Machine Learning, Lecture 6: Logistic regression

S. Nõmm

[1]Department of Computer Science, Tallinn University of Technology

12.03.2015

Generative approach *versus* Discriminative approach

- *Generative* approach - create a model of the form $p(y, \boldsymbol{x})$ and then derive $p(y \mid \boldsymbol{x})$.
- *Discriminative* approach - fit the model of the form $p(y \mid \boldsymbol{x})$ directly.

Logistic regression

- Linear regression model $p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}(y \mid \mu(\boldsymbol{x}))$
  - Replace Gaussian distribution for y with a Bernoulli distribution (more appropriate for the binary response)

    $$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y \mid \mu(\boldsymbol{x}))$$

    where $\mu(\boldsymbol{x}) = \mathbb{E}[y \mid x] = p(y = 1 \mid x)$.
  - Ensure that $0 \leq \mu(\boldsymbol{x}) \leq 1$ by

    $$\mu(\boldsymbol{x}) = \mathrm{sigm}(\boldsymbol{\theta}^T x)$$

    where $\mathrm{sigm}(\eta)$ is the *sigmoid* or *logistic* or *logit* function:

    $$\mu(\boldsymbol{x}) = \frac{1}{1 + e^{-\eta}} = \frac{e^{\eta}}{e^{\eta} + 1}$$

- $$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y \mid \mathrm{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}))$$

Some important properties

- For the logistic function

$$g(\eta) = \frac{1}{1 + e^{-\eta}}$$

$$\begin{aligned} g(\eta) &= 0.5 \qquad \text{if} \quad \eta = 0 \\ g(\eta) &> 0.5 \qquad \text{if} \quad \eta > 0 \\ g(\eta) &< 0.5 \qquad \text{if} \quad \eta < 0 \end{aligned}$$

- Derivative of the logistic function

$$g'(\eta) = g(\eta)(1 - g(\eta))$$

Probabilistic interpretation

- Let us compute the probabilities of $y = 1$ and $y = 0$

$$\begin{aligned} P(y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) &= \operatorname{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}) \\ P(y = 0 \mid \boldsymbol{x}, \boldsymbol{\theta}) &= 1 - \operatorname{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}) \end{aligned}$$

  Could you write this statement in a more compact form?

$$P(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = ?$$

- The meaning of $\boldsymbol{\theta}^T \boldsymbol{x}$

$$g(\boldsymbol{\theta}^T \boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}^T \boldsymbol{x}}}$$

  after the straight but tedious calculations one gets

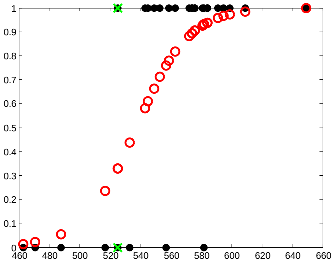
$$\boldsymbol{\theta}^T \boldsymbol{x} = \log \frac{g(\boldsymbol{\theta}^T \boldsymbol{x})}{1 - g(\boldsymbol{\theta}^T \boldsymbol{x})}$$

  here and after referred as *log -odds*, probability of event occurring is divided by the probability of not occurring.

Example

Denote x_i to be the SAT score of the student i and y_i is whether they passed or failed a class.

$$p(y_i = 1 \mid x_i \boldsymbol{w}) = \mathrm{sigm}(\omega_0 + \omega_1 x_i)$$



Likelihood

- Likelihood of the parameters (probability of the entire data set)

$$\mathcal{L}(\boldsymbol{\theta}) = P(Y \mid \boldsymbol{X}; \boldsymbol{\theta}) = \prod_{i=1}^{m} (\text{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}))^{y_i} (1 - \text{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}))^{1-y_i}$$

- We use log- likelihood which leads:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log \mathcal{L}(\boldsymbol{\theta}) \\
&= \log \prod_{i=1}^{m} (\text{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}))^{y_i} (1 - \text{sigm}(\boldsymbol{\theta}^T \boldsymbol{x}))^{1-y_i} \\
&= \sum_{i=1}^{m} \big(y_i \log \text{sigm}(\boldsymbol{\theta}^T x_i) + (1 - y_i) \log(1 - \text{sigm}(\boldsymbol{\theta}^T x_i)) \big)
\end{aligned}$$

Likelihood maximization

- Gradient descent to minimize the negative log-likelihood. Update step:

$$\theta_j^{k+1} = \theta_j^k - \alpha \frac{\partial}{\partial \theta_j^k} \ell(\boldsymbol{\theta})$$

- Gradient ascent to maximize log likelihood. Update step:

$$\theta_j^{k+1} = \theta_j^k + \alpha \frac{\partial}{\partial \theta_j^k} \ell(\boldsymbol{\theta})$$

- By derivation the log -likelihood one gets the gradient ascend update for the logistic regression:

$$\theta_j^{k+1} = \theta - j^k + \alpha \sum_{i=1}^{m} (y_i - \mathrm{sigm}(\boldsymbol{\theta}^T x_i)) x_{i,j}$$

simultaneously for each $\theta_j, \quad j = 0, \ldots n$.