Linear Discrimination

464.561A Models and Technologies for Information Services

Jonghun Park
jonghun@snu.ac.kr
Dept. of Industrial Eng.
Seoul National University

4/19/11

linear regression

$$y = \sum_{i=0}^{N} w_i \times f_i = \mathbf{w} \cdot \mathbf{f}$$

where $f_0 = 1$

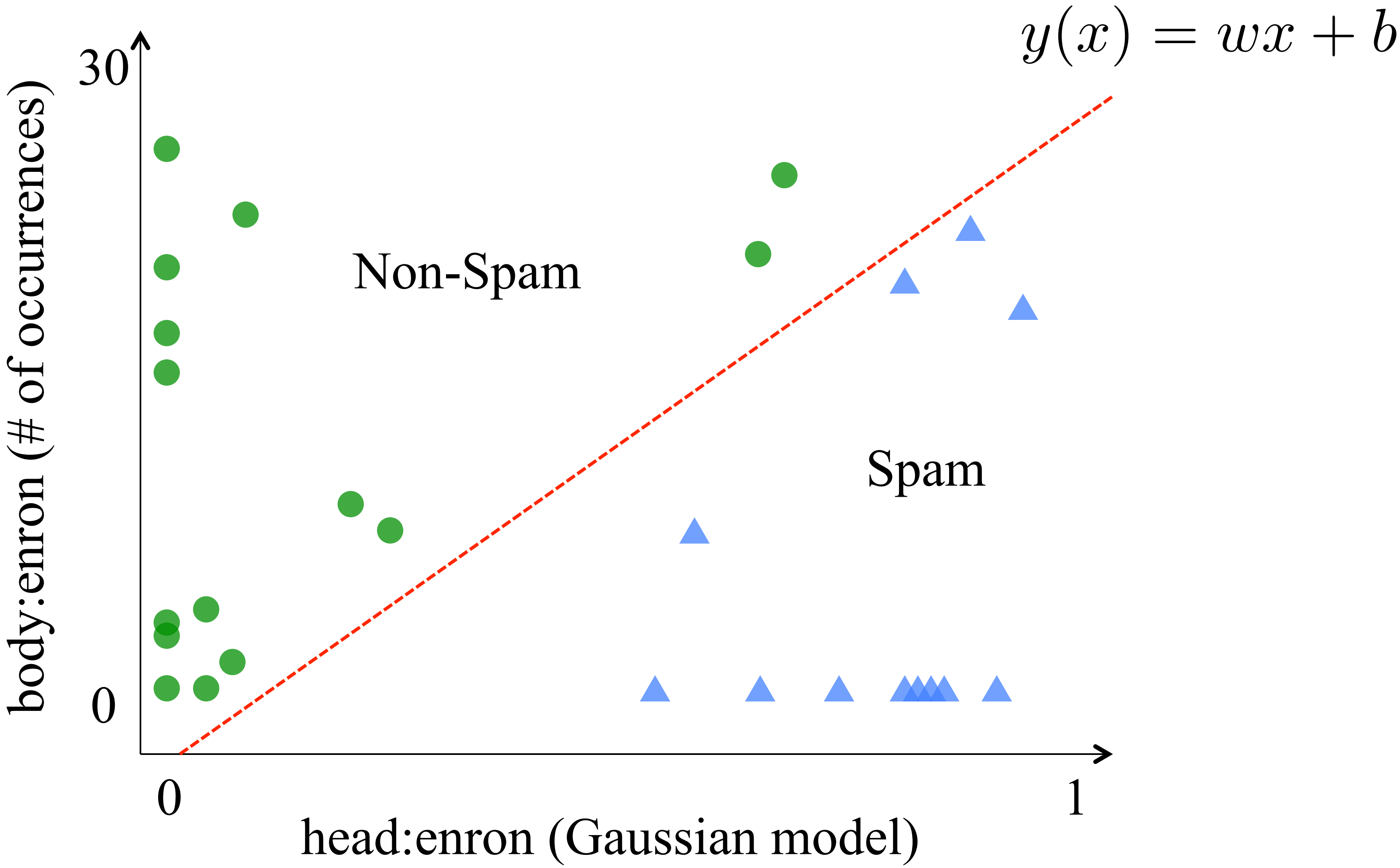
learning in linear regression

$$y_{pred}^{(j)} = \sum_{i=0}^{N} w_i \times f_i^{(j)}$$

$$cost(\mathbf{w}) = \sum_{j=0}^{M} (y_{pred}^{(j)} - y_{obs}^{(j)})^2$$

$$\mathbf{w_{ML}} = (X^T X)^{-1} X^T \mathbf{y}$$

Spam Email Classification – ENRON case



logistic regression

$$P(y = \text{true}|x) = \mathbf{w} \cdot \mathbf{f}$$

$$\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)} = \mathbf{w} \cdot \mathbf{f}$$

$$\ln\left(\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)}\right) = \mathbf{w} \cdot \mathbf{f}$$

$$\text{logit}(p(x)) \triangleq \ln\left(\frac{p(x)}{1 - p(x)}\right)$$

logistic regression

$$\ln\left(\frac{P(y=\mathrm{true}|x)}{1-P(y=\mathrm{true}|x)}\right)=\mathbf{w}\cdot\mathbf{f}$$

$$\frac{P(y=\mathrm{true}|x)}{1-P(y=\mathrm{true}|x)}=e^{\mathbf{w}\cdot\mathbf{f}}$$

$$P(y=\mathrm{true}|x)+P(y=\mathrm{true}|x)e^{\mathbf{w}\cdot\mathbf{f}}=e^{\mathbf{w}\cdot\mathbf{f}}$$

$$P(y=\mathrm{true}|x)=\frac{e^{\mathbf{w}\cdot\mathbf{f}}}{1+e^{\mathbf{w}\cdot\mathbf{f}}}$$

$$\therefore P(y=\mathrm{false}|x)=\frac{1}{1+e^{\mathbf{w}\cdot\mathbf{f}}}$$

logistic regression

$$P(y = \text{true}|x) = \frac{e^{\mathbf{w}\cdot\mathbf{f}}}{1 + e^{\mathbf{w}\cdot\mathbf{f}}} = \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{f}}}$$

$$P(y = \text{false}|x) = \frac{e^{-\mathbf{w}\cdot\mathbf{f}}}{1 + e^{-\mathbf{w}\cdot\mathbf{f}}}$$

classification using logistic regression

our observation should labeled 'true' if

$$P(y = \text{true}|x) > P(y = \text{false}|x)$$

$$\frac{P(y = \text{true}|x)}{P(y = \text{false}|x)} > 1$$

$$\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)} > 1$$

$$e^{\mathbf{w}\cdot\mathbf{f}} > 1$$

$$\Rightarrow \mathbf{w}\cdot\mathbf{f} > 0$$

learning in logistic regression

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \prod_j P(y^{(j)}|x^{(j)})$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_j \log P(y^{(j)}|x^{(j)})$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_j \log \begin{cases} P(y^{(j)} = 1|x^{(j)}), & \text{if } y^{(j)} = 1 \\ P(y^{(j)} = 0|x^{(j)}), & \text{if } y^{(j)} = 0 \end{cases}$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_j y^{(j)} \log P(y^{(j)} = 1|x^{(j)}) + (1 - y^{(j)}) \log P(y^{(j)} = 0|x^{(j)})$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_j y^{(j)} \log \frac{1}{1 + e^{-\mathbf{w}\cdot\mathbf{f}}} + (1 - y^{(j)}) \log \frac{e^{-\mathbf{w}\cdot\mathbf{f}}}{1 + e^{-\mathbf{w}\cdot\mathbf{f}}}$$

References

- E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2004.
- D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd Ed., Prentice Hall, 2009.