The logistic regression model

Logistic regression is a discriminative, linear model for binary classification. That is, it models the probability distribution $p(y \mid \mathbf{x})$ where y is the class label of the item (either -1 or 1), and $\mathbf{x}$ is its feature representation. Once $p(y \mid \mathbf{x})$ is learned, the model will classify a new item as belonging to class 1 if $p(y = 1 \mid \mathbf{x}) > t$ and -1 otherwise where t is a threshold that can be determined by the user (usually, we choose $t = 0.5$). Stated differently, logistic regression finds a hypothesis of the form

$$h(\mathbf{x}) = \begin{cases} 1 : p(y = 1 \mid \mathbf{x}) > t \\ -1 : \text{otherwise} \end{cases}$$

where the probability distribution $p(y \mid \mathbf{x})$ is represented as

$$\begin{aligned} p(y = 1 \mid \mathbf{x}) &= \frac{1}{1 + e^{-\boldsymbol{\beta}^\top \mathbf{x}}} \\ &= \sigma\left(\boldsymbol{\beta}^\top \mathbf{x}\right) \end{aligned}$$

where σ is the **sigmoid function**

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

and $\boldsymbol{\beta}$ is the weight-vector. Thus, each hypothesis in the considered hypothesis space is characterized by a $\boldsymbol{\beta}$ vector. Choosing a hypothesis, then, is akin to finding an appropriate $\boldsymbol{\beta}$. For example, we can choose a $\boldsymbol{\beta}$ to be the maximum likelihood estimate of a training dataset. As another example, a Bayesian method can be employed to derive a posterior distribution over $\boldsymbol{\beta}$.

Motivation

For the sake of brevity, let

$$\theta := p(y = 1 \mid \mathbf{x})$$

. Logistic regression is motivated by the attempt to model θ as a linear combination of the components of $\mathbf{x}$. That is, we believe that

$$\theta = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

. The problem with this model is that it does not inherently guarantee that θ will be between zero and one and thus there is no guarantee that it will be a proper probability. Is there a way to force θ to stay between zero and one while allowing it to depend on a linear combination of the features? The trick to accomplishing this is to model the *log odds* of θ as this function rather than θ directly:

$$\log\left(\frac{\theta}{1-\theta}\right) = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

. Recall the log odds is given by the logit function. Thus, the model becomes

$$\text{logit}(\theta) = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

. Recall that the inverse of the logit function is the sigmoid function. Thus, we can express θ as

$$\begin{aligned}\theta &= \text{logit}^{-1}\left(\alpha + \sum_{i=1}^{k} \beta_i x_i\right) \\ &= \sigma\left(\alpha + \sum_{i=1}^{k} \beta_i x_i\right)\end{aligned}$$

. Furthermore, we note that since the sigmoid function lies between zero and one, the math works out to ensure that θ is a valid probability. Lastly, we note that if we let the first element of the $\boldsymbol{\beta}$ vector be α and let the first element of any feature vector $\mathbf{x}$ be 1, then we have come to the logistic regression model:

$$\theta = \sigma\left(\boldsymbol{\beta}^\top \mathbf{x}\right)$$

.

Logistic regression is a linear model

Logistic regression is a linear classifier due to the fact that the decision boundary is a hyperplane. This arises from the fact that $p(y \mid \mathbf{x})$ is modeled as a monotonic function of $\mathbf{x}^\top \boldsymbol{\beta}$. We note that in order to classify an item $\mathbf{x}$ as 1, we need $p(y \mid \mathbf{x}) > 0.5$. This will occur if $\boldsymbol{\beta}^\top \mathbf{x} > 0$. Thus, the decision boundary is

$$\boldsymbol{\beta}^\top \mathbf{x} = 0$$

, which is a hyperplane. Figure 1 illustrates the decision boundary of a logistic regression classifier learned on 2-dimensional data.

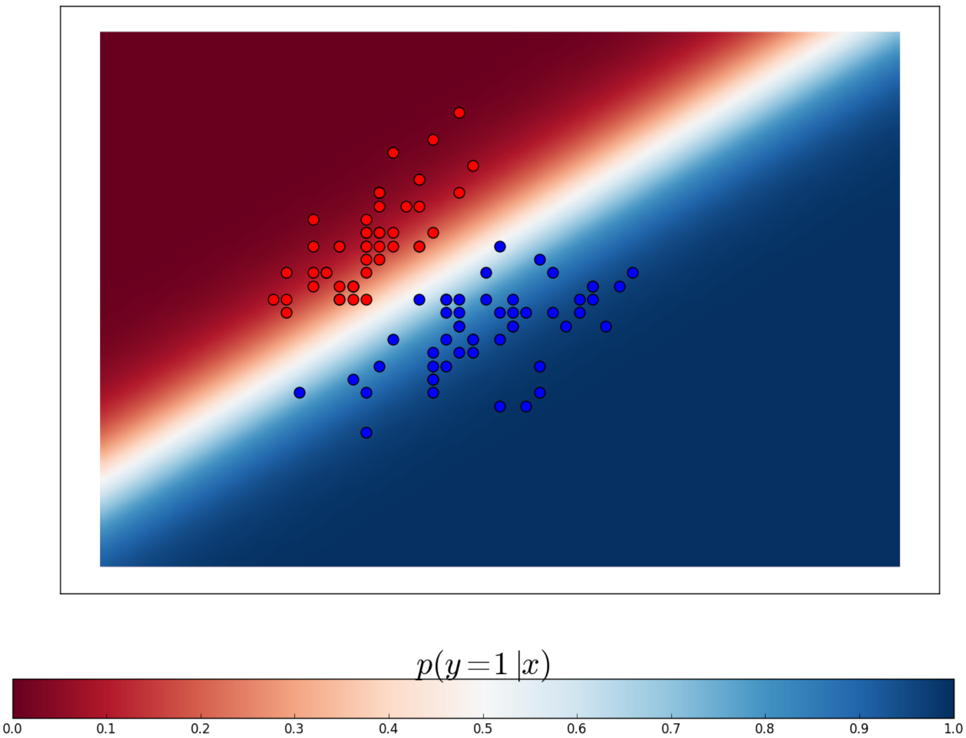

Figure 1: Plotting $p(y = 1 \mid \mathbf{x})$. Note that the decision boundary at $p(y = 1 \mid \mathbf{x}) = 0.5$ is linear.

