STAT 391 3/4/25

Lecture 17

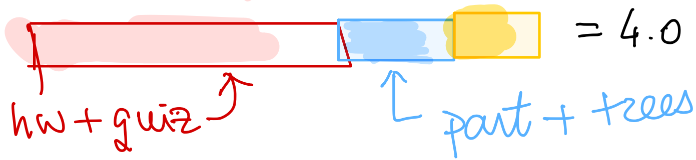
l8-regression.pdf

HW7 TB posted
optional

Logistic Regression + Classification

$$\text{score for HW + Quiz} = \max\begin{cases} \dfrac{\text{total pts hw + best quiz}}{\text{total possible from hw + best q.}} \\[2ex] \dfrac{\text{pts}(\text{hw} \setminus \text{worst}) + \text{points } q1 + q2}{\text{total possible}} \end{cases}$$

. current participation TB posted



= 4.0

Logistic Regression

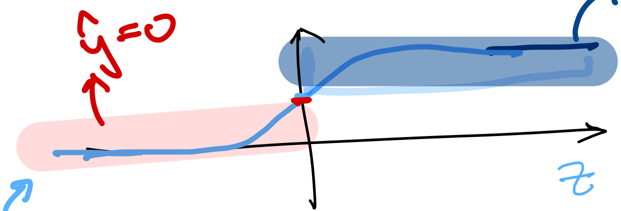
Model $\left\{ f(x) = \dfrac{e^{\beta^T x}}{1+e^{\beta^T x}} \right\} = \mathcal{F}$

$x \in \mathbb{R}^d$
$\beta \in \mathbb{R}^d$ parameter

$\begin{bmatrix} \boxed{\beta_0} \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$ — Intercept

$x \leftarrow \begin{bmatrix} \boxed{1} \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$

Prediction $\hat{y} = \begin{cases} 1 & f(x) \geq 0.5 \\ 0 & f(x) < 0.5 \end{cases}$



$\dfrac{e^z}{1+e^z}$ = logistic fction

$f(x) \in (0,1)$

$\min\{f(x), 1-f(x)\} \leq 0.5$

confidence in $\hat{y}$

Classifying with logistic regression

x new input observation (with features) : $f(x)$ = model belief about $y(x)$

> 0.5 predict $\hat{y} = 1$

< 0.5 —"— $\hat{y} = 0$

Training logistic r. (= estimating β by ML)

Given $\mathcal{F}$, $\mathcal{D} = \{(x^1, y^1), \dots (x^n, y^n)\}$ data $\Longrightarrow$ wanted β^{ML}

Estimating β by Max Likelihood

Likelihood $L(\beta) = \prod_{i=1}^{n} \underbrace{P[y^i \mid \beta, x^i]}_{\substack{P[y=1 \mid x^i, \beta] \text{ if } y^i=1 \\ P[y=0 \mid x^i, \beta] \text{ if } y^i=0}} = \prod_{i=1}^{n} \underbrace{f(x^i)^{y^i} (1-f(x^i))^{1-y^i}}_{\substack{p \text{ if } y^i=1 \\ 1-p \text{ if } y^i=0}}$

$$= \prod_{i=1}^{n} \frac{e^{\beta^T x^i \cdot y^i}}{1+e^{\beta^T x^i}}$$

$e^{\beta^T x^i}$ $y^i=1$
1 $y^i=0$

$P[y \mid \beta, x] = ?$

$$g(x) = \beta^T x = \ln \frac{P[y=1 \mid x, \beta]}{\underbrace{P[y=0 \mid x, \beta]}_{\to \text{odds}}}$$

log · odds

p_x $\quad 1-p_x$

$$g = \ln \frac{p}{1-p} \longrightarrow \text{solve for } p$$

$$e^g = \frac{p}{1-p} \Rightarrow \boxed{p = \frac{e^g}{1+e^g}} = f(x)$$

is exactly model's $P[y=1 \mid x, \beta]$!!

$$\boxed{1-p = \frac{1}{1+e^g}}$$

1. log-likelihood

$$\ell(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \left[y^i \cdot \beta^T x^i - \ln\left(1 + e^{\beta^T x^i}\right) \right]$$

2. $\max_{\beta} \ell(\beta)$

STAT

~~CALCULUS~~

GRADIENT ASCENT

The gradient of $\ell(\beta)$

$$\ell(\beta) = \sum_{i=1}^{n} \left[y^i \cdot \beta^T x^i - \ln\left(1 + e^{\beta^T x^i}\right) \right]$$

$z^i = \beta^T x^i \in \mathbb{R}$

- $\left[y \cdot z - \ln(1 + e^z) \right]' = y - \underbrace{\frac{e^z}{1+e^z}}_{f} = y - f(x) = \begin{cases} 1 - \Pr[y^i = 1 \mid x^i, \beta] & y = 1 \\ 0 - \Pr[y^i = 1 \mid x^i, \beta] = \Pr[y = 0 \mid x, \beta] - 1 & \text{for } y = 0 \end{cases}$

$= -(1 - \Pr[y = 0 \mid x, \beta])$

- $\frac{\partial}{\partial \beta}(\beta^T x) = x \in \mathbb{R}^d$

$\beta, x \in \mathbb{R}^d$

$$\frac{\partial}{\partial \beta} \ell = \sum_{i=1}^{n} \left[\quad z^i \quad \right]' \frac{\partial}{\partial \beta}(\beta^T x^i) = \sum_{i=1}^{n} \overbrace{\left(\underbrace{y^i - f_\beta(x^i)}_{\text{scalar}} \right)}^{\text{small for high confidence (and correct!)}} \cdot \underbrace{x^i}_{\text{vector}} \in \mathbb{R}^d$$

+ for $y^i = 1$
− for $y^i = 0$

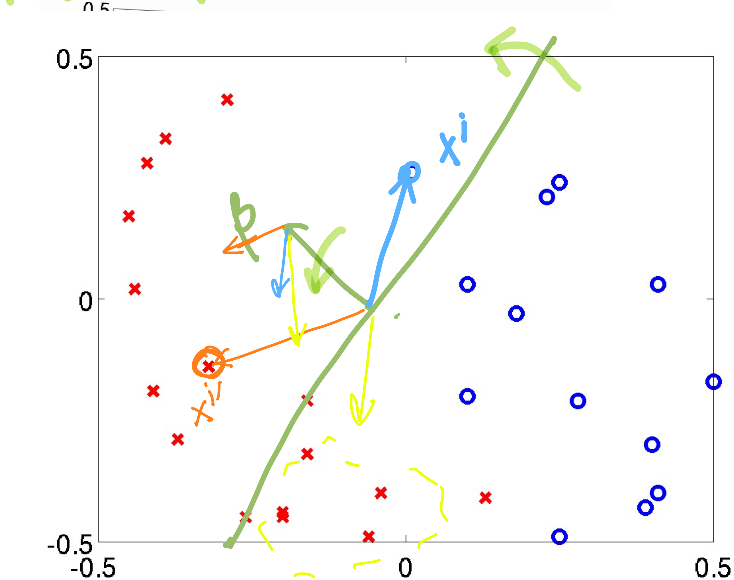
$d = 2$

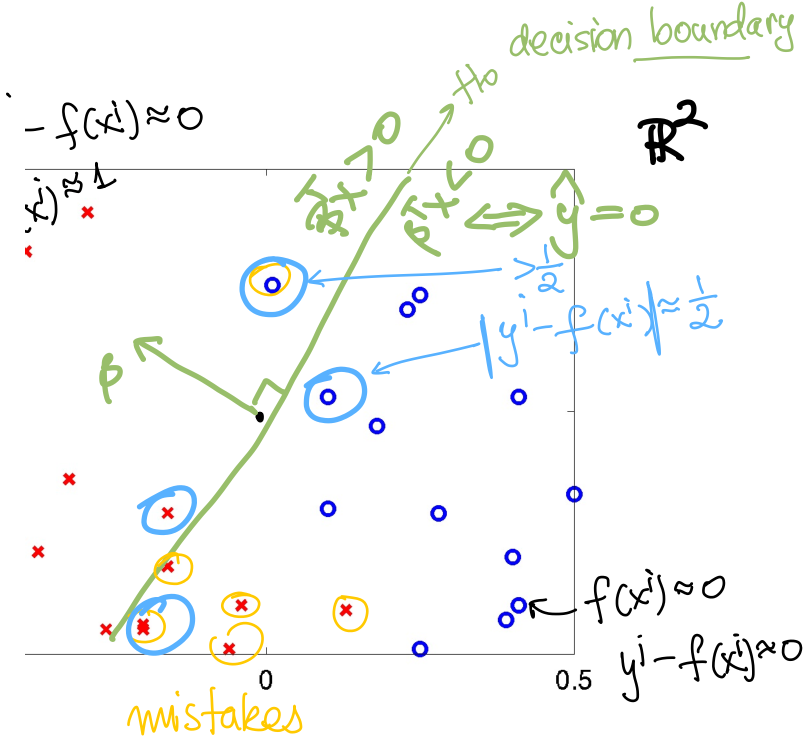
Assume $\beta = 0$ at initialization

GRAD. STEP $\quad \beta^t \leftarrow \beta^{t-1} + \underbrace{\eta}_{\text{small stepsize}} \cdot \frac{\partial \ell}{\partial \beta}(\beta^{t-1})$

$\mathcal{H}_0 = \{x \mid \beta^T x = 0\}$ decision boundary $\longleftrightarrow$ $\beta^T x \gtrless 0 \iff f(x) \gtrless 0.5$

rotate $\beta \iff$ rotate $\mathcal{H}_0$





H_0 = hyperplane in $\mathbb{R}^d$ $\iff$ $f(x)$ = LINEAR CLASSIFIER ← Because "Linear" decision boundary

NOT linear function of x !!

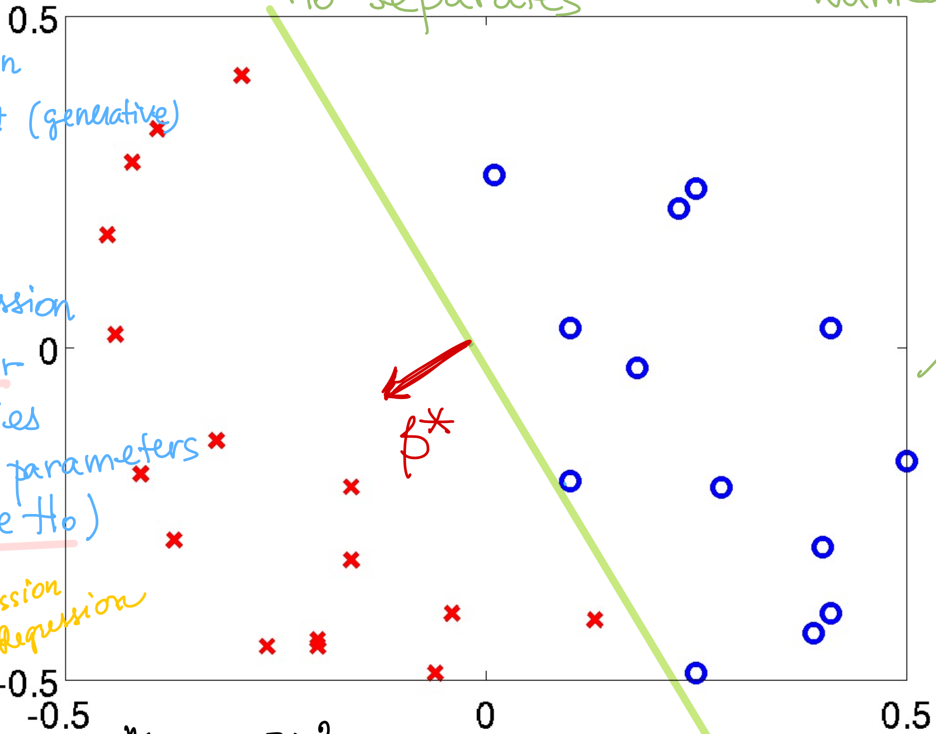
H_0 separates $\{y=0\}$ from $\{y=1\}$ wanted

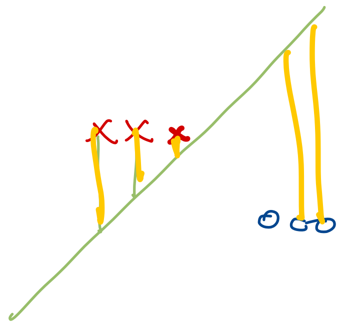
Linear classifiers

- logistic regression
- Fisher discriminant (generative)
- SVM
- perceptron
- ...
- Linear Regression

↳ All have linear Decision boundaries

BUT: different β parameters ($\Rightarrow$ not the same H_0)





β^* at convergence

outlier

x Big influence on Linear Regression

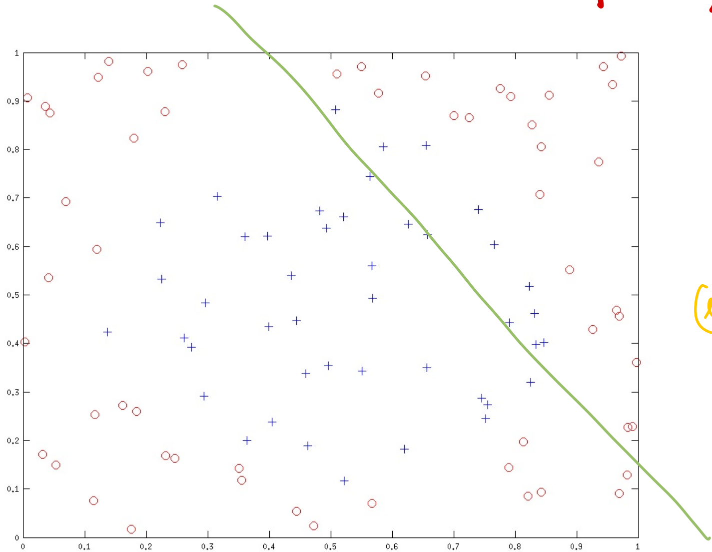
x no influence on Logistic Regression

min LSquares $\iff$ min $\sum_{i=1}^{n} \underbrace{(y^i - \beta^T x^i)}_{\text{res}}^2$, $y^i \in \{0,1\}$

H_0 wanted

$$z = \beta_0 + \beta_1 x$$

$$f = \frac{e^z}{1+e^z}$$



(underfitting!)

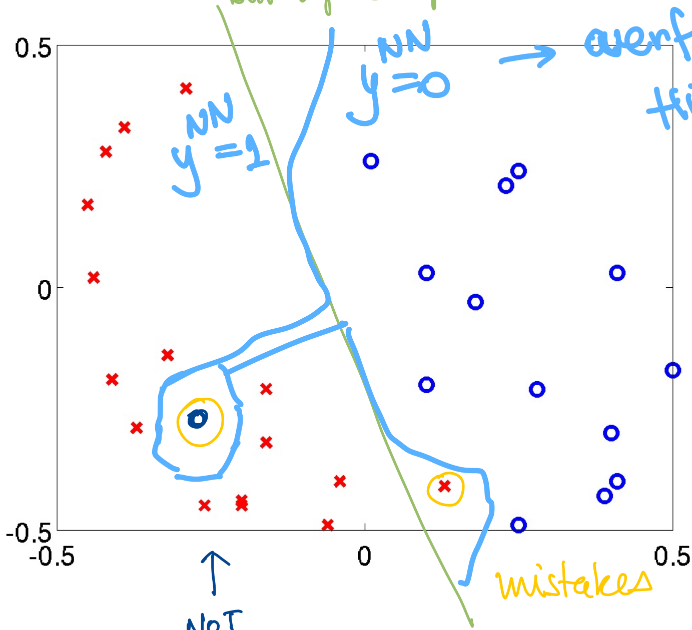
Best? Logistic regression

Best logistic regression → Bias > 0
Low Var

NN = nearest neighbor

$y^{NN} = 1$

$y^{NN} = 0$ → overfit
high var
0 bias !!



mistakes

NOT
linearly
separable !!

which? CV = answer