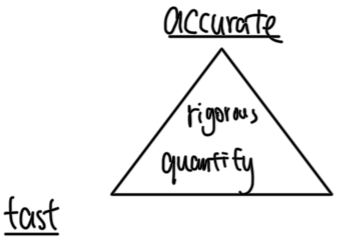
Lecture 1

Goal: Algorithms make "good" decisions based on data



trade off.

Interlude: Basic linear algebra

- $\langle u, v \rangle = \sum_{i=1}^{n} u_i v_i = u^T v$
- $uv^T = \begin{pmatrix} u_1 v_1 & \cdots & u_1 v_n \\ \vdots & \ddots & \vdots \\ u_n v_1 & \cdots & u_n v_n \end{pmatrix}$
- $\|u\|^2 = \langle u, u \rangle = Tr(uu^T)$
- $Tr(M) = \sum_{i=1}^{n} M_{ii}$
- $Tr(uv^T) = \sum_{i=1}^{n} u_i v_i = \langle u, v \rangle$
- $Tr(A + B) = Tr(A) + Tr(B)$ * linearity
- $Tr(A \cdot B) = \sum_i (AB)_{ii} = \sum_i \sum_j A_{ij} B_{ji} = \sum_j (BA)_{jj} = Tr(B \cdot A)$ * cyclicity

---

Regression data

response values $y_1 \cdots y_n \in \mathbb{R}$

feature vectors $x_1 \cdots x_n \in \mathbb{R}^d$

Example housing sale data

- n houses
- y_i = sales price of house i
- x_i = d numerical attributes of house i
  - living area
  - basement area
  - parking lot
  - .....

Learn from this data

Least Squares (LS) fit linear functions

minimize $f(\beta) = \sum_{i=1}^{n} (y_i - \beta^T x_i)^2$ such that $\beta \in \mathbb{R}^d$

Is LS solution $\hat{\beta}$ meaningful?

Matrix notation:

$$f(y) = \|y - X\beta\|^2, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}$$

($x_i^T \in \mathbb{R}^d$)

Simplistic model:

$y = X\beta^*$ for unknown $\beta^* \in \mathbb{R}^d$

Better model? Allow stochastic noise!

$\rightarrow \hat{\beta} = \beta^*$ assuming $\text{rank}(X) = d$

(Statistical) linear regression model

$$\underline{y} = X\beta^* + \underline{w}, \quad \underline{w} \sim N(0, \sigma^2 I_n)$$

Gaussian, mean, covariance

$\Leftrightarrow w_1, \dots, w_n \sim N(0, \sigma^2)$

* Double underliner for random variables.

assume $\text{rank}(X) = d$

w.l.o.g. $\frac{1}{n} X^T X = I_d$ * Columns of X are orthogonal to each other, $\|\cdot\|^2 = n$

Theorem LS estimator $\hat{\beta} = \arg\min_{\beta} \|\underline{y} - X\beta\|^2$

satisfies $\mathbb{E}[\|\beta^* - \hat{\beta}\|^2] = \sigma^2 \cdot \frac{d}{n}$ → goes to 0 if $n \geq w(d)$

Proof:

gradient: $\frac{1}{2n} \nabla_\beta f(\beta) = \nabla_\beta \left(\frac{1}{2n} \|X\beta - \underline{y}\|^2 \right) = \frac{1}{2n} X^T (X\beta - \underline{y}) \cdot 2 = \frac{1}{n} X^T (X\beta - X\beta^* - \underline{w})$

$$= \underbrace{\frac{1}{n} X^T X}_{I_d} (\beta - \beta^*) - \frac{1}{n} X^T \underline{w}$$

Optimality $\nabla f(\hat{\beta}) = 0$

$\rightarrow \hat{\beta} - \beta^* = \frac{1}{n} X^T \underline{w}$ $\longrightarrow \frac{1}{2n} \nabla_\beta f(\beta) = 0$ Minimize

$$\rightarrow \|\hat{\beta} - \beta^*\| = \frac{1}{n^2} \|X^T w\|^2 = \frac{1}{n^2} \text{Tr}(X^T w w^T X)$$

$$= \frac{1}{n^2} \text{Tr}(X X^T w w^T)$$

$$\rightarrow \mathbb{E}[\|\hat{\beta} - \beta^*\|^2] = \frac{1}{n^2} \mathbb{E}[\text{Tr}(X X^T w w^T)]$$

$$= \frac{1}{n^2} \text{Tr}(X X^T \underbrace{\mathbb{E}[w w^T]}_{\sigma^2 \cdot I_d})$$

* linearity of expectation

* $w \sim N(0, \sigma^2 I_d)$

$$= \frac{1}{n^2} \text{Tr}(X X^T) \cdot \sigma^2$$

$$= \frac{\sigma^2}{n} \text{Tr}\left(\frac{1}{n} X^T X\right)$$ * cyclicity

$$= \frac{\sigma^2}{n} \text{Tr}(I_d)$$ * $\frac{1}{n} X^T X = I_d$

$$= \frac{\sigma^2}{n} \cdot d$$

Do we have better algorithms?

Minimax optimality of LS

$\sigma^2 = 1$, matrix $X \in \mathbb{R}^{n \times d}$, $\frac{1}{n} X^T X = I_d$

arbitrary estimator: $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$, $y \mapsto \hat{\beta}(y)$

risk: $R(\hat{\beta}; \beta^*) = \mathbb{E}[\| \hat{\beta}(\underbrace{X\beta^* + W}_{Y \text{ observation}}) - \underbrace{\beta^*}_{\text{truth}} \|^2]$

We showed $R(\hat{\beta}_{LS}; \beta^*) = \frac{d}{n}$

Theorem: $\forall$ estimator $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$

$$\sup_{\beta^* \in \mathbb{R}^d} \underbrace{R(\hat{\beta}; \beta^*)}_{\text{Least upper bound on } R(\hat{\beta}; \cdot)} \geq \frac{d}{n}$$

Proof Strategy:

We show $\forall t \geq 0$, $\forall$ est. $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$

$$\mathbb{E}[R(\hat{\beta}; \beta^*)] \geq \frac{d}{n + \frac{1}{t}} \xrightarrow[t \to \infty]{} \frac{d}{n}$$

$\beta \sim N(0, t \cdot I_d)$

Bayes-Optimal estimator

Posterior-mean estimator

For dist. P over $\mathbb{R}^d$ and $\beta \sim P$

$$\hat{\beta}_P(y) = \mathbb{E}[\beta \mid \underbrace{X\beta + W}_{Y}]$$

$W \sim N(0, I_d)$

Theorem

$\forall$ dist. P, $\forall$ est. $\hat{\beta} : \mathbb{R}^n \to \mathbb{R}^d$

$\mathbb{E}[R(\hat{\beta}; \beta)] \geq \mathbb{E}[R(\hat{\beta}_P; \beta)]$

* If we have some prior, then less risks.
* Remember $R(\hat{\beta}; \beta^*) = \mathbb{E}[\| \beta^* - \hat{\beta}(X\beta^* + W) \|^2]$

Proof

Let $Y = X\beta + W$, $\hat{\beta} = \hat{\beta}(Y)$

Let $A = \hat{\beta} - \mathbb{E}[\beta \mid Y]$, $B = \mathbb{E}[\beta \mid Y] - \beta$

$\mathbb{E}[\|B\|^2] = \mathbb{E}[R(\hat{\beta}_P; \beta)]$

$\mathbb{E}[\|A + B\|^2] = \underbrace{\mathbb{E}[\|A\|^2]}_{\geq 0} + \mathbb{E}[\|B\|^2] + 2\,\mathbb{E}[\langle A, B \rangle]$

$\mathbb{E}[\|\hat{\beta} - \beta\|^2]$

$= \mathbb{E}[\| \underbrace{\hat{\beta} - \mathbb{E}[\beta \mid Y]}_{A} + \underbrace{\mathbb{E}[\beta \mid Y] - \beta}_{B} \|^2]$

$= \mathbb{E}[\langle A, B \rangle \mid Y] = \langle A, \mathbb{E}[B \mid Y] \rangle$

$= \langle A, \mathbb{E}[\mathbb{E}[\beta \mid Y] - \beta \mid Y] \rangle$

$= \langle A, \mathbb{E}[\beta \mid Y] - \mathbb{E}[\beta \mid Y] \rangle$

$= \langle A, 0 \rangle = 0$

Gaussian prior

$X \in \mathbb{R}^{n \times d}$, $\frac{1}{n} X^T X = I_d$, $W \sim N(0, I_n)$, $\beta \sim N(0, t \cdot I_n)$, $\beta \perp W$, $Y = X\beta + W$

$\hat{\beta}_{mean} = \mathbb{E}[\beta \mid Y]$

to show: $\mathbb{E}[\| \hat{\beta}_{mean} - \beta \|^2] = \frac{d}{n + \frac{1}{t}}$

Lemma: $\{\beta \mid y\} = N\left(\frac{1}{n+1/t} X^T y, \frac{1}{n+1/t} I_d\right)$

$\hat{\beta}_{mean} = \mu(y)$

The lemma implies $\mathbb{E}\left[\|\hat{\beta}_{mean} - \beta\|^2 \mid y\right] = d \cdot \frac{1}{n+1/t}$.