I, – Canvas

– Syllabus, no office hours today

– Classroom capture

– HW expectations

– Zoom office hour expectations

---

statistical learning vs. machine learning

Data examples

- spam
- CU FCQ (old ones)
  - Instructor effectiveness
  - Hours/week spent on HW
  - prior interest in class

Notation

$\underline{X}$, x = input / features / covariates / independent variables / predictors

$\underline{Y}$, y = response / output / dependent variable

Note! Different features will be denoted with subscripts, e.g.

x_1 = Hrs/week on HW x_2 = prior interest in class

Generic model! Given p predictors, the model is

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

where

f = systematic explainable variation

ε = mean zero random error representing unexplainable variation

1.1 Typical Goals

- Find relationships btwn features + response
- Interpret this relationship
- Predict new data outcomes

Often boils down to estimating f in

$$Y = f(X_1, \ldots, X_p) + \varepsilon$$

Prediction

Usually the Xs are easily available or controllable, but Y is the main quantity of interest.

Based on data, if we can estimate f, say $\hat{f}$ (statistic!), then at new Xs we can predict Y via:

$$\underset{\text{predicted response}}{\hat{Y}} = \underline{\hat{f}}(\underbrace{X_1, \ldots, X_p}_{\text{new features}})$$

For example, would doubling homework help my instructor rating?

$$\hat{Y} = \hat{f}(\underline{2X_1}, X_2)$$

Inference

Refers to estimating f + characterizing & using probabilistic models. Some questions:

1. Which predictors are associated with a response?
2. Are relationships linear or nonlinear?
3. Is there evidence of interactions?

Soap box Everything is inference, even prediction. If you can't say how confident you are in your prediction you shouldn't be predicting.

Jargon

Parametric vs. nonparametric

Supervised vs. unsupervised

Regression vs. classification

quantitative $\to$ $Y \in \mathbb{R}$ or continuous $Y \geq 0$

Y qualitative, $Y \in \{s_1, \dots, s_n\}$ takes on only finitely many unordered outcomes

$\boxed{\text{Theme: bias-variance trade-off}}$

Common to use mean squared error (MSE) to quantify the quality of predictions.

Model:

$$Y = f(\underline{X}) + \varepsilon, \qquad \varepsilon \text{ mean zero}$$

based on data we have $\hat{f}$. For a new feature, x_0, we have

$$Y = f(x_0) + \varepsilon,$$

where we use $\hat{f}(x_0)$ to predict y.

Assume ε & $\hat{f}$ are uncorrelated.

Predictive MSE is:

$$E\left[(y-\hat{f}(x_0))^2\right] = E\left[(y-\hat{f})^2\right]$$

$$= E\left[(f+\varepsilon-\hat{f})^2\right] = E\left[\left((f-\hat{f})+\varepsilon\right)^2\right]$$

$$= E\left[(f-\hat{f})^2 + \varepsilon^2 + 2\varepsilon(f-\hat{f})\right]$$

$$= E\left[(f-\hat{f})^2\right] + \underbrace{E\varepsilon^2} + \underbrace{E\left[2\varepsilon(f-\hat{f})\right]}$$

$$2E\left[\varepsilon(f-\hat{f})\right]$$

$$2(E\varepsilon)E(f-\hat{f})$$

$$=0$$