LOGISTIC REGRESSION

Consider a binary classification problem with labels $y = 0, 1$. The Bayes classifier may be expressed

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

where

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

LR in a nutshell

1. Assume $\eta(x) = \dfrac{1}{1 + e^{-(w^T x + b)}}$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$

2. Compute the MLE $\hat{\theta} = \begin{bmatrix} \hat{b} \\ \hat{w} \end{bmatrix}$ of $\theta = \begin{bmatrix} b \\ w \end{bmatrix} \in \mathbb{R}^{d+1}$.

3. Plug the estimate

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

into the formula for the Bayes classifier.

The function $\frac{1}{1+e^{-t}}$ is called a <u>logistic</u> or <u>sigmoid</u> function.

Denote the LR classifier

$$\hat{f}(x) = \mathbb{1}_{\{\hat{\eta}(x) \geq \frac{1}{2}\}}$$

Observe that

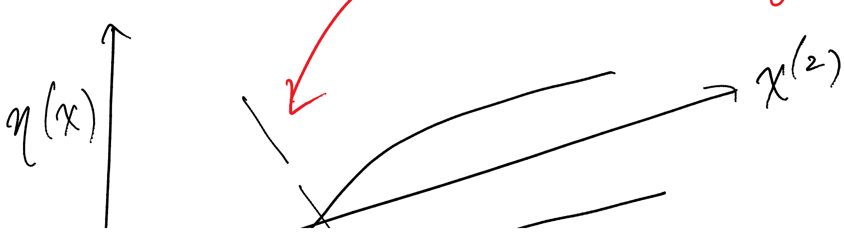
$$\hat{f}(x) = 1 \iff \frac{1}{1+e^{\hat{w}^T x + \hat{b}}} \geq \frac{1}{2}$$

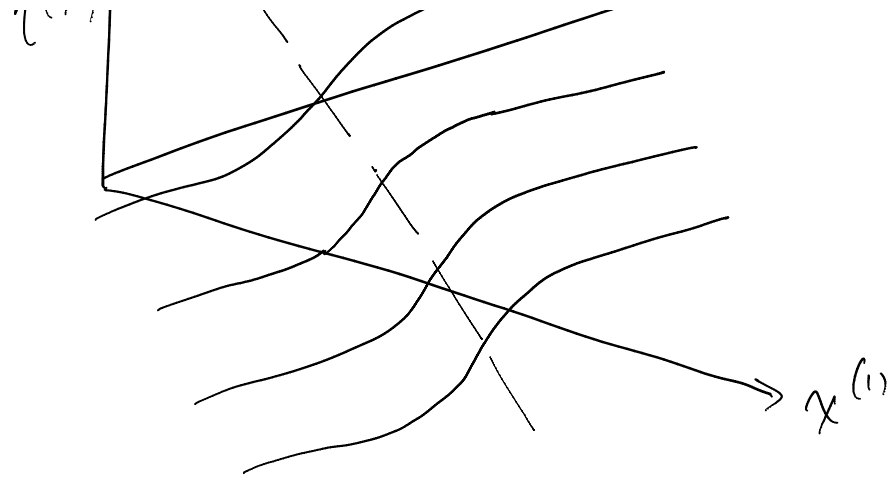
$$\iff e^{\hat{w}^T x + \hat{b}} \geq 1$$

$$\iff \hat{w}^T x + \hat{b} \geq 0.$$

Therefore $\hat{f}(x) = \mathbb{1}_{\{\hat{w}^T x + \hat{b}\}}$ and we see LR is a <u>linear</u> method.

decision boundary, $\hat{w}^T x + \hat{b} = 0$





Maximum Likelihood Estimation

Let $(x_1, y_1), \ldots, (x_n, y_n)$. LR does not model the marginal distribution of X, so we will treat X as fixed and maximize the conditional (log) likelihood.

Thus, let $p(y \mid x ; \theta)$ denote the condition pmf of y given x. Then the conditional likelihood of θ is

$$L(\theta) := \prod_{i=1}^{n} p(y_i \mid x_i ; \theta)$$

where we have assumed conditional independence of the labels given the feature vectors.

What is $p(y|x)$ in terms of $\eta(x;\theta)$?

Well, $Y|X$ is Bernoulli with success probability $\eta(x;\theta)$, so

$$p(y|x;\theta) = \begin{cases} \eta(x;\theta) & \text{if } y=1 \\ 1-\eta(x;\theta) & \text{if } y=0 \end{cases}$$

$$= \eta(x;\theta)^y \left(1-\eta(x;\theta)\right)^{1-y}.$$

Thus,

$$L(\theta) = \prod_{i=1}^{n} \eta(x_i;\theta)^{y_i} \left(1-\eta(x_i;\theta)\right)^{1-y_i}$$

and the log-likelihood $\ell(\theta) := \log L(\theta)$ is

$$\ell(\theta) = \sum_{i=1}^{n} y_i \log\left(\eta(x_i;\theta)\right) + (1-y_i)\log\left(1-\eta(x_i;\theta)\right)$$

Let's introduce some more notation:

$$\tilde{x} = [1 \;\; x^{(1)} \;\; \cdots \;\; x^{(d)}]^T$$

$$\theta = [b \;\; w^{(1)} \;\; \cdots \;\; w^{(d)}]^T$$

Then

Then

$$\ell(\theta) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{1}{1+e^{-\theta^T \tilde{x}_i}}\right) + (1-y_i)\log\left(\frac{e^{-\theta^T \tilde{x}_i}}{1+e^{-\theta^T \tilde{x}_i}}\right)\right]$$

Exercise Show that if we modify the label convention to $y \in \{-1, +1\}$, then

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

Regularized Logistic Regression

Unless $n \gg d$, it is preferable to minimize the modified objective function

$$J(\theta) = -\ell(\theta) + \lambda \|\theta\|^2$$

where $\lambda > 0$ is a fixed, user-specified constant called the <u>regularization parameter.</u>

Why introduce the regularization term? In brief:

- if $n < d$, $\nabla^2 \ell(\theta)$ won't be invertible
- $\ell_\lambda(\theta)$ is strictly convex, so it has a

- $\ell_\lambda(\theta)$ is strictly convex, so it has a unique minimizer
- Newton's method has nice convergence properties — see Boyd and Vandenberghe
- The regularization term encourages a small (intuitively, "simple") solution, which can prevent overfitting to the training data — important when the sample size n is small relative to d.

We'll talk more about regularization later.

Newton's Method

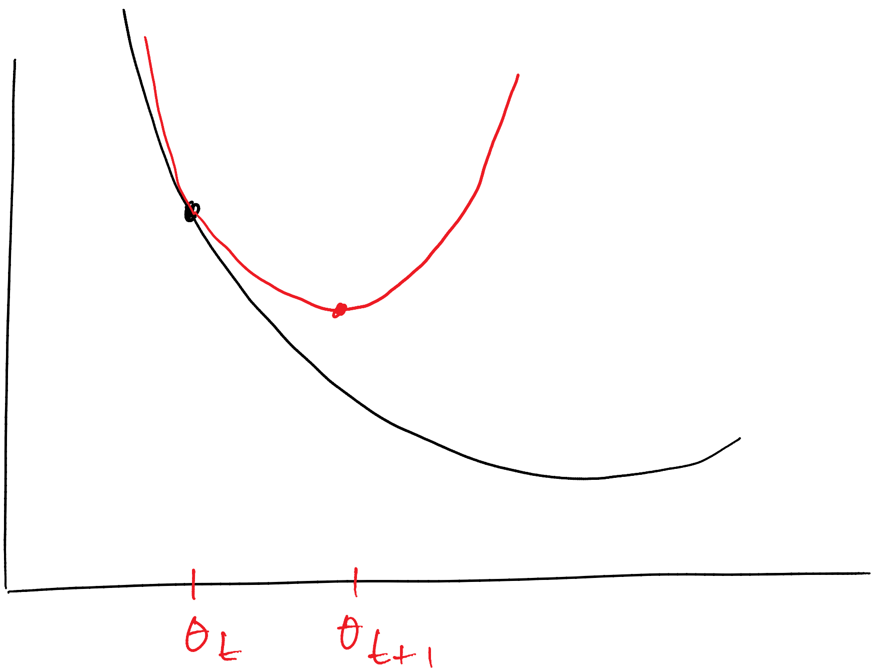
Solving $J'(\theta) = 0$ analytically is impossible (try it!). However, $J(\theta)$ is convex, and so we can use numerical methods. A common approach is Newton's method aka the Newton-Raphson algorithm:

Newton-Raphson algorithm:

$$\theta_{t+1} = \theta_t - \left(\nabla^2 J(\theta_t)\right)^{-1} \nabla J(\theta_t)$$

Newton's method can be viewed as minimizing the second order approximation

$$J(\theta) \approx J(\theta_t) + \nabla J(\theta_t)^T (\theta - \theta_t)$$
$$+ (\theta - \theta_t)^T \nabla^2 J(\theta_t) (\theta - \theta_t)$$



Final Thought

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