Logistic Regression

Aarti Singh (Instructor), HMW-Alexander Noter)()

February 6, 2017

Back to Index

Contents

Resources

- Lecture

- Output need to be [0,1]
- It is a linear classifier, not a real regression

Sigmoid function

$$P(Y=0|X) = \frac{1}{1+\exp(w_0 + \sum w_i X_i)}$$

Decision boundar:

- Class 0: $P(Y=0|X) > P(Y=1|X)$, $\sum w_i X_i < 0$
- Class 1: $P(Y=0|X) < P(Y=1|X)$, $\sum w_i X_i > 0$

1 Training Logistic Regression

How to learn the parameters $w_0, \dots, w_d$ from training data $\{(X^{(i)}, Y^{(i)}\}$

1.1 Maximum Conditional Likelihood Estimate

$$\hat{\vec{w}}_{MLE} = \arg\max_{\vec{w}} \prod_{i=1}^{n} P(X^{(i)}, Y^{(i)}|\vec{w})$$

But we only know the model of $P(Y|X)$

Discriminative philosophy: Don't waste effort learning $P(X)$, focus on $P(Y|X)$ – that's all that matters for classification!

$$
\begin{aligned}
P(Y=0|X) &= \tfrac{1}{1+\exp(w_0+\sum w_i X_i)} \\
P(Y=1|X) &= \tfrac{\exp(w_0+\sum w_i X_i)}{1+\exp(\sum w_0+w_i X_i)}
\end{aligned}
\left.\vphantom{\begin{aligned} a \\ b \end{aligned}}\right\} \frac{\exp(w_0 y_i + \sum w_i X_i y_i)}{1+\exp(w_0+\sum w_i X_i)} \tag{1}
$$

$$
l(w) = ln \prod_i P(y^i|x^i, w) = \sum_i [y^i(w_0 + \sum w_j x^i_j) - \ln(1+\exp(w_0 + \sum w_j x^i_j))]
$$

No close-form solution to maximize w, but $l(w)$ is concave function of w.
Gradient Ascent Algorithm:

- Initialize: Pick w at random
- Gradient: $\nabla_w l(w) = [\frac{\partial l(w)}{\partial w_0}, \ldots, \frac{\partial l(w)}{\partial w_d}]^T$
- Update rule: $\Delta w = \eta \nabla_w l(w)$, and $w_i^{t+1} \leftarrow w_i^t + \Delta w$

$$
\frac{\partial l(w)}{\partial w_0} = \sum_i [y^i - P(Y^i = 1|x^i, w)]
$$

$$
\frac{\partial l(w)}{\partial w_j} = \sum_i x^i_j [y^i - P(Y^i = 1|x^i, w)]
$$

1.2 Maximum Conditional A Priori Estimate

Define priors on w

$$
p(w) = \prod_i \frac{1}{\kappa\sqrt{2\pi}} \exp -\frac{w_i^2}{2\kappa^2}
$$

$$
l(w) = lnP(w) \prod_i P(y^i|x^i, w) = \sum_i [y^i(w_0 + \sum w_j x^i_j) - \ln(1+\exp(w_0 + \sum w_j x^i_j))]
$$

2 Logistric Regression for More Than 2 classes

for $k < K$

$$
P(Y = y_k|X) = \frac{\exp(w_{k0} + \sum w_{ki} X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum w_{ji} X_i)}
$$

for $k = K$

$$
P(Y = y_K|X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum w_{ji} X_i)}
$$

The decision boundaries are still linear and they will intersect at one point. (Why?)