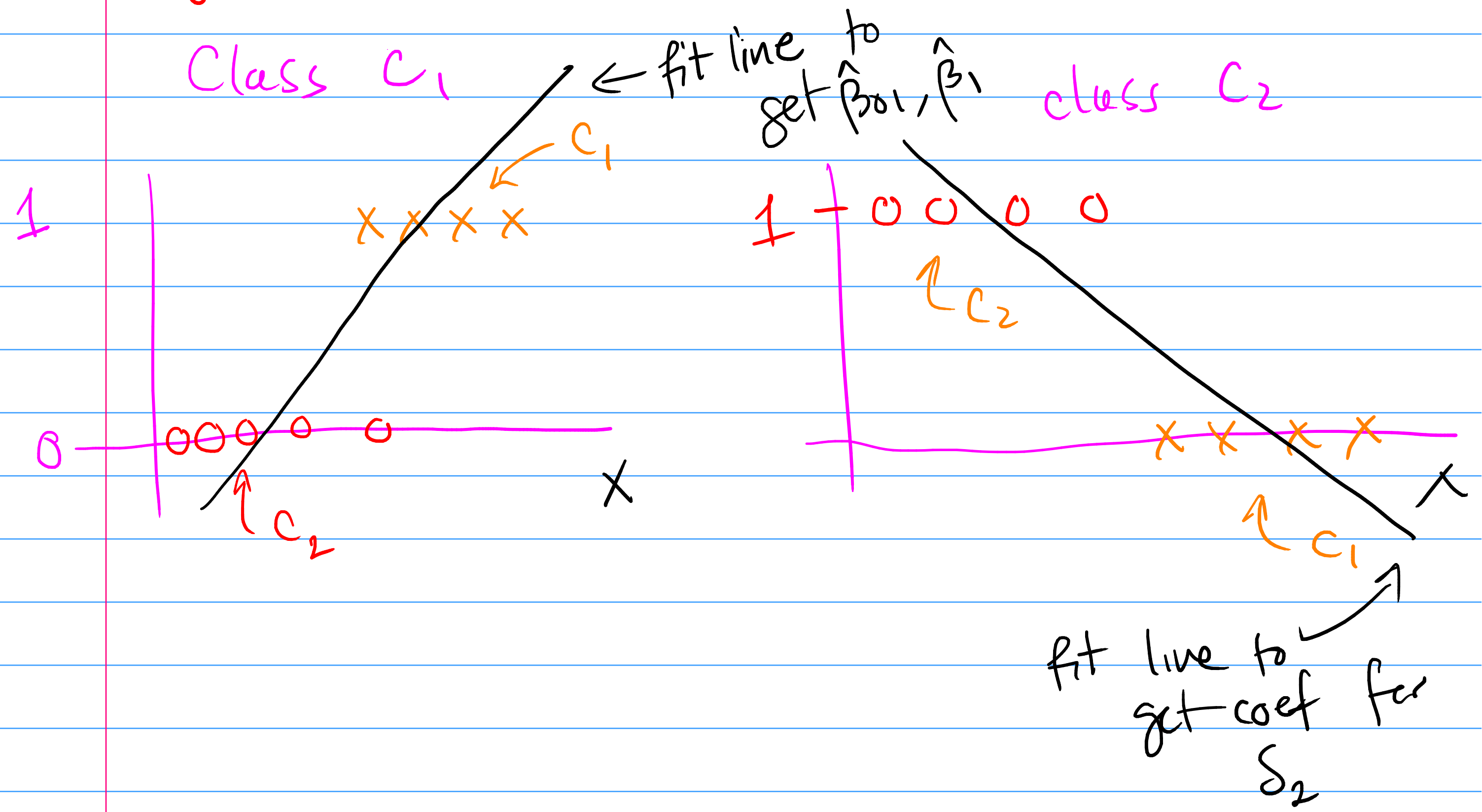
Lecture 11: Logistic Regression

LDA is a linear classifier

$$\delta_c(\underset{\sim}{x}) = \hat{\beta}_{0c} + \hat{\beta}_c^T \underset{\sim}{x}$$

Why not just fit δ_c using linear regression?

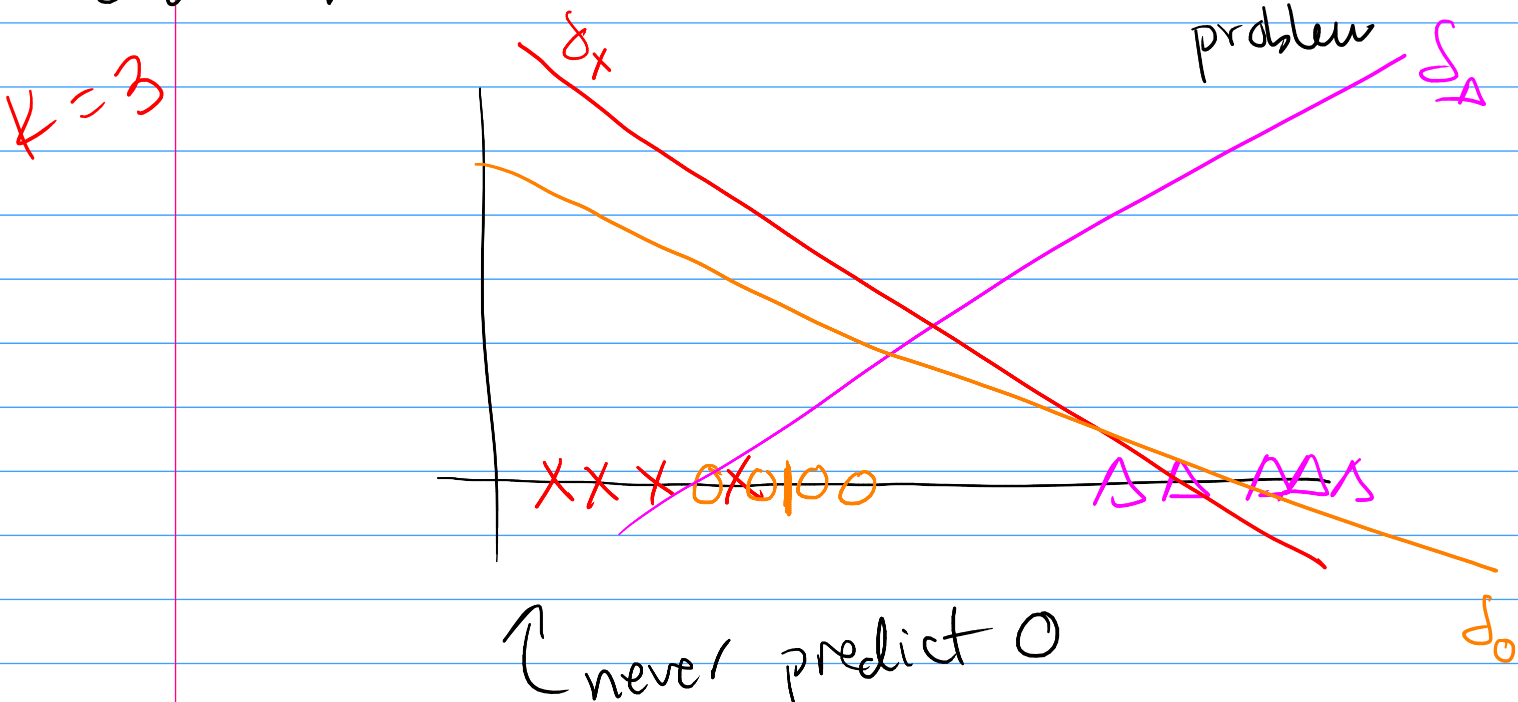
Binary example $Y = c_1$ or $Y = c_2$



Punchline: - reasonable when K = # classes is small

($K=2$ exactly gives LDA)

When K is "large" can have a masking problem



Logistic Regression

LDA: $\delta_c(\underset{\sim}{x}) = \mathbb{P}(Y = c \mid \underset{\sim}{X} = \underset{\sim}{x})$

$$\propto \mathbb{P}(\underset{\sim}{X} = \underset{\sim}{x} \mid Y = c)\,\mathbb{P}(Y = c)$$

Logistic Reg.:

directly model $\delta_c(\underset{\sim}{x}) = \mathbb{P}(Y = c \mid \underset{\sim}{X} = \underset{\sim}{x})$

Binary Classification ($K = 2$)

So $Y = 0$ or $Y = 1$ $\quad$ ($Y = \pm 1$)

$$\delta_0(\underset{\sim}{x}) = \mathbb{P}(Y = 0 \mid \underset{\sim}{X} = \underset{\sim}{x})$$

$$\delta_1(\underset{\sim}{x}) = \mathbb{P}(Y = 1 \mid \underset{\sim}{X} = \underset{\sim}{x}) = 1 - \mathbb{P}(Y = 0 \mid \underset{\sim}{X} = \underset{\sim}{x}) = 1 - \delta_0(\underset{\sim}{x})$$

So I really only need δ_1

$$\hat{f}(\underset{\sim}{x}) = \arg\max_c \delta_c(\underset{\sim}{x})$$

$$\hat{f}(\underset{\sim}{x}) = 1 \quad \text{when} \quad \delta_1(\underset{\sim}{x}) > \delta_0(\underset{\sim}{x})$$
$$> 1 - \delta_1(\underset{\sim}{x})$$

so

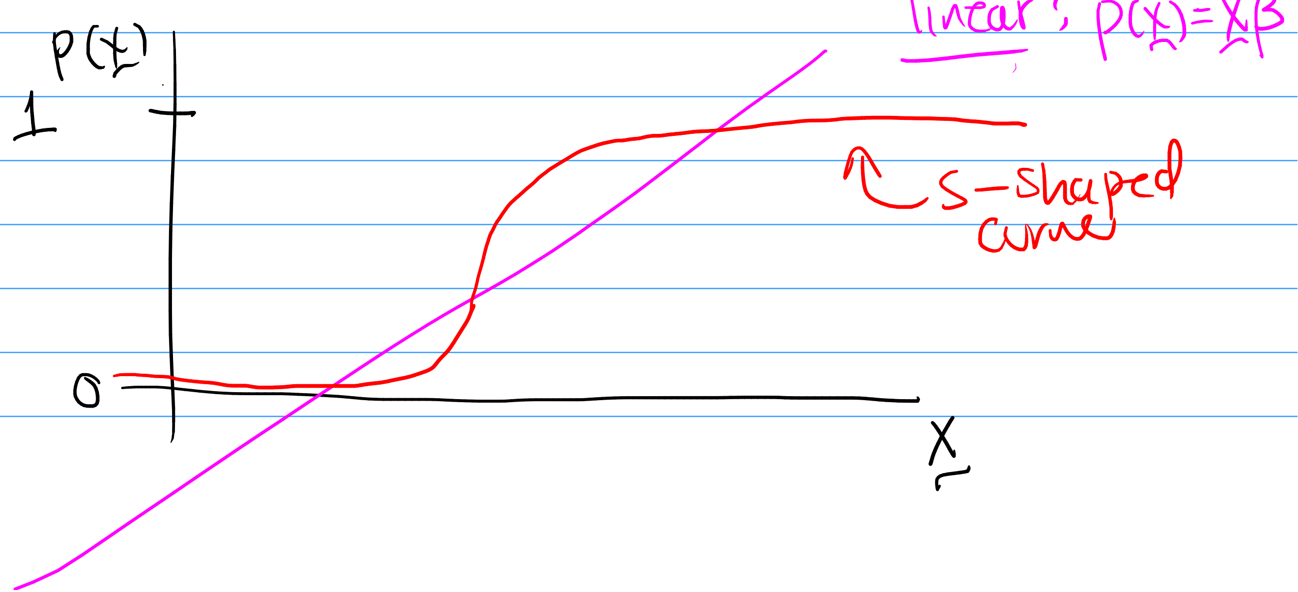
$$\hat{f}(\underset{\sim}{x}) = 1 \quad \text{when} \quad \delta_1(\underset{\sim}{x}) > 1/2$$

Traditionally $p(\underset{\sim}{x}) = \delta_1(\underset{\sim}{x}) = \mathbb{P}(Y=1 \mid \underset{\sim}{X} = \underset{\sim}{x})$

Given $\underset{\sim}{X} = \underset{\sim}{x}$, $Y=0$ or $Y=1$

$$Y \mid \underset{\sim}{X} = \underset{\sim}{x} \sim \text{Bern}(p(\underset{\sim}{x}))$$

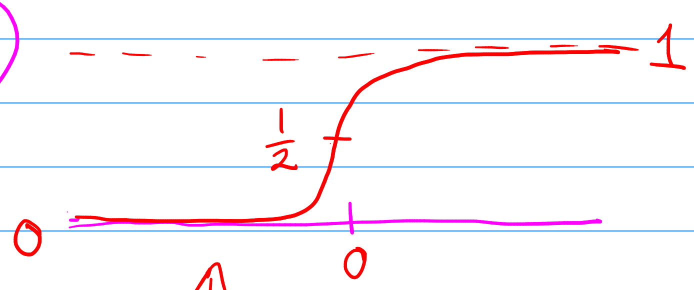
Game: reasonable model for $p(\underset{\sim}{x})$



Logistic regression says:

$$p(\underset{\sim}{x}) = \text{logistic}(\underset{\sim}{x}^T \hat{\beta})$$

$$\text{logistic}(x) = 1/(1+e^{-x})$$



logistic fn

Inverse:

$$\text{logit}(x) = \text{logistic}^{-1}(x)$$
$$= \log\left(x/(1-x)\right)$$

$$p(\underset{\sim}{x}) = \text{logistic}(\underset{\sim}{x}^T\hat{\beta})$$
$$= \frac{1}{1+\exp(-\underset{\sim}{x}^T\hat{\beta})}$$
$$= \frac{1}{1+\exp\left(-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_p x_p)\right)}$$

Notice: $\delta_1(\underset{\sim}{x}) = p(\underset{\sim}{x}) = \text{logistic}(\underset{\sim}{x}^T\hat{\beta})$

so

$$\text{logit}(\delta_1(\underset{\sim}{x})) = \underset{\sim}{x}^T\hat{\beta}$$

linear

a linear classifier

How do we get $\hat{\beta}$?

Training data: $Y_n \mid \underset{\sim}{X}_n = \underset{\sim}{x}_n \overset{\text{indep}}{\sim} \text{Bern}(p_\beta(\underset{\sim}{x}_n))$

$$p_\beta(\underset{\sim}{x}) = \frac{1}{1+\exp(-\underset{\sim}{x}^T\beta)}$$

Maximum Likelihood Ests (MLEs)

parameter θ in a model $\mathbb{P}_\theta(\underset{\sim}{x})$

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \; \mathbb{P}_\theta(\underset{\sim}{x})$$

Logistic regression gets $\hat{\beta}$ as MLE of β under this model.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \; \underbrace{\mathbb{P}_\beta(Y_1, Y_2, Y_2, \dots, Y_N \mid \underset{\sim}{X}_1, \underset{\sim}{X}_2, \dots, \underset{\sim}{X}_N)}_{L(\beta)}$$

$$L(\beta) = \prod_{n=1}^{N} \mathbb{P}_\beta(Y_n \mid \underset{\sim}{X}_n = \underset{\sim}{x}_n)$$

$$= \prod_{n=1}^{N} p_\beta(\underset{\sim}{x}_n)^{y_n} (1 - p_\beta(\underset{\sim}{x}_n))^{1-y_n}$$

$$= \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\underset{\sim}{x}_n^T\beta)}\right)^{y_n} \left(1 - \frac{1}{1+\exp(-\underset{\sim}{x}_n^T\beta)}\right)^{1-y_n}$$

Bernoulli (p)

$p(x) = p^x (1-p)^{1-x}$

for $x = 0$ or 1

No closed form soln, need numerical opt to get $\hat{\beta}$.

Multi-nomial Logistic Regression

When $K > 2$

K-1 probs

$$Y_n \mid \underset{\sim}{X}_n = \underset{\sim}{x}_n \overset{\text{indep}}{\sim} \text{Categorial}(p_1(\underset{\sim}{x}_n), \ldots, p_{K-1}(\underset{\sim}{x}_n))$$

$$\delta_k(\underset{\sim}{x}) = \mathbb{P}(Y = k \mid \underset{\sim}{X} = \underset{\sim}{x}) = p_k(\underset{\sim}{x})$$

$$= \text{multi-variate logistic}(\underset{\sim}{x}^T \hat{\beta}_k)$$

soft-max function

$$= \frac{\exp(\hat{\beta}_k^T \underset{\sim}{x})}{1 + \sum_{k=1}^{K-1} \exp(\hat{\beta}_k^T \underset{\sim}{x})}$$

Similarly, fit each/all $\hat{\beta}_k$s by MLE.

LDA v. Logistic Regression

LDA	Logistic Regression
1. models X|Y and Y using normality assumption on X	1. models Y|X no model of X
2. easier to fit	2. harder to fit

3. both linear