Lecture 11 - Logistic Regression

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

$$= \frac{\mathbb{P}(X=x \mid Y=c)\,\mathbb{P}(Y=c)}{\mathbb{P}(X=x)}$$

$$\propto \underbrace{\mathbb{P}(X=x \mid Y=c)}_{N(\mu_c,\sigma^2)}\underbrace{\mathbb{P}(Y=c)}_{\pi_c}$$

Logistic Regression

$$\delta_c(x) = \mathbb{P}(Y=c \mid X=x)$$

model this directly.

Binary Logistic Regression $(K=2)$

So here $Y=0$ or $Y=1$

discr. fns

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

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

So in the binary case only need one discr function (we can always get the other)

Call $p(x) = \delta_1(x) = \mathbb{P}(Y=1|X=x)$

<u>Rule:</u> classify as <u>class 1</u> if $p(x) > 1/2$

since iff $p(x) = \delta_1(x) > 1/2$ since δ_1 and δ_0 sum to 1 then $\delta_0(x) < 1/2$,

hence $\delta_1(x) > \delta_0(x)$

Given $X=x$ notice that $Y=0$ or $Y=1$.

We call $Y|X=x$ a Bernoulli RV.

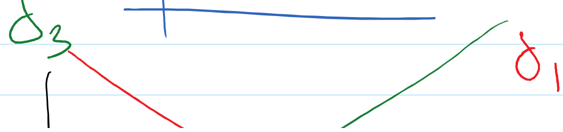
We could write our setup simply as

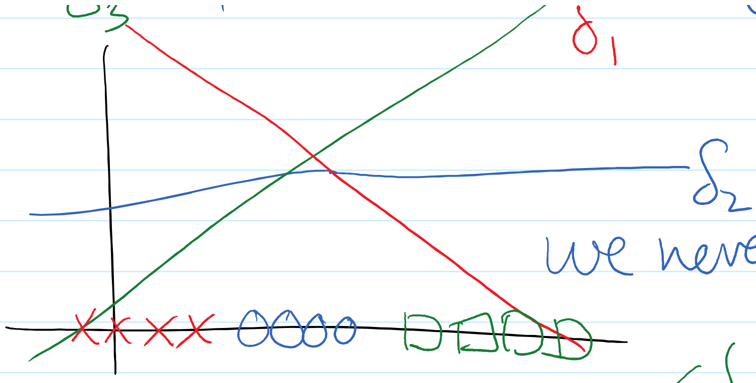
$$Y|X=x \sim \text{Bernoulli}(p(x))$$

<u>Game</u>: How do we model $p(x)$?

<u>Way 1</u>: $p(x) = \hat{\beta}^T x$ could learn $\hat{\beta}$s from regression.

<u>problems:</u> masking



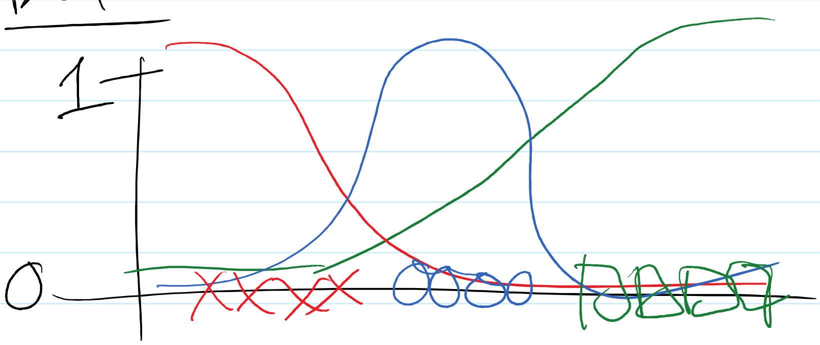


we never predict class O

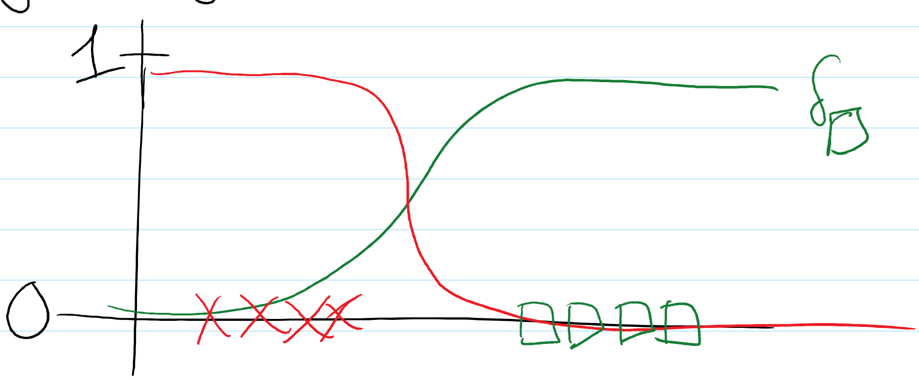
$p(x)>1$ here

$p(x)<0$ here

Better



Way 2: Logistic Regression (Binary)



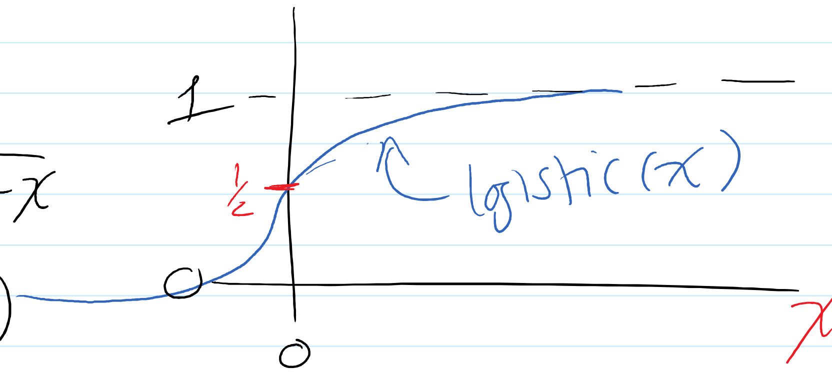
logistic regression says let

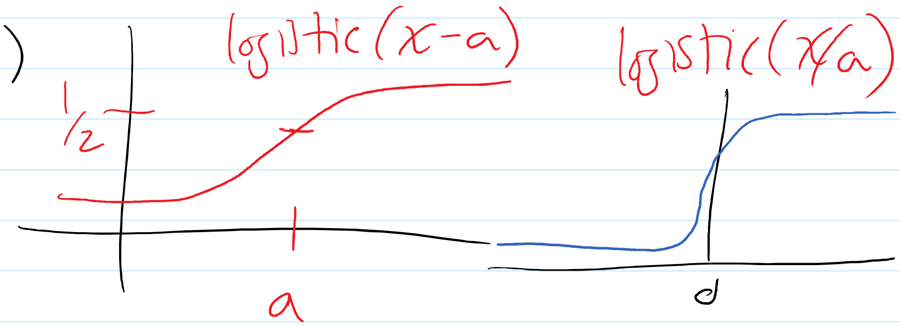
$$\boxed{p(x) = \text{logistic}(\beta^T x)}$$

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

$$\text{logit}(y) = \log\left(\frac{y}{1-y}\right)$$

$$= \text{logistic}^{-1}(y)$$





---

logistic regression (in Binary case)

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

$$\text{and } p(x) = 1/(1+e^{-\beta^T x})$$

How do we learn $\hat{\beta}$? Maximum Likelihood Estimation

$$y_n \mid x_n \overset{indep}{\sim} \text{Bernoulli}(p_\beta(x))$$

$$p_\beta(x) = \text{logistic}(\beta^T x).$$

Joint density of y_ns conditioned on x_ns

likelihood

$$L(\beta) = P(y_1, y_2, y_3, \ldots, y_N \mid x_1, x_2, x_3, \ldots, x_N)$$
$$= \prod_{n=1}^{N} P(y_n \mid x_n)$$
$$= \prod_{n=1}^{N} p_\beta(x)^y (1 - p_\beta(x))^{1-y}$$
$$= \prod_{n=1}^{N} \left(\frac{1}{1+e^{-\beta^T x_n}}\right)^{y_n} \left(1 - \frac{1}{1+e^{-\beta^T x_n}}\right)^{1-y_n}$$

$y \sim \text{Benoulli}(p)$

$$f(x) = p^y (1-p)^{1-y} = \begin{cases} p, & 1 \\ 1-p, & 0 \end{cases}$$

<u>MLE:</u>

$$\hat{\beta} = \arg\max_\beta L(\beta)$$

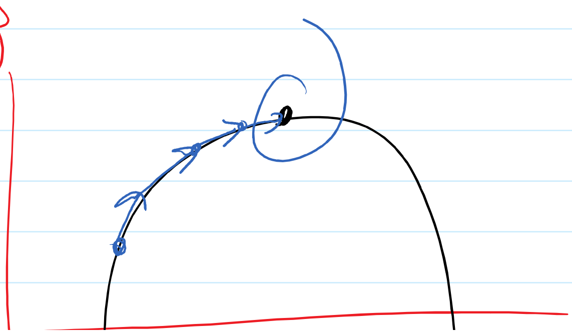
→ No analytical solution — give to friends in OR.

→ Solve w/ gradient descent (iterative)

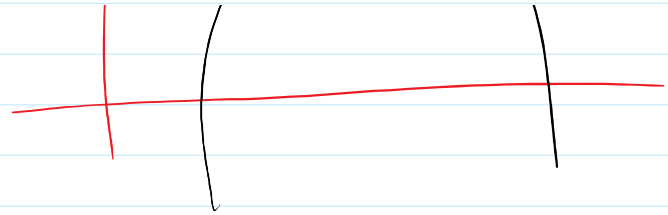
Can cast GD as a series of (weighted) regression.

(IRLS)

(iteratively



(iteratively reweighted least-squares



Multinomial Logistic Regression

Multi-Class $(K > 2)$

generalization of Binomial when I have K discrete outcomes

$$y_n \mid x_n \overset{\text{indep}}{\sim} \text{Multinomial}\big(\underbrace{p_1(x), p_2(x), \ldots, p_{K-1}(x)}_{\text{Sum to 1 so only need } K-1}\big)$$

For $k = 1, \ldots, K-1$

$$\delta_k(x) = p_k(x) = P(Y = k \mid X = x)$$

$$= \text{MVLogistic}_k(\ldots\ldots)$$

$$= \frac{e^{\beta_k^T x}}{1 + \sum_{\ell=1}^{K} e^{\beta_\ell^T x}}$$

β for each class.

Now we have $\beta_1, \beta_2, \ldots, \beta_K$

and we estimate these as MLEs

$$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_K = \underset{\beta_1 \cdots \beta_K}{\arg\max}\, L(\beta_1, \ldots, \beta_K)$$

Back to Binary $(K = 2)$

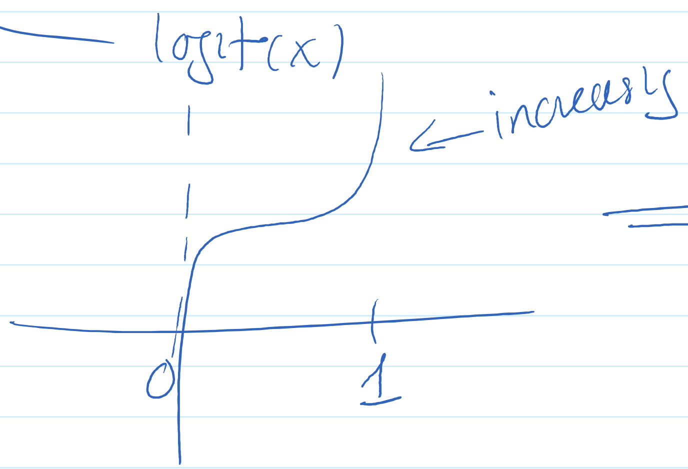
$$\delta_1(x) = p(x) = \frac{1}{\ldots} = \text{logistic}(\beta^T x)$$

$1+e^{-\beta^T x}$

$$\operatorname{logit} \delta_1(x) = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta^T x$$

$$1-p(x) = \frac{1+e^{-\beta^T x} - 1}{1+e^{-\beta^T x}} = \frac{e^{-\beta^T x}}{1+e^{-\beta^T x}}$$

$$p(x) / 1-p(x) = \frac{1+e^{-\beta^T x}}{e^{-\beta^T x}} \cdot \frac{1}{1+e^{-\beta^T x}} = e^{\beta^T x}$$



$\Rightarrow$ increasing transf. of my δ_1 is linear

$\Rightarrow$ logistic Regression is a linear classifier.

LDA v. Logistic Regression

LDA	Logistic Reg
① models $P(X\mid Y)$ and $P(Y)$ using a normality ...	① models $Y\mid X$ directly — no normality assumption about X (more general)

Models $\pi(x|Y)$ and $\pi(Y)$ using a normality assumption about $X|Y$
(X continuous)

② normality is make estimates easier to get
(more efficient if Xs truly normal)

no normality assumption about X (more general)
(can have categorical covariates)

② estimating βs is comp. more expensive.

$$\boxed{\text{Both } \delta_1(x) = \hat{\beta}^T x}$$