likelihood

\[ p(\text{class} = C_k \mid \mathbf{x}) \]

Fixed basis functions

Assume fixed nonlinear transformation
Transform inputs using a vector of basis functions \( \phi(x) \)
The resulting decision boundaries will be linear in the feature space \( \phi \)
Outline

- Probabilistic discriminative models
- Logistic regression
- Maximum likelihood solution
- Iterative reweighted least squares

Logistic regression

Logistic regression model
Posterior probability of a class for two-class problem:

\[ P(C_1|\phi) = y(\phi) = \sigma(w^T\phi) \]

\[ P(C_2|\phi) = 1 - P(C_1|\phi) \]

\[ \sigma(a) = \frac{1}{1 + \exp(-a)} \]
Logistic function

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

\[
\frac{d\sigma}{da} = \frac{e^{-a}}{(1 + e^{-a})^2}
\]

\[
= \sigma(a) \left\{ \frac{e^{-a}}{1 + e^{-a}} \right\}
\]

\[
= \sigma(a) \left\{ \frac{1 + e^{-a}}{1 + e^{-a}} - \frac{1}{1 + e^{-a}} \right\}
\]

\[
= \sigma(a)(1 - \sigma(a)).
\]

Outline

- Probabilistic discriminative models
- Logistic regression
- Maximum likelihood solution
- Iterative reweighted least squares
Maximum Likelihood

- Determining \( \mathbf{w} \) using ML
  - Likelihood function:
    \[
    P(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{(1-t_n)}
    \]
  - \( \mathbf{t} = (t_1, \ldots, t_N)^T \)  \( y_n = p(C_1|\phi_n) \)
    \( C_1: t_n = 1 \)
    \( C_2: t_n = 0 \)
  - Cross-entropy error function (negative log likelihood)
    \[
    E(\mathbf{w}) = -\ln P(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{ t_n \ln y_n + (1-t_n) \ln (1-y_n) \}
    \]
    \( y_n = \sigma(a_n), a_n = \mathbf{w}^T \phi_n \)
  - The gradient of the error function w.r.t. \( \mathbf{w} \)
    \[
    \nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \phi_n
    \]

Computing the gradient

\[
E(\mathbf{w}) = -\ln P(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{ t_n \ln y_n + (1-t_n) \ln (1-y_n) \}
\]

\[
y_n = \sigma(a_n), a_n = \mathbf{w}^T \phi_n
\]

\[
\frac{\partial E}{\partial y_n} = \frac{1-t_n}{1-y_n} \cdot \frac{t_n}{y_n} = \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)}
\]

\[
\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n).
\]
Computing the gradient

\[ \nabla E = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla a_n \]

\[ = \sum_{n=1}^{N} (y_n - t_n) \phi_n \]

\[ \nabla E(w) = \sum_{n=1}^{N} (y_n - t_n) \phi_n \] No longer a closed-form solution

Outline

- Probabilistic discriminative models
- Logistic regression
- Maximum likelihood solution
- Iterative reweighted least squares
Iterative reweighted least squares

The Newton-Raphson update to minimize a function $E(w)$

$$w^{(new)} = w^{(old)} - H^{-1} \nabla E(w)$$

Where $H$ is the Hessian matrix, the second derivatives of $E(w)$

\[
\begin{align*}
\nabla E(w) &= \sum_{n=1}^{N} (y_n - t_n) \phi_n = \phi^T(y - t) \\
H &= \sum_{n=1}^{N} y_n (1 - y_n) \phi_n \phi_n^T = \phi^T R \phi \\
R_{nn} &= y_n (1 - y_n)
\end{align*}
\]

$$w^{(new)} = (\phi^T R \phi)^{-1} \phi^T R z$$

$$z = \phi w^{(old)} - R^{-1} (y - t)$$