Statistical Learning - Introduction

Julian D. Karch

March 3, 2023

Outline

1. About This Course
2. Inference vs. Prediction
3. Bias-Variance Tradeoff
4. k-nearest Neighbors

About This Course

This Course

- New methodology for data analysis
- Different focus, Prediction!
- Machine Learning / Computer Science
- Statistics and Machine Learning: Statistical Learning

This Course (cont'd)

Meetings

Lecture, Wednesday **09:00**-10:45

Working Group, Friday **09:00**-10:45

Rooms

Different rooms and even buildings! See `https://rooster.universiteitleiden.nl/`

This Course (cont'd)

- Course book: ISLR.jpg
- Use of corresponding online lectures
- Watch video lectures before lecture
- Read book chapters between lecture and workgroup
- Weekly take-home exercise (pass or fail) after workgroup

Three Professors

- Dr. Anikó Lovik - unsupervised learning
- Dr. Marjolein Fokkema - advanced supervised learning
- Dr. Julian Karch - basic supervised learning, coordinator

Schedule

https://brightspace.universiteitleiden.nl//content/enforced/208559-4433STLT6Y_2223_S2/schedule.pdf

Assignments

Your course grade will be determined based on:

- **Homework assignment 1 (1/3)**
- **Homework assignment 2 (1/3)**
- **Presentation assignment (1/3)**

To pass the course, you must also pass 9 out of 12 weekly assignments. Details can be found at `https://brightspace.universiteitleiden.nl/d2l/le/lessons/208559/topics/2281907`.

Programming Language

- Course instructors will employ R for exercises.
- You may use Python for exercises and assignments, but instructors may not be able to assist with errors or problems.

- Statistical learning refers to vast set of tools for understanding data.
  - Supervised: $Y \leftarrow f(X_1, \ldots, X_p)$; predict Y on the basis of X
  - Unsupervised: $X_1, \ldots, X_p$; finding structure in X (underlying dimensions/groups)

Inference vs. Prediction

Introduction

General Setup

- $Y = f(X) + \epsilon$, with $Y =$ outcome variable, $X_1, \dots, X_p$, p predictors, $\epsilon =$ error term
- f describes the true relationship between predictors and outcome.

Concrete Example

- Test Score $= 3 \times$ IQ $+ 10 \times$ Motivation $+ \epsilon$
- Thus if we have two people that differ by one in both IQ and Motivation, *on average*, their test scores will differ by 13

Introduction (cont'd)

Not Causal!

f is not (necessarily) causal! An increase of 1 in motivation does not necessarily lead to an increase of 10 in test score. ??

Different Goals: Inference

Both inference and prediction aim to find a $\hat{f}$ as a substitute for the true f but with different goals.

Inference

Establish how predictors are *related* to test scores in the population:

estimation

1. $\hat{f}$ should match f as closely as possible.
2. $\hat{f}$ should be interpretable.
3. We want to quantify how close $\hat{f}$ is to f.

Different Goals: Prediction

Prediction

1. Find $\hat{f}$ that makes most accurate predictions for unseen Ys
2. Estimate how well $\hat{f}$ predicts unseen Ys

Optimal Answer is the Same true f

Although these different goals lead to different statistical approaches, for both inference and prediction, the optimal $\hat{f}$ is the true f.

Linear Regression Model

- The linear regression model

$$\hat{Y} = f(X_1, \ldots, X_p) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

can be used for inference and/or prediction.

Inferential Regression

- Suppose we have data and obtain estimates:

$$\hat{Y} = 2 + 2\text{IQ} + 9\text{Motivation}$$

- If the regression assumptions are met:
  1. Estimated coefficients are best estimate of true coefficients (for a particular definition of best; MVU)
  2. Estimated coefficients can be interpreted: An increase of 1 in motivation is associated with an increase of 9 in the test score
  3. Confidence / credibility intervals indicate how far the estimates ?? are from true cofficients
  4. Statistical tests can provide effidence for whether a predictor is really related to the outcome variable, given the other variables.

Optimal Estimation

Classical

- Require unbiasedness, that is, $\mathbb{E}[\hat{\beta}_j] = \beta_j$ for all j
- Among the unbiased estimators, search for lowest MSE

$$MSE_{\text{inf}} = \sum_j \mathbb{E}[(\hat{\beta}_j - \beta_j)^2]$$

- Minimum-variance unbiased (MVU): unbiased + *always* lowest MSE (among unbiased)

Existence Common

Often a MVU estimator exist. For example, OLS regression coefficients, sample means, ...

Optimal Estimation

Modern

- Give up unbiasedness in favor of lower MSE
- Example: Stein's paradox: for more than 3 dimensions, there is a (biased) estimator that *always* has a lower MSE than the sample mean (`https://www.youtube.com/watch?v=cUqoHQDinCM&t=756s`)

Existence Essentially Impossible

A estimator that always has lowest MSE does not exist for any (meaningful) problem

Predictive Regression

- Suppose we have data and obtain estimates:

$$\hat{Y} = 2 + 2\text{IQ} + 9\text{Motivation}$$

- Suppose we have a *new* observation
$x_1 = [\text{IQ} = 100 \quad \text{Motivation} = 3]$
- With these values we can predict Y, i.e.,
$2 + 2 \times 100 + 9 \times 3 = 229$
- We do not care to recover parameters that generated the data, but want to obtain a $\hat{f}$ that yields as accurate as possible $\hat{f}(X) = \hat{Y}$.
- I.e., minimize

$$MSE_{\text{pred}} = \mathbb{E}(\hat{f}(X) - Y)^2$$

How far, on average, are our predictions $\hat{f}(X)$ from the true values Y

R Example

See R slides

Bias-Variance Tradeoff

No Free Lunch Theorem

Optimally

Method that based on training set D, returns $\hat{f} = f$ minimizing MSE_{pred}

Impossible

Does not exist; No method can return true f based on finite data set D. Even worse, we do not know (beforehand) which method performs best for a particular data set.

Solution

- Apply multiple methods, e.g., linear and polynomial regression to training set
- Use test set to estimate MSE_{pred}
- How to best select the methods? Should I try a flexible method or not?

Method MSE

- Instead of the performance of a fixed prediction function $\hat{f}$, we consider the performance of a method (e.g. linear regression) repeatedly applied to data from the same population.
- We then ask which statistical method, on average, leads to the best prediction function $\hat{f}$

Formally

Probability distribution P^*, $(X, Y) \sim P^*$, training set of n i.i.d realizations from (X, Y), and $\hat{f}(X; D) = \hat{Y}$ is a statistical method.

D: training set

$$\text{EPE} = E_{X,Y}\left[E_{\mathcal{D}}\left[\{Y - \hat{f}(X; \mathcal{D})\}^2\right]\right] \tag{1}$$

Bias-Variance Tradeoff Formal

$$EPE = (\text{Bias})^2 + \text{Variance} + \text{Irreducible error}$$

$$(\text{Bias})^2 = E_X\left[\left\{E_{\mathcal{D}}\left[\hat{f}(X;\mathcal{D})\right] - Y\right\}^2\right]$$

$$\text{Variance} = E_X\left[E_{\mathcal{D}}\left[\left\{\hat{f}(X;\mathcal{D}) - E_{\mathcal{D}}\left[\hat{f}(X;\mathcal{D})\right]\right\}^2\right]\right]$$

$$\text{Irreducible error} = E_{X,Y}\left[\{Y - f(X)\}^2\right] = \sigma_\epsilon^2.$$

D_1

D_2

D_3

$$T\left\{\begin{array}{ll} X_1 & Y_1 \\ X_2 & Y_2 \\ X_3 & Y_3 \end{array}\right.$$

Bias-Variance Tradeoff Text

- $(\text{Bias})^2$ = Consider a fixed value of $X = x_0$. Obtain predictions for this value of X using the model trained on infinitely many training sets of size n. Average these predictions and compare the result to the true value. Repeat for all x_0 values and average those results. $\Rightarrow$ *How far are the average predictions from the true values?*
- Variance = Fix $X = x_0$ and obtain predictions for each of the infinitely many training sets. Compute the variance of these predictions. This is the variance for x_0. The total variance is the average of the variances across all possible X values. $\Rightarrow$ *How much do the predictions differ from one training set to another?*

Bias-Variance Composition Intuition

huge training set

- Low Bias, High Variance $\Rightarrow$ Averaging across training sets leads to perfect prediction. However, for a particular training set we are likely far away from this perfect prediction $\Rightarrow$ High EPE

small training set

- High Bias, Low Variance $\Rightarrow$ For a particular training set we are likely close to the average prediction. However, the average prediction is far away from the perfect prediction $\Rightarrow$ High EPE
- Low bias, Low Variance $\Rightarrow$ Averaging across training sets leads to perfect prediction and for a particular training set, we are likely close to the perfect prediction $\Rightarrow$ Low EPE

Low Bias and Variance?

Warning!

Both variance and bias and relative to the population, especially f(X).

- If we have good knowledge about $f(X)$ (say it's linear), we can identify a method with low bias, and low variance: Linear regression (with shrinkage)
- Typically bias and variance of a method are discussed as a property of the method, independent of the population.
- Implicit assumption: $f(X)$ is rather complex, nonlinear

Bias Variance Tradeoff

- Flexible methods $\Rightarrow$ low bias, high variance
- Inflexible methods $\Rightarrow$ high bias, low variance

Overfitting + Underfitting

See Rscript (also on Brightspace).

k-nearest Neighbors

Linear Model

Often we fit a linear model, assuming that f is linear.

This assumption is most likely false! Why does it often work so well?

Population: Nonlinear Regression

The true f corresponds to the conditional means at each point x

Sample Data

Using sample data, we want to obtain an estimate $\hat{f}(X)$ of $f(X)$.

- Due to sparsity, cannot estimate a conditional mean at all points $(X = x)$.
- Thus, take a small neighbourhood around $X = x$ and take neighbourhood mean as predicted value, i.e. *nearest neighbour averaging*.
  What happens to bias and variance if size of neighbourhood increases?

Curse of Dimensionality

- With multiple predictors the observations are further spread out through the space
- Essential reason: with each predictor "volume" of space is multiplied
- Nearest neighbours might not be near at every point
- This is known as the *curse of dimensionality*
- More structure in f is needed
- *How can we impose structure?*

Curse_Captain_Potrait.png

Conclusion

- Larger noise increases variance $\Rightarrow$ favors inflexible method (does not overfit noise as dramatically)
- More dense sampling of feature space allows distinguishing noise from signal $\Rightarrow$ favors flexible method
- Larger sample size $\Rightarrow$ favors flexible method
- Larger amount of predictors $\Rightarrow$ favors inflexible method
- Very nonlinear f $\Rightarrow$ favors flexible method