Lecture 9 — LS, Classification, Ortho Bases, Regularization

LS & Classification

Given training $\{\underline{x}_i, y_i\}_{i=1}^n$, $\underline{x}_i \in \mathbb{R}^p$ (e.g. $x_i = \begin{bmatrix} \text{left mouth to brow} \\ \text{right " " " } \\ \text{mouth width} \end{bmatrix}$)

$y_i \in \mathbb{R}$ (degree of happiness)

or

$y_i \in \{-1, 1\}$ (happy or say)

Linear prediction

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip}$$

$$= \langle \underline{w}, \underline{x}_i \rangle = \underline{w}^T \underline{x}_i$$

(A) Build a "data matrix" or "feature matrix" and label vector

$$X = \begin{bmatrix} — \ \underline{x}_1^T \ — \\ \vdots \\ — \ \underline{x}_n^T \ — \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad \underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Linear model $\hat{\underline{y}} = X\underline{w}$

(B) Solve least squares opt:

$$\hat{\underline{w}} = \underset{\underline{w}}{\operatorname{argmin}} \ \|\underline{y} - X\underline{w}\|_2^2 \longrightarrow \sum_{i=1}^{n} (y_i - \underline{x}_i^T \underline{w})^2$$

if cols of X are linearly independent $\Rightarrow$ X^TX is positive definite

$\Rightarrow \underline{a}^T X^T X \underline{a} > 0 \ \forall \ \underline{a} \neq 0 \Rightarrow$ X^TX is invertible

e.g. $X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \rightarrow \underline{a}^T X^T X \underline{a}$ w/ $\underline{a} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \Rightarrow X\underline{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \underline{a}^T X^T X \underline{a} = 0$

$\Rightarrow X^TX$ is NOT pos. def.

if $X^T X$ is pos. def. then $\exists$ unique LS solution

$$\hat{w} = (X^T X)^{-1} X^T \underline{y}$$

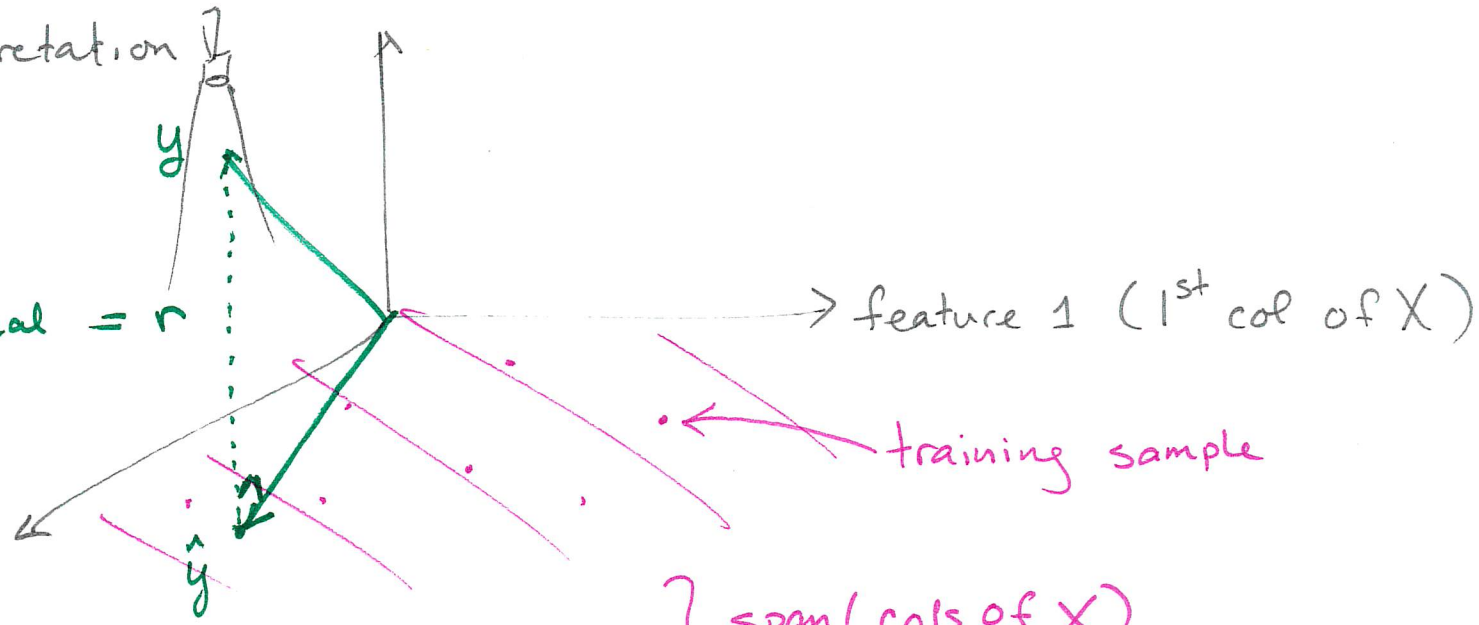
(c) Validate w/ test / hold out data

$$\{ \tilde{\underline{x}}_i, \tilde{y}_i \}_{i=1}^{n_T}$$

$$\sum_{i=1}^{n_T} (\tilde{y}_i - \tilde{\underline{x}}_i^T \hat{w})^2 = \text{error on test data}$$

Geometric Interpretation

$p = 2$



$$\hat{\underline{w}} = (X^T X)^{-1} X^T \underline{y}$$

$$\hat{\underline{y}} = X \hat{\underline{w}}$$

$$= \underbrace{X (X^T X)^{-1} X^T}_{\text{projection matrix } P_x} \underline{y}$$

Binary classification

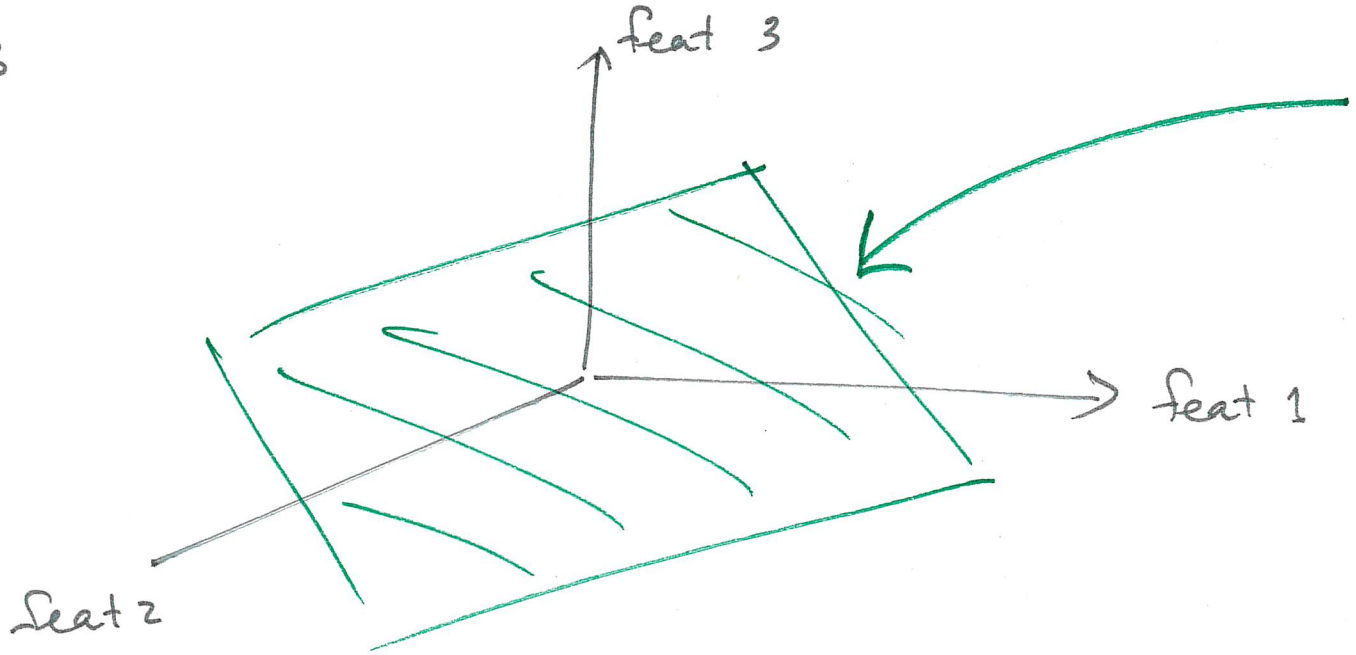
$$y_i = \begin{cases} +1 & \text{class 1} \\ -1 & \text{class 2} \end{cases}$$

Classification rule: $\tilde{y}_i = \begin{cases} +1 & \text{if } \hat{y}_i \text{ is closer to } +1 \text{ than } -1 \\ -1 & \text{otherwise} \end{cases}$

$$= \text{sign}(\underline{x}_i^T \hat{\underline{w}})$$

Decision Boundary in Feature Space

$p = 3$



$\{ \underline{x} : \underline{x}^T \hat{\underline{w}} = 0 \}$

What sort of set is $B = \{ x : \underline{x}^T \hat{\underline{w}} = 0 \}$?

① $\underline{x} = 0 \Rightarrow \underline{x} \in B$ $\quad (\underline{0} \in B)$

② if $\underline{x}_1$ and $\underline{x}_2 \in B$, then $\underline{x}_1 + \underline{x}_2 \in B$

$x_1^T w = 0 = x_2^T w$ $\quad (x_1 + x_2)^T w = x_1^T w + x_2^T w = 0 + 0 = 0$

③ if $\underline{x} \in B$, $\alpha \in \mathbb{R}$, $\alpha \underline{x} \in B$

$x^T \hat{w} = 0 \Rightarrow \alpha \underline{x}^T \hat{\underline{w}} = 0$

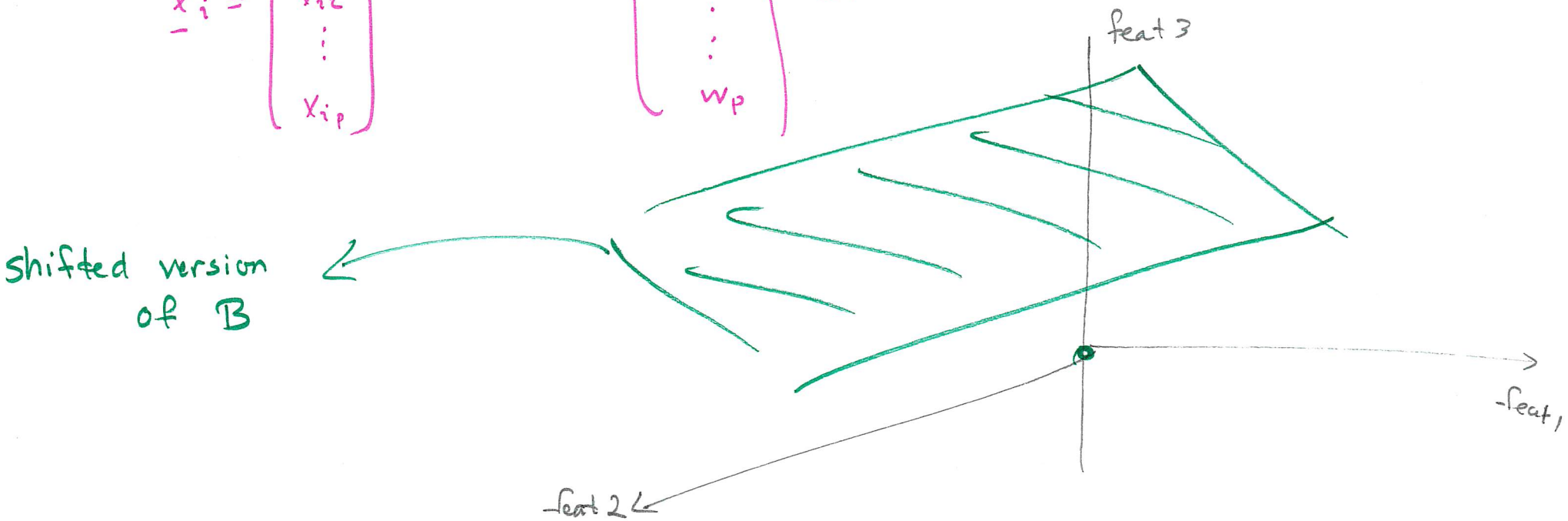
B is a subspace

What if we include an offset?

$$\hat{y}_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip} + w_0$$

$$= \underline{w}^T \underline{x}_i + w_0 = \underset{\sim}{w}^T \underset{\sim}{x}_i$$

$$\underset{\sim}{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \qquad \underset{\sim}{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p+1}$$



$$X = \begin{bmatrix} 98 & 102 & 300 \\ 102 & 98 & 300 \\ 96 & 104 & 100 \\ 104 & 96 & 100 \end{bmatrix}, \quad \underline{y} = \begin{bmatrix} 350 \\ 300 \\ 250 \\ 250 \end{bmatrix}$$

↑ left eye to lip dist. ↑ right eye to lip dist. ↑ mouth width

cols of X are linearly independent $\Rightarrow X^T X$ invertible

$$\hat{\underline{w}} = (X^T X)^{-1} X^T \underline{y} = \begin{bmatrix} 1 \\ 1 \\ 1/2 \end{bmatrix}$$

U is an orthonormal basis matrix $\Rightarrow U^T U = I$

$$U = \begin{bmatrix} 1/2 & -1/\sqrt{10} & 1/2 \\ 1/2 & 1/\sqrt{10} & 1/2 \\ 1/2 & -2/\sqrt{10} & -1/2 \\ 1/2 & 2/\sqrt{10} & -1/2 \end{bmatrix}, \quad \hat{\underline{v}} = (U^T U)^{-1} U^T \underline{y}$$

$$= U^T \underline{y} = \begin{bmatrix} 600 \\ 0 \\ 100 \end{bmatrix}$$

$$U = XA, \quad A = \begin{bmatrix} 1/400 & 1/4\sqrt{10} & -1/200 \\ 1/400 & -1/4\sqrt{10} & -1/200 \\ 0 & 0 & 1/200 \end{bmatrix}$$

$\underline{u}_1 = \dfrac{\underline{x}_1 + \underline{x}_2}{400} \approx$ avg of left + right eye to lip distance

$\underline{u}_2 = \dfrac{\underline{x}_1 - \underline{x}_2}{4\sqrt{10}} \approx$ diff between left & right

$\underline{u}_3 \approx$ mouth width − avg of L + R eye to lip dist

$= \dfrac{x_3 - (x_1 + x_2)}{200}$

w/ X, get weight vect. $X\hat{w}$ good predictions

suggests that left and right features equally important

w/ U, clear that avg. of L+R important

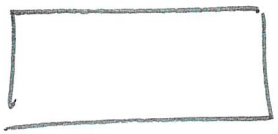
diff between L+R unimportant.

but u_3 less interpretable.

Q: how should we represent data/features?

Regularization

LS: $\hat{\underline{w}} = \underset{\underline{w}}{\operatorname{argmin}} \|\underline{y} - X\underline{w}\|_2^2$

what if $X =$  (short & fat)

$n < p$

fewer equations than unknowns.

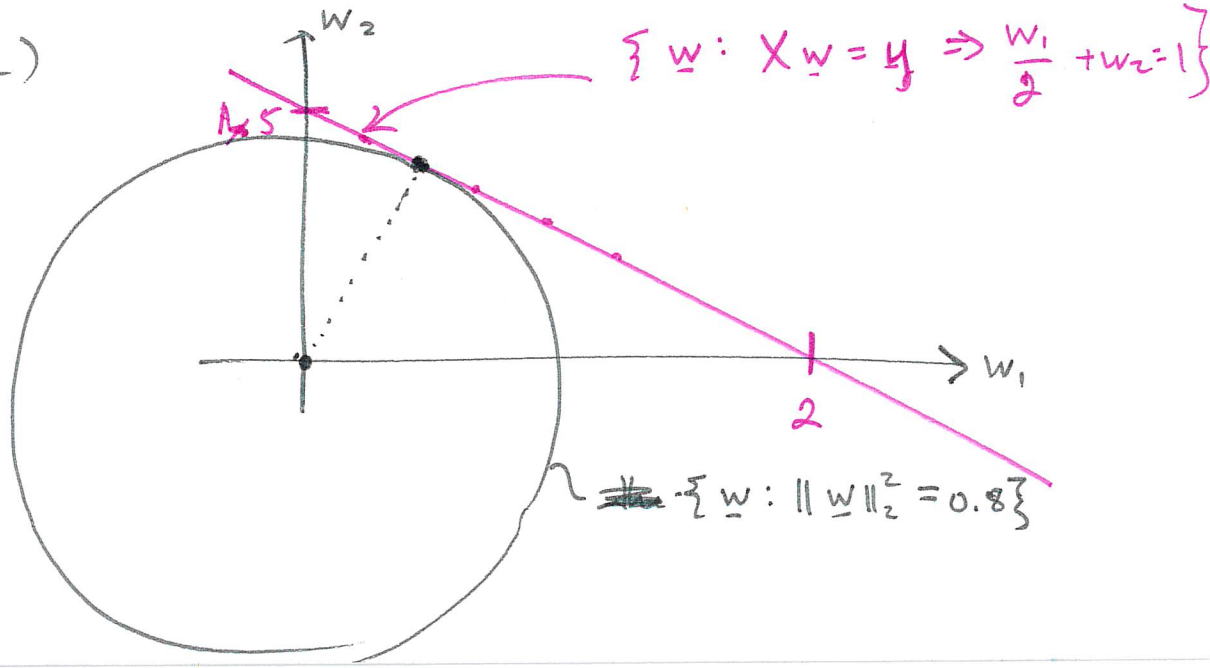
$X^T X$ is not invertible

one approach: find $\underline{w}$ that has small norm $\|w\|_2^2$

$X = [\,\tfrac{1}{2} \quad 1\,]$ $(n=1, p=2)$

$y = 1$

$\frac{w_1}{2} + w_2 = 1$



$$\underset{\underline{w}}{\text{minimize}} \ \|\underline{w}\|_2^2 \quad \text{subject to} \quad y = X\underline{w}$$

alt: find $\underline{w}$ so that $\|y - X\underline{w}\|_2^2$ is small and $\|\underline{w}\|_2^2$ small

$$\hat{\underline{w}} = \underset{\underline{w}}{\text{argmin}} \ \|y - X\underline{w}\|_2^2 + \lambda \|\underline{w}\|_2^2 \quad \text{for} \quad \lambda > 0$$

λ tells us relative importance of two terms.

$\lambda \|\underline{w}\|_2^2$ = "regularizer"

~~Tikhonov~~ Tikhinov regularization

Ridge regression

$$\|y - Xw\|_2^2 + \lambda \|w\|_2^2 = \left\| \begin{bmatrix} y - Xw \\ w \end{bmatrix} \right\|_2^2$$

$$= \left\| \underbrace{\begin{bmatrix} y \\ 0 \end{bmatrix}}_{\tilde{y}} - \underbrace{\begin{bmatrix} Xw \\ \sqrt{\lambda}\, w \end{bmatrix}}_{\begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} w} \right\|_2^2 = \|\tilde{y} - \tilde{X} w\|_2^2$$

$$\hat{w} = \underset{w}{\operatorname{argmin}} \|\tilde{y} - \tilde{X} w\|_2^2 = \underbrace{(\tilde{X}^T \tilde{X})^{-1}}\, \tilde{X}^T \tilde{y}$$

always invertible!

$$\left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T \tilde{y} =$$

$$= \left(X^T X + \lambda I\right)^{-1} X^T y$$

$$\tilde{X}^T \tilde{X} = \begin{bmatrix} X^T & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}$$

$$= X^T X + \lambda I$$