101C

ADMIN STUFF

IMPORTANT STUFF

- Read the syllabus (on CCLE)
- Office Hours: Thursday 5:00-6:00pm
- Midterm in class: November 12
- Final: TBD
- Send all academic questions to Piazza, not email. Feel free to answer/participate in discussions.

101C

PREAMBLES

STRATEGY VS TACTICS

Strategy without tactics is the slowest route to victory.
Tactics without strategy is the noise before defeat.

– Sun Tzu

LECTURE R CODE

- Please visit
- https://github.com/davezes/fall2019_101C
- Clone/DL repo

101C

1

HEADS UP

- Advanced

WHAT'S THE POINT

- Inference
- 2 views, "System", Population
  - "Probability space"
  - "System"

SUPERVISED VS UNSUPERVISED

- Data modeling can be divided into two paradigms.
- **Supervised**. Simply, we have at least one or more responses that we wish to predict from explanatory variables.
- **Unsupervised**. We do not have a response, but rather seek to find "patterns" amongst our variables.
- Most of ISLR, and this course, is dedicated to exploring the **supervised** paradigm.

COST FUNCTIONS

- **How do we assess the "quality" of our prediction?**
- It depends.
- We don't spend a great deal of time studying/considering different cost functions in 101C.
- Actuarial sciences.
- Quantitative Response, MSE, RMSE (same minimum)
- Example, AC "attainment"

COST FUNCTIONS

- Very common cost function for quantitative response is the RMSE.
- RV version: $\sqrt{\mathrm{E}\left[\left(Y-\hat{Y}\right)^2\right]}$
- Data version: $\sqrt{\frac{1}{n}\sum_{i}^{n}\left[\left(y_i-\hat{y}_i\right)^2\right]}$

STRATEGERY

- Our general business in this course, and others like it, is to create $\hat{f}(\mathbf{x})$ from data, $(\mathbf{X}, \mathbf{y})$.
- This process itself is a *function*.

FUNCTIONS THAT CREATE FUNCTIONS

- For example, consider the multivariate linear solution:
- $\hat{f}(\mathbf{x}) = \mathbf{x}\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} = \mathbf{x}\,\widehat{\boldsymbol{\beta}}$
- Notice that this itself is a function of our data, $(\mathbf{X}, \mathbf{y})$.

FUNCTIONS THAT CREATE FUNCTIONS

- We can think of this process as:
- $h(\mathbf{X}, \mathbf{y}) = x\,(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} = x\,\widehat{\boldsymbol{\beta}} = \hat{f}(x)$
- The important thing to notice is that h is a function of our observations, $(\mathbf{X}, \mathbf{y})$, whereas the function it creates, $\hat{f}$, need not be a function of observations.

FUNCTIONS THAT CREATE FUNCTIONS

- h may require constants for the creation of $\hat{f}$.
- Such constants may be called "hyper-parameters".
- Just for example, in Ridge regression, we have $h(\mathbf{X}, \mathbf{y}, \lambda) = x\,(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\,\mathbf{X}^T\mathbf{y} = x\,\widehat{\boldsymbol{\beta}} = \hat{f}(x)$.

FUNCTIONS THAT CREATE FUNCTIONS

- The ISLR book we're using this quarter, and others like it, are really a collection of different h 's. In some sense, there's little more to it than that.
- Understanding, at the most fundamental level, that estimation techniques are ultimately build atop functions of our data, makes dealing with "Testing", "Training", "Validation", much more natural.

FUNCTIONS THAT CREATE FUNCTIONS

- It also makes dealing with some of the key theoretical concepts much easier.
- If you don't like the word "theory", you can substitute "how/why it works".

KEYS

- Two of the keys to the theory of estimation are found in Chapter 2.
- Total Estimation Error (the square of which is called the Prediction Variance), pg 19
- The Bias-Variance Trade-Off, pg 34

SYSTEM

- $y = f(x) + \varepsilon \ , \quad \varepsilon \sim \mathcal{N}[0, \sigma]$

TOTAL ESTIMATION ERROR

- Total Estimation Error (the square of which is called the Prediction Variance):
- $\mathrm{E}\left[\left(Y-\hat{Y}\right)^2\right] = \mathrm{E}\left[\left(f(x)+\varepsilon-\hat{f}(x)\right)^2\right] = \left(f(x)-\hat{f}(x)\right)^2 + \mathrm{Var}[\varepsilon]$
- This is easily demonstrated by expanding and factoring terms …
- However, to pull this off, we need to recognize two basic properties …

TOTAL ESTIMATION ERROR

- The first, that $\mathrm{E}\left[f(x)\ \varepsilon\right] = 0$. This follows directly from the definition of Y.
- The second is much more interesting, that $\mathrm{E}\left[\hat{f}(x)\ \varepsilon\right] = 0$.
- Technically, this requirement **may not in general be true**.
- $\hat{f}$ **could be** a stochastic function of some correlate of ε.
- It is not — not in 101C.

TOTAL ESTIMATION ERROR

- This becomes entirely clear if we turn back to our good friend, the meta function, h.
- There are two ways to conceptualize how h creates $\hat{f}$.
- One, likely the most comfortable, is to view h as a function of **data**, $h(\mathbf{X}, \mathbf{y})$, and perhaps some hyper-parameters. While these data may be realized through a stochastic process, once they are realized they are considered fixed – and so too the function $\hat{f}$ that h has created.

TOTAL ESTIMATION ERROR

- The other is to allow h to be a function of Y as a random variable, $h(\mathbf{X}, Y)$, so that, interestingly, $\hat{f}(x)$ can be thought of as a random variable.
- Either way, the result is the same:
- Since ε are **mutually independent**, then $\mathrm{E}\left[\hat{f}(x)\ \varepsilon\right] = 0.$

TOTAL ESTIMATION ERROR

$$\mathrm{E}\left[\left(Y-\hat{Y}\right)^2\right] = \underbrace{\left(f(x)-\hat{f}(x)\right)^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}[\varepsilon]}_{\text{Irreducible}}$$