CPSC 536F Nov 21, 2025

- Last time: $A \in \mathbb{R}^{n \times n}$ symmetric

$$A = \Lambda = \begin{bmatrix} \lambda_1 & & 0's \\ & \ddots & \\ 0's & & \lambda_n \end{bmatrix} \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$$

Then $\forall\ V \subset \mathbb{R}^n$, $\dim(V) = k$

$\exists\ \vec{w} \in V$ s.t. $\mathcal{R}_A(\vec{w}) \leq \lambda_k$.

- Today:

  - Finish max-min, min-max

  - SVD

  - "Dual" of ridge regression

Thm: Say our model has data

$\mathcal{D}: (\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)$

$\vec{x}_i \in \mathbb{R}^d, \quad y_i \in \mathbb{R}$

$\left(\vec{x}_i \in \mathbb{R}^d \text{ used to be } \vec{f}_i \in \mathbb{R}^d \right)$

Let

$\lambda = C > 0$

$$\mathcal{E}_{\mathcal{D}}(\vec{w}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \left(y_i - \vec{x}_i \cdot \vec{w} \right)^2 + \lambda \|\vec{w}\|^2$$

$$= \| X\vec{w} - \vec{y} \|_{L^2}^2 + \lambda \|\vec{w}\|^2$$

Hence $(X^T X + \lambda I_d)\vec{w}^* = X^T \vec{y}$

Then ("dually") for

$$(X^T X + \lambda I_n)\,\alpha^* = \vec{y}$$

we have

$$\vec{w}^* = X^T \alpha^* .$$

dim d d×n n dim

n, d is large

Rem: If $\vec{V} \subset \mathbb{R}^n$, $\dim(\vec{V}) = k$, then $\exists\ \vec{w} \in \vec{V}$

s.t.

$$\vec{w} = \begin{array}{cl} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ w_k \\ \vdots \\ w_n \end{bmatrix} & \begin{array}{l} \left.\vphantom{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}\right\} k-1 \\ \vphantom{\begin{matrix} w_k \\ \vdots \\ w_n \end{matrix}} \end{array} \end{array}$$

$$\Longrightarrow \quad R_A(\vec{w}) \leq \lambda_k$$

$$\begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

If $\vec{V} = \text{span}\langle \vec{e}_1, \vec{e}_2, \ldots, \vec{e}_k \rangle$

so

$$\vec{V} = \left\{ \begin{array}{c|c} \begin{bmatrix} \left.\begin{matrix} * \\ * \\ \vdots \\ * \end{matrix}\right\} k \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} & * = \text{any } \mathbb{R} \end{array} \right\}$$

Then $\vec{w} \in \vec{V}$

$R_A(\vec{w})$

nonneg weights add to 1

$$= \lambda_1 (\quad) + \lambda_2 (\quad) + \ldots + \lambda_k (\quad) \geq \lambda_k$$

$\dim(\bar{V}) = k$:

$$\min_{\vec{w} \in \bar{V}} R_A(\vec{w}) \le \lambda_k$$

and there is $\tilde{V}$, $\dim(\tilde{V}) = k$ s.t.

$$\min_{\vec{w} \in \tilde{V}} R_A(\vec{w}) = \lambda_k .$$

So

$$\max_{\substack{\bar{V} \subset \mathbb{R}^n \\ \dim(\bar{V}) = k}} \left(\min_{\vec{w} \in \bar{V}} R_A(\vec{w}) \right) = \lambda_k$$

and ---

min $\qquad$ max $\qquad = \lambda_{n-k}$

SVD (Singular Value Decomposition):

generalizes the spectral theorem for symmetric matrices:

$A \in \mathbb{R}^{m \times n}$

diagonal, but $m \times n$

$A = U \Sigma V^T$

orthogonal

V is orthogonal

$n \times n$

$$U \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & 0 \\ & & \ddots & & \\ & & & \sigma_m & \end{bmatrix} V^T$$

$m \times m$

$(m < n)$

$m \times n$

Proof: Imagine this is true ...

$$AA^T \overset{\text{hopefully}}{=} (U\Sigma V^T)(V\Sigma^T U^T)$$

$$= U\Sigma \underbrace{V^T V}_{I} \Sigma^T U^T$$

$$= U(\Sigma\Sigma^T)U^T$$

"Reverse engineer":

$AA^T \in \mathbb{R}^{m\times m}$ is symmetric

$$(AA^T)^T = (A^T)^T (A)^T = AA^T$$

must $=$ UDU^T, some diag D, $n \times n$

One application:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{m\cancel{ax}in(m,n)}$$

Gives a "low rank approximation" of A

Idea of typical application:

A = pixels in a picture of

clown,
astronomical,

SVD, plus top ~~few~~ $\sigma_1, \dots, \sigma_k$ k small

gives a good approximation to A

picture 1 ← $\vec{x}_1 \in \mathbb{R}^d$

picture 2 ← $\vec{x}_2 \in \mathbb{R}^d$

$\vdots$ $\qquad\qquad\qquad\qquad$ $\vdots \qquad \vdots$

picture n ← $\vec{x}_n \in \mathbb{R}^d$

but each picture an element of $\mathbb{R}^d$

$d \gg n$.

To each picture, we have

$$(\vec{x}_i, y_i)$$

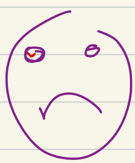
input feature list

output what we hope to infer

$$\min_{\vec{w} \in \mathbb{R}^d} \underbrace{\mathcal{E}_D(\vec{w})}_{\|X\vec{w} - \vec{y}\|^2 + \lambda \|w\|^2}$$

$\Rightarrow$ $\vec{w}^*$ is a minimizer, then

$$\underbrace{X^T X \vec{w} + \lambda \vec{w}}_{(X^T X + \lambda I_d)\vec{w}} = X^T \vec{y}$$



$d >> n$

$d \approx 12$ million, $n = 100, 1000,$

Class Ends

$$\vec{w}^* = \underbrace{\left(X^T X + \lambda I_d\right)}_{d \times d \text{ matrix}}{}^{-1} X^T y$$

☹

Lecture 10: Kernel Methods