Lecture 2:
The classic paradigm
vs
the predictive paradigm

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 13, 2020

Agenda

Recap Last Class

Motivation

- We discussed the examples of Google Flu and Facebook face detection
  - Take away, the success was driven by an empiric approach
  - Given data estimate a function $f(x)$ that predicts y from x
- This is basically what we do as economists everyday so:
  - Are these algorithms merely applying standard techniques to novel and large datasets?
  - If there are fundamentally new empirical tools, how do they fit with what we know?
  - As empirical economists, how can we use them?

The Classic Paradigm

$$Y = f(X) + u \tag{1}$$

- Interest lies on inference
- "Correct" $f()$ to understand how Y is affected by X
- Model: Theory, experiment
- Hypothesis testing (std. err., tests)

Example: OLS and the classical model

Set

$$f(X) = X\beta \tag{2}$$

- $Y = X\beta + u$
- Interest in β
- The model is given.
- Problem: how to estimate β in the given model?
- Minimize SSR

$$\hat{\beta} = (X'X)^{-1}X'y \tag{3}$$

- Gauss-Markov: under the classical assumptions it is BLUE
- Classical assumptions: how they affect properties (omitted variables, endogeneity, heteroscedasticity, etc.)

The Predictive Paradigm

$$Y = f(X) + u \tag{4}$$

- Interest on predicting Y
- "Correct" $f()$ to be able to predict (no inference!)
- Model?

Statistical Decision Theory: A bit of theory

- We need a bit of theory to give us a framework for choosing f
- A decision theory approach involves an **action space** $\mathcal{A}$
- The **action space** $\mathcal{A}$ specify the possible "actions we might take"
- Some examples

Table 1: Action Spaces

Inference	Action Space
Estimation $\theta, g(\theta)$	$\mathcal{A} = \Theta$
Prediction	$\mathcal{A} = \textit{space of } X_{n+1}$
Model Selection	$\mathcal{A} = \{\textit{Model I}, \textit{Model II}, ...\}$
Hyp. Testing	$\mathcal{A} = \{\textit{Reject} \mid \textit{Accept } H_0\}$

Statistical Decision Theory: A bit of theory

- After the data $X = x$ is observed, where $X \sim f(X|\theta)$, $\theta \in \Theta$
- A decision is made
- The set of allowable decisions is the action space ($\mathcal{A}$)
- The loss function in an estimation problem reflects the fact that if an action a is close to θ,
  - then the decision a is reasonable and little loss is incurred.
  - if it is far then a large loss is incurred

$$L : \mathcal{A} \to [0, \infty] \tag{5}$$

Statistical Decision Theory: A bit of theory

Loss Function

- If θ is real valued, two of the most common loss funcitons are
  - Squared Error Loss:

$$L(a,\theta) = (a-\theta)^2 \tag{6}$$

  - Absolute Error Loss:

$$L(a,\theta) = |a-\theta| \tag{7}$$

- These two are symmetric functions. However, there's no restriction. For example in hypothesis testing a "0-1" Loss is common.
- Loss is minimum if the action is correct

Statistical Decision Theory: A bit of theory

Risk Function

In a decision theoretic analysis, the quality of an estimator is quantified by its risk function, that is, for an estimator $\delta(x)$ of θ, the risk function is

$$R(\theta,\delta) = E_{\theta}L(\theta,\delta(X)) \tag{8}$$

at a given θ, the risk function is the average loss that will be incurred if the estimator $\delta(X)$ is used

- since θ is unknown we would like to use an estimator that has a small value of $R(\theta,\delta)$ for all values θ
- Loss is minimum if the action is correct
- If we need to compare two estimators (δ_1 and δ_2) then we will compare their risk functions
- If $R(\delta_1,\theta) < R(\delta_2,\theta)$ for all $\theta \in \Theta$, then δ_1 is preferred becasue it performs better for all θ

How to choose f for prediction

- In a prediction problem we want to predict Y from $f(X)$ in such a way that the loss is minimum
- Assume also that $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ with joint distribution $Pr(X, Y)$

$$R(Y, f(X)) = E[(Y - f(X)^2] \tag{9}$$

$$= \int (y - f(x)^2 Pr(dx, dy) \tag{10}$$

conditioning on X we have that

$$R(Y, f(X)|X) = E_X E_{Y|X}[(Y - f(X))^2|X] \tag{11}$$

this risk is also know as the **mean squared (prediction) error** $Err(f)$

Mean square error

It suffices to minimize the $Err(f)$ point wise so

$$f(x) = argmin_m E_{Y|X}[(Y - m)^2 | X = x) \tag{12}$$

Y a random variable and m a constant (predictor)

$$min_m E(Y - m)^2 = \int (y - m)^2 f(y) dy \tag{13}$$

Result: The best prediction of Y at any point $X = x$ is the conditional mean, when best is measured by mean squared error.

Mean square error

Proof
FOC

$$\int -2(y-m)f(y)dy = 0 \tag{14}$$

Divided by -2 and clearing

$$m\int (y)dy = \int yf(y)dz = 0 \tag{15}$$

$$m = E(Y|X = x) \tag{16}$$

The best prediction of Y at any point $X = x$ is the conditional mean, when best is measured by mean squared error.

Reducible and Irreducible Error

$$Y = f(X) + u \tag{17}$$

- If f were known and X were observable, the problem comes down to predicting u
- Given that u is not observable, the best prediction in MSE is its expectation. u is the irreducible error
- When $f(.)$ is also unknown, the prediction problem is reduced to knowing $f(.)$
- The ’reducible’ error refers to the discrepancy between $\hat{f}(.)$ and $f(.)$

Reducible and irreducible error

- Let's think about our usual problem $f(.)$ is unknown
- Consider a given estimate $\hat{f}$ and a set of predictors
- this predictors yield $\hat{Y} = \hat{f}(x)$.
- For now assume $\hat{f}$ and X are fixed (Hastie et al. make this assumption any idea why?)
- Then we can show that the mean square error

$$E(Y - \hat{Y})^2 = E(f(X) + u - \hat{f}(X))^2 \tag{18}$$

$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{Reducible} + \underbrace{Var(u)}_{Irreducible} \tag{19}$$

Reducible and irreducible error

$$E(Y-\hat{Y})^2 = \underbrace{[f(X)-\hat{f}(X)]^2}_{Reducible} + \underbrace{Var(u)}_{Irreducible} \tag{20}$$

- The focus the is on techniques for estimating f with the aim of minimizing the reducible error
- It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y
- This bound is almost always unknown in practice

Variance Decomposition/ Bias

Remember

- $Bias(\hat{f}(X)) = E(\hat{f}(X)) - f = E(\hat{f}(X) - f(X))$
- $Var(\hat{f}) = E(\hat{f} - E(\hat{f}))^2$

Result (very important!)

$$MSE = Bias^2(\hat{f}(X)) + V(\hat{f}(X)) \tag{21}$$

Proof: as an exercise

The econometric approach

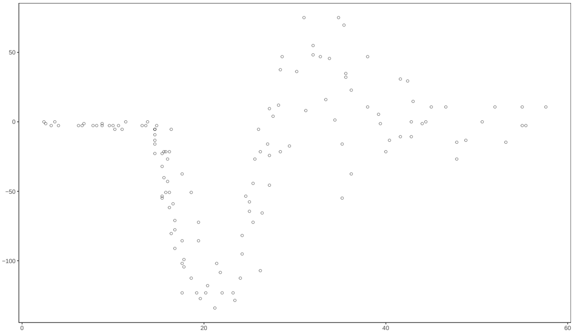
$$MSE = Bias^2(\hat{f}(X)) + V(\hat{f}(X)) \tag{22}$$

- When $\hat{f}(X)$ is unbiased, minimize MSE $\hat{f}(X)$ is reduced to minimize $V(\hat{f}(X))$
- The best kept secret: tolerating some bias is possible to reduce V($\hat{f}(X)$) and lower MSE
- If the goal is to predict, it is not a problem to tolerate biased estimates
- It could be the case that the MSE is minimum for biased predictors

How to estimate $f()$

- Parametric methods $\rightarrow$ assume the functional form $\rightarrow$ from economic theory?

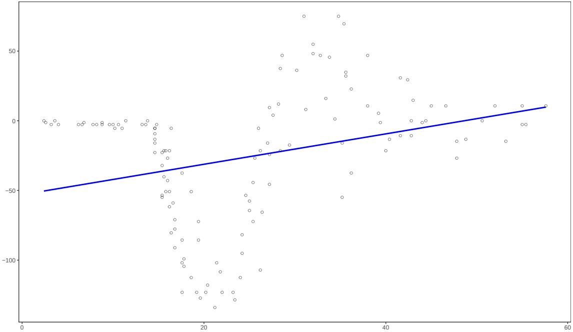
How to estimate $f(.)$



Source: motorcycle data from `https://www.stata-press.com/data/r12/r.html`

How to estimate $f(.)$

- Linear $f(X) = X\beta$



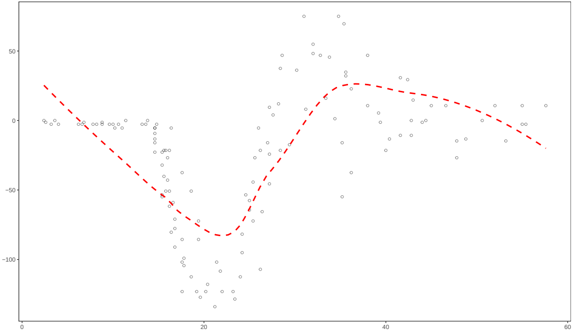
Source: motorcycle data from https://www.stata-press.com/data/r12/r.html

How to estimate $f()$

- Parametric methods $\rightarrow$ assume the functional form $\rightarrow$ from economic theory?
- Non-Parametric methods $\rightarrow$ no assumption about $f()$ let the data speak

How to estimate $f(.)$

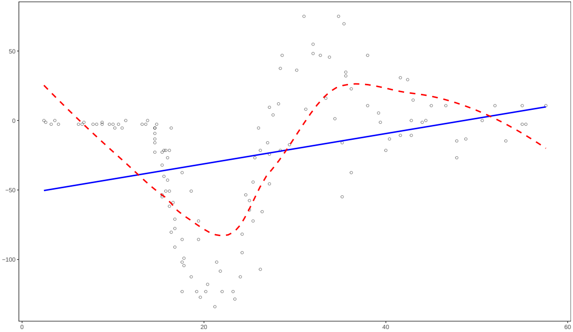
- Local Polynomial Regression



Source: motorcycle data from https://www.stata-press.com/data/r12/r.html

How to estimate $f(.)$

- Linear vs Local Polynomial Regression



Source: motorcycle data from https://www.stata-press.com/data/r12/r.html

Accuracy, complexity and interpretability

Source: https://imgs.xkcd.com/comics/machine_learning.png

Accuracy, complexity and interpretability

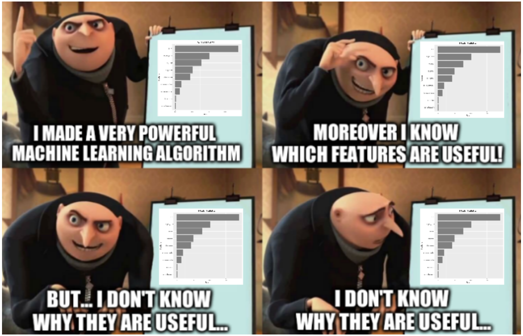
Recall the problem of interpretation in

$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + u \tag{23}$$

- We have lost the interpretation of β_2 as a marginal effect
- In a non-linear model the interpretations are no longer trivial
- Machine learning: we quickly lose interpretability in predictive quality post
- Is this a problem?

Accuracy, complexity and interpretability

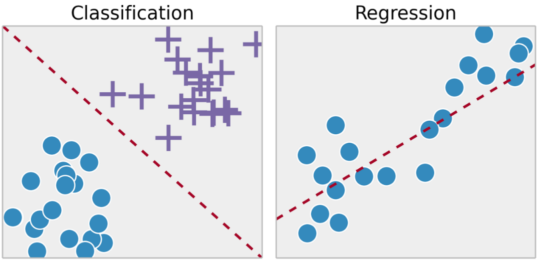
- Is this a problem?



Source: shorturl.at/gm013

Supervised vs Unsupervised

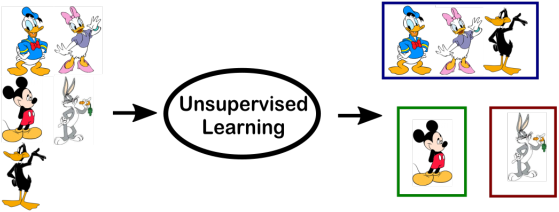
- Supervised Learning
  - for each predictor x_i a 'response' is observed y_i.
  - everything we have done in econometrics is supervised



Source: shorturl.at/opqKT

Supervised vs Unsupervised

- Unsupervised Learning
  - observed x_i but no response.
  - example: cluster analysis



Source: shorturl.at/opqKT

Recap

- We start shifting paradigms
- Tools are not that different (so far)
- Decision Theory: Risk with square error loss $\rightarrow$ MSE
- Objective minimize the reducible error
- Irreducible error our unknown bound
- Machine Learning best kept secret: some bias can help lower MSE

Next

- GitHub Demo
- Questions about software installation
  - `R` and `RStudio`
  - `Conda`?

Further Readings

- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

- Mullainathan, S. and Spiess, J., 2017. Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), pp.87-106.