Lec 4

Tuesday, September 10, 2019 10:59

Recap:

let's pick the best linear model

$$f \in \{ f_\beta : \beta \in \mathbb{R}^p \}$$

$$f_\beta(x) = \beta^T x$$

best in terms of minimal avg squared errors

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Last time:

$$\hat{R}_n(f_\beta) = \frac{1}{n} \| \underline{Y} - \underline{X}\beta \|_2^2$$

$$\underline{Y} \in \mathbb{R}^n \quad \underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\underline{X} \in \mathbb{R}^{n \times p} \quad \underline{X} = \begin{pmatrix} - x_1 - \\ \vdots \\ - x_n - \end{pmatrix}$$

$$\underline{X}\beta = \begin{pmatrix} f_\beta(x_1) \\ \vdots \\ f_\beta(x_n) \end{pmatrix} \in \mathbb{R}^n$$

$$\nabla_\beta \hat{R}_n(f_\beta) = \begin{pmatrix} \partial \hat{R}_n(f_\beta)/\partial \beta_1 \\ \vdots \\ \partial \hat{R}_n(f_\beta)/\partial \beta_p \end{pmatrix} = \frac{2}{n} \underline{X}^T (\underline{Y} - \underline{X}\beta)$$

$\hat{R}_n(f_\beta)$ is diff'able & convex in β

so minimizers = critical pts

Want to solve

$$\frac{2}{n} \mathbb{X}^T (\mathbb{Y} - \mathbb{X}\hat{\beta}) = 0 \qquad \text{for } \hat{\beta}$$

$$\mathbb{X}^T \mathbb{Y} = \mathbb{X}^T \mathbb{X} \hat{\beta}$$

Get: $\hat{\beta} = \underbrace{(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T}_{\text{called the (left) pseudoinverse of } \mathbb{X}} \mathbb{Y}$

Wanted to solve $\mathbb{Y} \approx \mathbb{X}\beta$

can't solve exactly

pseudoinverse gives the closest-possible soln

$\hat{\beta}$ is called OLS

ordinary least squares

Conditional Expectation is the Best Regression Model

kNN, OLS are some regr models

Which is the BEST?

should minimizes

X, Y r.v.s representing new unseen random examples

$$R(f) = \mathbb{E}[\ell(Y, f(X))]$$
$$= \mathbb{E}[(Y - f(X))^2]$$

Preliminary warm up:

Given a r.v. Y, which $c \in \mathbb{R}$ minimizes $\mathbb{E}[(Y-c)^2]$? $\quad c = \mathbb{E}[Y]$

$$\frac{\partial}{\partial c} \mathbb{E}[(Y-c)^2] = \mathbb{E}[\frac{\partial}{\partial c}(c-Y)^2]$$

$$= \mathbb{E}[2(c-Y)]$$

$$= 2c - 2\mathbb{E}[Y] = 0 \Rightarrow c = \mathbb{E}Y$$

In words:

The mean (avg) is the single no. that is simultaneously closest to all vals of a random variable in avg squared dist.

Now consider minimizing

$$R(f) = \mathbb{E}[(Y - f(X))^2]$$

$$= \mathbb{E}[\mathbb{E}[(Y - f(X))^2 | X]]$$

expectation over Y drawn from $(Y|X)$

What should $f(X)$ be?

$$f^*(x) = \mathbb{E}[Y | X = x]$$

the optimal prediction

OLS regression estimating the conditional mean in some sense:

- we're replacing $\mathbb{E}$ with empirical avgs
- we're restricting to linear models

recap: Conditional means, modes, probs are the targets of supervised learning

Linear Model for Classification

Log Odds

Focus for now on binary classification

$G = \{0, 1\}$

Recall, Bayes classifier declare $\hat{Y}=1$

When $\mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x)$

Same as

$$O.R. = \underbrace{\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}}_{\displaystyle \parallel \atop \displaystyle \frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)} \in [0,\infty]} > 1$$

log odds: $\log(\text{O.R.}) = \log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) \in [-\infty,\infty]$

$$= \text{logit}(\mathbb{P}(Y=1|X=x))$$

for $p \in [0,1]$

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

$$\text{logit}: \underbrace{[0,1]}_{\text{domain}} \to \underbrace{[-\infty,\infty]}_{\text{co-domain}}$$

$\text{logit}(\mathbb{P}(Y=1|X=x))$ is score for declaring $\hat{Y}=1$

that's symmetric & takes vals in $[-\infty,\infty]$

→ When it's positive ⇒ declare $\hat{Y}=1$

→ — " — neg ⇒ declare $\hat{Y}=0$

<u>Logistic regression</u>

Posit $\text{logit}\, \mathbb{P}(Y=1|X=x) = \beta^T x$

$\mathbb{P}(Y=1|X=x) = \text{logit}^{-1}(\beta^T x)$

$$\pi(1|X=x) = \text{logit}^{-1}(\beta^T x)$$
$$= \sigma(\beta^T x)$$

Logistic sigmoid fn

$$\sigma(z) = \text{logit}^{-1}(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$$

σ gives us a way to transform a score into a prob

$$\sigma : \underbrace{[-\infty, \infty]}_{\text{domain}} \to \underbrace{[0,1]}_{\text{co-domain}}$$

$$\sigma(0) = \frac{1}{2}$$
$$\sigma(\infty) = 1$$
$$\sigma(-\infty) = 0$$
$$\sigma(-z) = 1 - \sigma(z)$$
$$\partial\sigma / \partial z = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$$
$$= \sigma(z) \cdot (1-\sigma(z)) = \sigma(z) \cdot \sigma(-z)$$

Fitting Logistic regression: Maximum Likelihood

Logistic regression provides a generative model for the data

— i.e., a model that specifies how (probabilistically) the data was generated

Given $X=x$, logistic regression says

$$Y \sim \text{Bernoulli}(\sigma(\beta_0^T x))$$

for some special β

β_0

What is β_0, or what β makes the data look the most likely under our generative model.