Logistic Regression

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Implement the fundamental learning algorithm Logistic Regression

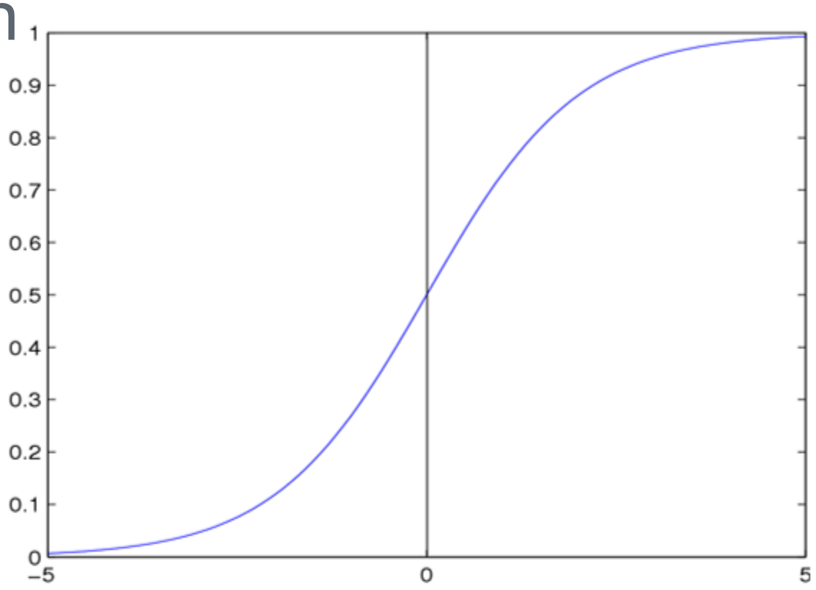
Discriminative Model: Example

- Again, we are given a training set of n labelled samples $<\mathbf{x}^{(i)}, y^{(i)}>$

- Why not directly model/learn $P(y|\mathbf{x})$?
  - Discriminative model

- Further assume $P(y|\mathbf{x})$ takes the form of a logistic sigmoid function

→ **Logistic Regression**



Logistic Regression

- Logistic regression: use the logistic function for modeling $P(\mathrm{y}|\mathbf{x})$, considering only the case of $\mathrm{y} \in \{0, 1\}$

- The *logistic function*

$$\sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^{t}}{1+e^{t}}$$

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}$$

$$P(y=1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}$$

Logistic Regression → Linear Classifier

- Given a sample **x**, we classify it as 0 (i.e., predicting y=0) if

$$P(\mathrm{y}=0|\mathbf{x}) \geq P(\mathrm{y}=1|\mathbf{x})$$

➔ This is a linear classifier.

The Parameters of the Model

What are the model parameters in logistic regression?

Given a parameter **w**, we have $P(y|\mathbf{x})$ =

$$[\sigma(w^t x)]^y[1-\sigma(w^t x)]^{1-y}$$

Suppose we have two different sets of parameters, $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, whichever giving a larger $P(y|\mathbf{x})$ should be a better parameter.

The Conditional Likelihood

Given n training samples, <$\mathbf{x}^{(i)}$, $y^{(i)}$>, i=1,…,n, how can we use them to estimate the parameters?

➔ For a given $\mathbf{w}$, the probability of getting all those $y^{(1)}$, $y^{(2)}$ …,$y^{(n)}$ from the corresponding data $\mathbf{x}^{(i)}$, i=1,…,n, is

$$P\left[y^{(1)}, y^{(2)}, \ldots, y^{(n)} \mid x^{(1)}, x^{(2)}, \ldots, x^{(n)}, w\right] = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; w)$$

$$= \prod_{i=1}^{n} \left[\sigma(w^t x^{(i)})\right]^{y^{(i)}} \left[1 - \sigma(w^t x^{(i)})\right]^{1-y^{(i)}}$$

➔ Call this $L(\mathbf{w})$, the (conditional) likelihood.

The Conditional Log-likelihood

$$\ell(w) = \log \mathcal{L}(w) = \log \prod_{i=1}^{n} (\cdots)$$

$$= \sum_{i=1}^{n} \log\left[\sigma(w^t x^{(i)})^{y^{(i)}} (1 - \sigma(w^t x^{(i)}))^{1-y^{(i)}}\right]$$

$$= \sum_{i=1}^{n} \left[\log\left(\sigma(w^t x^{(i)})^{y^{(i)}}\right) + \log\left((1 - \sigma(w^t x^{(i)}))^{1-y^{(i)}}\right)\right]$$

Maximizing Conditional Log Likelihood

- Optimal parameters

$$\begin{aligned}\mathbf{w}^* &= \text{argmax}_{\mathbf{w}} l(\mathbf{w}) \\ &= \text{argmax}_{\mathbf{w}} \sum_{i=1}^{n} [y^{(i)} \mathbf{w}^t \mathbf{x}^{(i)} - \log\left(1 + \exp\left(\mathbf{w}^t \mathbf{x}^{(i)}\right)\right)]\end{aligned}$$

- We cannot really solve for **w*** analytically (no closed-form solution)
  - We can use a commonly-used optimization technique, gradient descent/ascent, to find a solution.

Finding the gradient of $l(\mathbf{w})$

Recall: $\frac{\partial (w^t x)}{\partial w} = x$, $\left(\frac{\partial \log f(x)}{\partial x} = \frac{1}{f(x)} \frac{\partial f(x)}{\partial x} \right)$

$\frac{\partial e^x}{\partial x} = e^x$

$$\nabla_w l(w) = \nabla_w \left[\sum_{i=1}^{n} \left(y^{(i)} w^t x^{(i)} - \log\left(1 + e^{w^t x^{(i)}}\right) \right) \right],$$

$$= \sum_{i=1}^{n} \left[y^{(i)} x^{(i)} - \frac{e^{w^t x^{(i)}} \cdot x^{(i)}}{1 + e^{w^t x^{(i)}}} \right]$$

(Setting this to 0 cannot really give us a closed-form solution for w.
So we will do gradient ascent.)

Gradient Ascent Algorithm

The algorithm

Iterate until converge

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \eta \nabla_{\mathbf{w}^{(k)}} l(\mathbf{w})$$

$\eta > 0$ is a constant called the learning rate.