EECS 553: Machine Learning (ECE) **University of Michigan**

Logistic Regression

Winter 2023 **Clayton Scott**

1 Logistic Regression

Consider a binary classification problem with labels $y \in \{0,1\}$. The Bayes classifier may be expressed

$$f^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \eta(\boldsymbol{x}) \geqslant \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where

$$\eta(\boldsymbol{x}) := \Pr(Y = 1 | \boldsymbol{X} = \boldsymbol{x}).$$

In a nutshell, *logistic regression*[1] is a plug-in method that assumes

$$\eta(\boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}^T\boldsymbol{x} + b)}} \tag{2}$$

for some $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$. Given training data, the maximum likelihood estimate (MLE)

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{b} \\ \widehat{\boldsymbol{w}} \end{bmatrix} \qquad \text{of} \qquad \boldsymbol{\theta} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

is computed and plugged in to (2), yielding the estimate

$$\widehat{\eta}(\boldsymbol{x}) = \frac{1}{1 + e^{-(\widehat{\boldsymbol{w}}^T\boldsymbol{x} + \widehat{b})}}.$$

The function $\frac{1}{1+e^{-t}}$ is called a *logistic* or *sigmoid* function. See Figure 1.

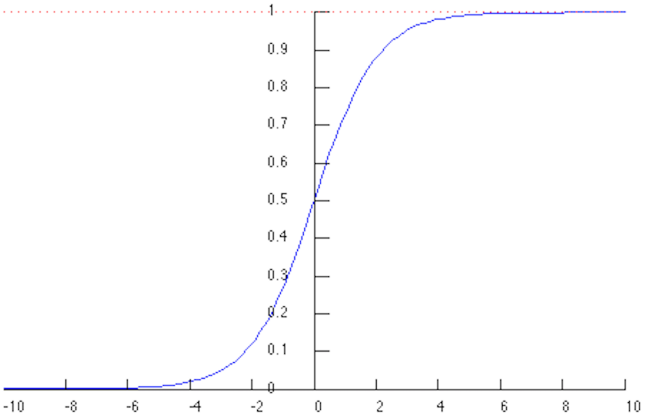


Figure 1: Sigmoid function.

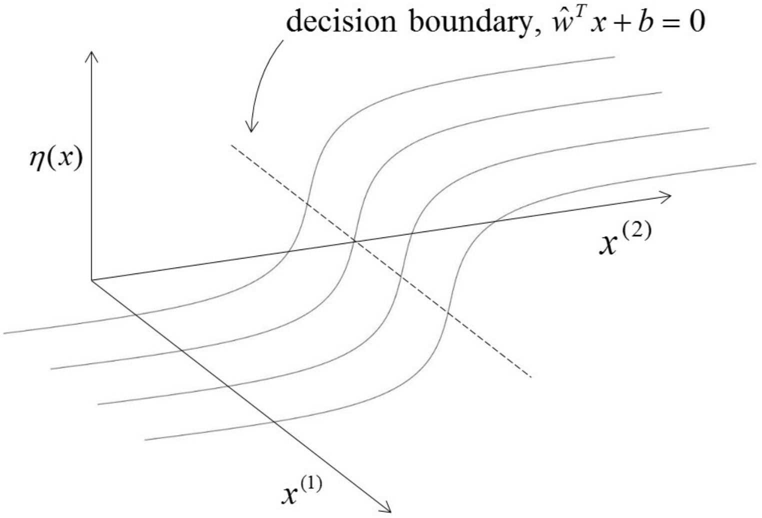
[1] Despite the name, logistic regression is a method for classification. Of course, classification may be viewed as a special case of regression, so the name still makes some sense. The name makes even more sense when considered as a variant of linear regression within the class of generalized linear models.

2 Linearity

Observe that

$$\begin{aligned}\widehat{f}(\boldsymbol{x}) = 1 &\iff \frac{1}{1+e^{-(\widehat{\boldsymbol{w}}^T\boldsymbol{x}+\widehat{b})}} \geqslant \frac{1}{2} \\ &\iff \widehat{\boldsymbol{w}}^T\boldsymbol{x} + \widehat{b} \geqslant 0.\end{aligned}$$

Therefore logistic regression produces a *linear* classifier.



3 Maximum Likelihood Estimation

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. Logistic regression does not model the marginal distribution of $\boldsymbol{X}$, so we will treat $\boldsymbol{x}_i$ as fixed and maximize the *conditional (log) likelihood.* Thus let $p(y \,|\, \boldsymbol{x}; \boldsymbol{\theta})$ denote the conditional pmf of y given $\boldsymbol{x}$. Then the conditional likelihood of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) := \prod_{i=1}^{n} p(y_i \,|\, \boldsymbol{x}_i; \boldsymbol{\theta}) \tag{3}$$

where, in taking the product, we have assumed conditional independence of the labels given the feature vectors.

Given $\boldsymbol{X} = \boldsymbol{x}$, Y may be viewed as a Bernoulli trial with success probability $\eta(\boldsymbol{x}; \boldsymbol{\theta})$, where the notation for η now reflects the dependence on $\boldsymbol{\theta}$. Therefore

$$p(y \,|\, \boldsymbol{x}; \boldsymbol{\theta}) = \eta(\boldsymbol{x}; \boldsymbol{\theta})^y (1 - \eta(\boldsymbol{x}; \boldsymbol{\theta}))^{1-y}, \qquad y = 0, 1,$$

which is just the pmf of a Bernoulli trial. Plugging this into (3), and using the notation

$$\begin{aligned}\tilde{\boldsymbol{x}} &= [1 \ x_1 \ \cdots \ x_d]^T \\ \boldsymbol{\theta} &= [b \ w_1 \cdots \ w_d]^T,\end{aligned}$$

we arrive at the log-likelihood

$$\begin{aligned}\ell(\boldsymbol{\theta}) &:= \log L(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1+e^{-\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i}} \right) + (1-y_i) \log \left(\frac{e^{-\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i}}{1+e^{-\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i}} \right) \right].\end{aligned}$$

Although the equation $\nabla \ell(\boldsymbol{\theta}) = \mathbf{0}$ does not have a closed form solution, we have an explicit function of $\boldsymbol{\theta}$ that can be optimized using iterative optimization procedures like gradient ascent and Newton's method.

4 Regularized Logistic Regression

It is often preferable to add a term to the negative log-likelihood. We write $\ell(\boldsymbol{w}, b)$ for $\ell(\boldsymbol{\theta})$ to indicate the dependence on $\boldsymbol{w}$ and b more explicitly. Then a common objective function for logistic regression is

$$J(\boldsymbol{w}, b) = -\ell(\boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2, \tag{4}$$

where $\lambda > 0$ is a fixed, used-specified constant called a *regularization parameter*. The second term is called the *regularizer*, *regularization term*, or *penalty*. Intuitively, the regularization term encourages the estimate $\widehat{\boldsymbol{w}}$ to be small, which can prevent *overfitting* to the training data. This is especially important when $n < d$, in which case $-\ell$ may have infinitely many minimizers. From an optimization perspective, the regularizer makes the objective strictly convex, which ensures a unique minimizer and makes iterative solvers more effective. Strict convexity also implies that the Hessian is always positive definite, which makes it possible to apply second order methods like Newton's method.

An alternate objective function is

$$J(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|^2, \tag{5}$$

where the regularizer differs in that the offset b is now included. There is no good statistical reason to include b in the penalty, but this makes it easier to prove that J is strictly convex. See exercises.

5 Class Probability Estimation

Logistic regression actually solves a more general problem than classification, namely, *class probability estimation*. Given a test point $\boldsymbol{x}$, $\eta(\boldsymbol{x}; \widehat{\boldsymbol{\theta}})$ is an estimate of the probability that $\boldsymbol{x}$ has a label of 1.

Exercises

1. (★) Show that if we change the label convention to $y \in \{-1, 1\}$, then

$$-\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i\right)\right).$$

2. Introduce the notation $\phi(t) = \log(1 + \exp(-t))$ so that, by the previous problem, (5) may be written

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \phi(y_i \boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i) + \lambda \|\boldsymbol{\theta}\|^2.$$

   (a) (★) Using this form, compute the gradient $\nabla J(\boldsymbol{\theta})$.

   (b) (★★) Based on your answer to part (a), find the Hessian $\nabla^2 J(\boldsymbol{\theta})$.

   (c) (★) Using your answer to (b), prove that J is convex, and strictly convex when $\lambda > 0$.

   (d) (★★★) There is no reason to penalize the magnitude of b, and so a preferable penalty is $\lambda \|\boldsymbol{w}\|^2$. Show that if $\lambda > 0$, then the objective is still strictly convex with this penalty.

3. (★★) Show that the LDA assumption implies the logistic regression assumption. That is, if the class-conditional densities are multivariate Gaussian with common covariance, then $\eta(\boldsymbol{x})$ has the form of (2) for certain $\boldsymbol{w}$ and b. Give formulas for $\boldsymbol{w}$ and b in terms of the LDA model parameters.