Linear Classification (Part III: Logistic Regression)

nanopoulos@ismll.de

1

Outline

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

Discriminant vs. Probabilistic Discriminative

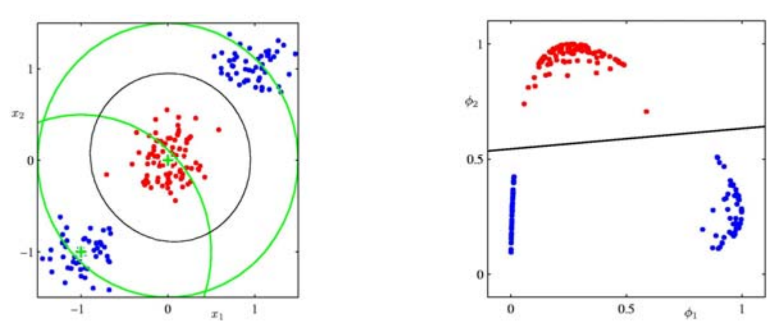
Use discriminant functions directly without probabilities:
- Fisher's LDA, Perceptron

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

$$p(class = C_k \mid \mathbf{x})$$

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 ϕ



Outline

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

Logistic regression

Logistic regression model

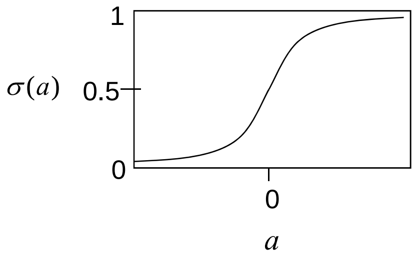
Posterior probability of a class for two-class problem:

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

$$P(C_2|\varphi) = 1 - P(C_1|\varphi)$$

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

Logistic function



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

$$\begin{aligned}
\frac{d\sigma}{da} &= \frac{e^{-a}}{(1+e^{-a})^2} \\
&= \sigma(a)\left\{\frac{e^{-a}}{1+e^{-a}}\right\} \\
&= \sigma(a)\left\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right\} \\
&= \sigma(a)(1-\sigma(a)).
\end{aligned}$$

Outline

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

Maximum Likelihood

- Determining **w** using ML
  - Likelihood function:

$$P(\mathbf{t}|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{(1-t_n)}$$

$$\mathbf{t} = (t_1, \ldots t_N)^T \qquad y_n = p(C_1|\phi_n) \qquad \begin{matrix} C_1: t_n = 1 \\ C_2: t_n = 0 \end{matrix}$$

  - Cross-entropy error function (negative log likelihood)

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

$$y_n = \sigma(a_n), a_n = w^T\phi_n$$

  - The gradient of the error function w.r.t. w

$$\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)\phi_n$$

9

Computing the gradient

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

$$y_n = \sigma(a_n), a_n = w^T\phi_n$$

$$\begin{aligned}\frac{\partial E}{\partial y_n} &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \\ &= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\ &= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \\ &= \frac{y_n - t_n}{y_n(1-y_n)}.\end{aligned}$$

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n). \qquad \nabla a_n = \phi_n$$

Computing the gradient

$$\begin{aligned}\nabla E &= \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla a_n \\ &= \sum_{n=1}^{N} (y_n - t_n) \phi_n\end{aligned}$$

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n) \phi_n$$

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)

Iterative reweighted least squares

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

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\phi_n = \Phi^T(\mathbf{y}-\mathbf{t})$$

Φ is NxM, n-th row = $(\varphi_n)^T$

$$H = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T = \Phi^T\mathbf{R}\Phi$$

$$R_{nn} = y_n(1-y_n)$$

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

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