Quiz 2 : next Wednesday (Oct 25th)
Hw 3 will be out this week

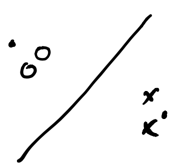
Outline ML estimation / MAP estimation / Laplacian Smooth
Logistic Regression

Recap: LDA / QDA / Naive Bayes

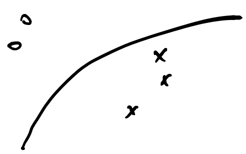
$$\hat{f}(x) = \arg\max_{k=1,\dots,K} P(Y=k \mid X=x)$$

$$= \arg\max_{k=1,\dots,K} P(Y=k) \cdot P(X=x \mid Y=k)$$

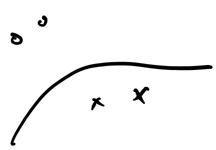
class conditional dist



LDA $N(\mu_k, \Sigma)$



QDA $N(\mu_k, \Sigma_k)$



NB $\prod_{j=1}^{d} g^{(j)}(x)$

(ex. Gaussian NB $g^{(j)} \sim N(\mu_j, \sigma_j^2)$)

NB discrete model (Laplacian Smoothing)

$$P(\underbrace{\text{"Michigan"}}_{w_1} \mid Y = \text{"sports"}) \approx \frac{\sum_{i \in \text{sports}} (c(d_i, w_1) + 1)}{\sum_j \sum_{i \in \text{sports}} (c(d_i, w_j) + 1)}$$

Why add 1?

Example coin tossing   Estimate $P(\text{heads}) = \theta$ from n independent coin tosses

$$P(x_i \mid \theta) = \begin{cases} \theta & x_i = 1 \\ 1-\theta & x_i = 0 \end{cases} = \theta^{x_i} \cdot (1-\theta)^{1-x_i}$$

Likelihood function: $p(\underline{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$

N_1: # heads

N_0: # tails

$$= \theta^{\sum_{i=1}^{n} x_i} \cdot (1-\theta)^{\sum_{i=1}^{n} 1-x_i}$$

$$= \theta^{N_1} \cdot (1-\theta)^{N_0}$$

Maximum Likelihood (ML) estimate of θ

$$\arg\max_{\theta} p(\underline{x} \mid \theta)$$

$$= \arg\max_{\theta} \log p(\underline{x} \mid \theta) = \arg\max_{\theta} N_1 \log\theta + N_0 \log(1-\theta)$$

Concave $\Rightarrow$ $\nabla \log p(x \mid \theta) = 0$

$$\frac{N_1}{\theta} + \frac{N_0}{1-\theta} = 0 \Rightarrow \theta = \frac{N_1}{N_0 + N_1} = \frac{\#\text{heads}}{n}$$

$\rightarrow$ Now suppose θ is random

$\theta \sim$ Beta distribution (α_1, α_0)

$$p(\theta) = \theta^{\alpha_1 - 1} \cdot (1-\theta)^{\alpha_0 - 1}$$

$$p(\theta \mid x_1, \ldots x_n) \propto p(x_1, \ldots x_n \mid \theta) \cdot p(\theta)$$

$$= \theta^{N_1} \cdot (1-\theta)^{N_0} \cdot \theta^{\alpha_1 - 1} \cdot (1-\theta)^{\alpha_0 - 1}$$

$N_1 + \alpha_1 - 1$ $N_0 + \alpha_0 - 1$

$$= \theta \quad \cdot (1-\theta)$$

Maximum a posteriori estimation (MAP)

$$\arg\max_{\theta} \; p(\theta \mid x_1, \dots, x_n) = \frac{N_1 + \alpha_1 - 1}{N_0 + N_1 + \alpha_1 + \alpha_0 - 2}$$

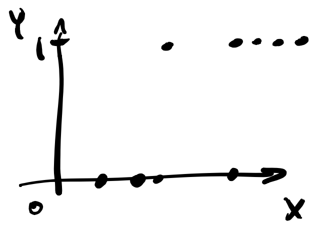
add $\alpha_1 - 1$ to heads
$\alpha_0 - 1$ to tails

$\alpha_1 = \alpha_0 = 2 \Rightarrow$ add 1
Laplacian Smoothing

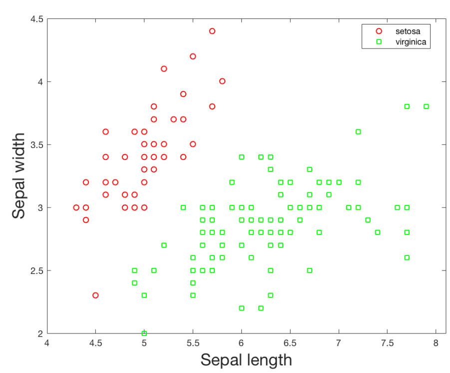
Multinomial case, dirichlet prior implies adding 1 to counts.

Logistic Regression

linear regression



In LDA/QDA/NB we made assumptions on $(x_i^{(1)}, \dots, x_i^{(d)})$



prob. model of
$P(Y=1 \mid X=x)$

Consider a binary classification problem w. labels $y \in \{0,1\}$

The Bayes Classifier

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

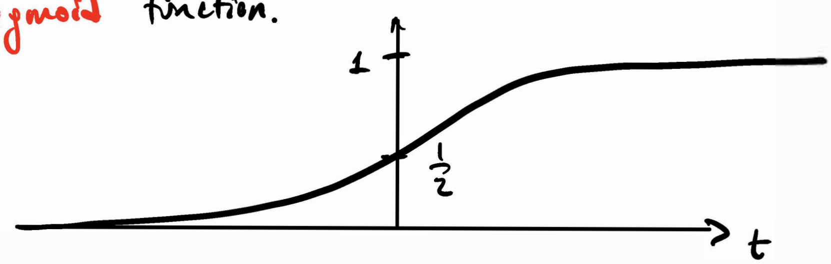
$\eta(x) = P(Y=1 \mid X=x)$ posterior dist.

In LR we assume $\eta(x) = \dfrac{1}{1+e^{-(w^T x+b)}}$

$w \in \mathbb{R}^d$, $b \in \mathbb{R}$ $\quad \underline{\theta} = \begin{bmatrix} w \\ b \end{bmatrix}$. Then we find ML estimates of w and b: $\hat{w}, \hat{b}$, plug in to the Bayes Classifier rule

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

The function $\dfrac{1}{1+e^{-t}}$ is called a logistic or sigmoid function.



$$p(Y=1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}}$$

$$\text{odds} = \frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)} = e^{+(w^T x+b)}$$

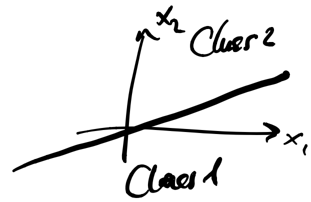
$$\Rightarrow \quad -\log \frac{P(Y=1 \mid x)}{1-P(Y=1 \mid x)} = w^T x + b = w_1 x_1 + \cdots + w_d x_d + b$$

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

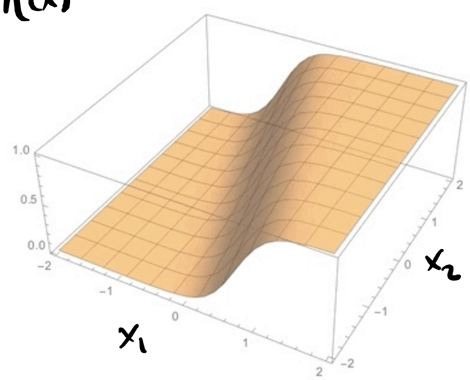
$$\hat{f}(x)=1 \iff \frac{1}{1+e^{-\hat{w}^T x-\hat{b}}} \geq \frac{1}{2}$$

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

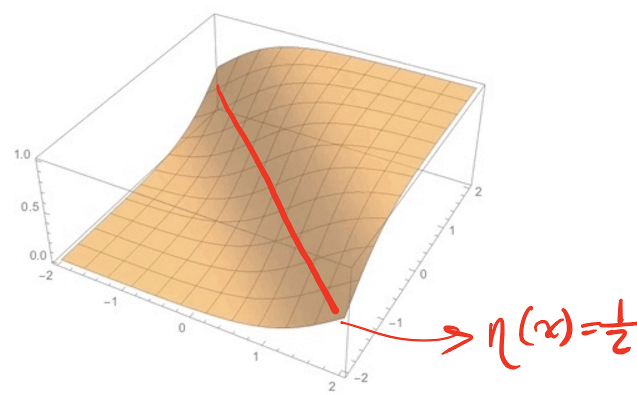
$\hat{f}(x)$ is a linear classifier



$\eta(x)$



$\begin{matrix} w_1 \\ w_2 \end{matrix} = \begin{pmatrix} 4.7 \\ 1.70587 \end{pmatrix}$



$\begin{matrix} w_1 \\ w_2 \end{matrix} = \begin{pmatrix} 2. \\ 2. \end{pmatrix}$

ML estimation

Let $(x_1, y_1), \dots, (x_n, y_n)$. LR does not model the distribution of x. We will treat x as fixed and maximize conditional likelihood

$p(y \mid x;\theta)$ denotes conditional pmf of y given x.

Conditional likelihood of θ is

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

(assumed labels are cond. indep. given x_i)

$$L(\theta) = \prod_{i=1}^{n} p(x_i)^{y_i} \cdot (1-p(x_i))^{1-y_i}$$

where $p(x_i) = P(Y=1 \mid X = x_i)$