An Overview of Statistical Learning

Nate Wells

Math 243: Stat Learning

September 4th, 2020

Outline

In today's class, we will. . .

Outline

In today's class, we will. . .

- Review matrix notation

Outline

In today's class, we will...

- Review matrix notation
- Discuss the goals of statistical learning algorithms

Outline

In today's class, we will...

- Review matrix notation
- Discuss the goals of statistical learning algorithms
- Survey some of the most common methods for statistical learning

Outline

In today's class, we will. . .

- Review matrix notation
- Discuss the goals of statistical learning algorithms
- Survey some of the most common methods for statistical learning
- Analyze data from the 'guess my age' activity

Section 1

Vectors and Matrices

atrices

- An $n \times p$ matrix $\mathbf{X}$ is an array of np numbers, arranged into n rows and p columns.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathbf{X} \text{ is } 3 \times 4$$

atrices

- An $n \times p$ matrix $\mathbf{X}$ is an array of np numbers, arranged into n rows and p columns.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathbf{X} \text{ is } 3 \times 4$$

- The (i,j)-entry of $\mathbf{X}$ is denote $x_{i,j}$ and is the entry in the ith row and jth column of $\mathbf{X}$

$$x_{1,2} = 2 \qquad x_{2,2} = 6 \qquad x_{3,4} = 12$$

atrices

- An $n \times p$ matrix $\mathbf{X}$ is an array of np numbers, arranged into n rows and p columns.

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathbf{X} \text{ is } 3 \times 4$$

- The (i, j)-entry of $\mathbf{X}$ is denote $x_{i,j}$ and is the entry in the ith row and jth column of $\mathbf{X}$

$$x_{1,2} = 2 \qquad x_{2,2} = 6 \qquad x_{3,4} = 12$$

- For us, rows will index samples or observations (from 1 to n), while columns will index variables (from 1 to p); this is consistent with the tidy dataframe structure

Vectors and Transposes

- The transpose of a matrix $\mathbf{X}$, denoted $\mathbf{X}^T$, is the matrix obtained switching rows and columns. (That is, the (i, j) entry of $\mathbf{X}^T$ is the (j, i) entry of $\mathbf{X}$)

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathbf{X}^T = \begin{pmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{pmatrix}$$

Vectors and Transposes

- The transpose of a matrix $\mathbf{X}$, denoted $\mathbf{X}^T$, is the matrix obtained switching rows and columns. (That is, the (i, j) entry of $\mathbf{X}^T$ is the (j, i) entry of $\mathbf{X}$)

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathbf{X}^T = \begin{pmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{pmatrix}$$

- An n-dimensional vector $\mathbf{v}$ is an ordered list of n numbers. By default, an n-dimensional vector is represented as a $n \times 1$ matrix

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Vectors and Transposes

- The transpose of a matrix $\mathbf{X}$, denoted $\mathbf{X}^T$, is the matrix obtained switching rows and columns. (That is, the (i,j) entry of $\mathbf{X}^T$ is the (j,i) entry of $\mathbf{X}$)

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathbf{X}^T = \begin{pmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{pmatrix}$$

- An n-dimensional vector $\mathbf{v}$ is an ordered list of n numbers. By default, an n-dimensional vector is represented as a $n \times 1$ matrix

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Rows and Columns

- We often are interested in the entries in the *i*th row of **X**, which we will denote using the vector x_i (recall vectors are by default, column vectors). It is the list of data on the *i*th individual in the sample

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^2 \\ \vdots \\ x_n^T \end{pmatrix}$$

Rows and Columns

- We often are interested in the entries in the ith row of $\mathbf{X}$, which we will denote using the vector x_i (recall vectors are by default, column vectors). It is the list of data on the ith individual in the sample

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad x_i^T = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^2 \\ \vdots \\ x_n^T \end{pmatrix}$$

- In other situations, we consider the jth column of a matrix, denoted $\mathbf{x}_j$. It is the list of values for jth variable in the sample

$$\mathbf{x}_i = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix}$$

Summary

- Vectors of length n (corresponding to the sample size) will be denoted using lower case bold letters: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}_1$.

Summary

- Vectors of length n (corresponding to the sample size) will be denoted using lower case bold letters: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}_1$.
- Vectors of length p (corresponding to the number of predictor variables) will be denoted using lower case normal font letters: x, y, x_1.

Summary

- Vectors of length n (corresponding to the sample size) will be denoted using lower case bold letters: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}_1$.
- Vectors of length p (corresponding to the number of predictor variables) will be denoted using lower case normal font letters: x, y, x_1.
- Individual numbers will also be denoted using lower case normal font letters (but usually with two subscripts): x_{ij}.

Summary

- Vectors of length n (corresponding to the sample size) will be denoted using lower case bold letters: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}_1$.
- Vectors of length p (corresponding to the number of predictor variables) will be denoted using lower case normal font letters: x, y, x_1.
- Individual numbers will also be denoted using lower case normal font letters (but usually with two subscripts): x_{ij}.
- Matrices will be denoted using capital bold letters: $\mathbf{X}$, $\mathbf{A}$

Summary

- Vectors of length n (corresponding to the sample size) will be denoted using lower case bold letters: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}_1$.
- Vectors of length p (corresponding to the number of predictor variables) will be denoted using lower case normal font letters: x, y, x_1.
- Individual numbers will also be denoted using lower case normal font letters (but usually with two subscripts): x_{ij}.
- Matrices will be denoted using capital bold letters: $\mathbf{X}$, $\mathbf{A}$
- We will use capital normal font letters to denote variables. X is usually used for predictor variables, and Y is used for response variables

Section 2

What is Stat Learning

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ and zero, one, or more response variables $Y, Y_1, \ldots$.

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ and zero, one, or more response variables $Y, Y_1, \ldots$.
- In the simplest case, we observe the values of a quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \dots, X_p$ and zero, one, or more response variables $Y, Y_1, \dots$.
- In the simplest case, we observe the values of a quantitative response Y, as well as p many predictors $X_1, \dots, X_p$.
- We assume there is a relationship between these observed values:

$$Y = f(X_1, \dots, X_p) + \epsilon$$

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ and zero, one, or more response variables $Y, Y_1, \ldots$.
- In the simplest case, we observe the values of a quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.
- We assume there is a relationship between these observed values:

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

- Here, ϵ represents a random or unobserved error term

The Setting

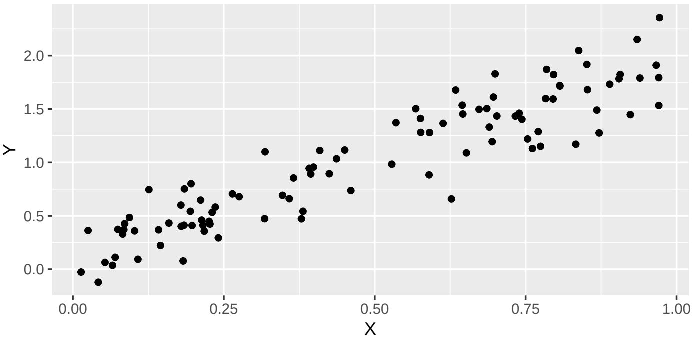
- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ and zero, one, or more response variables $Y, Y_1, \ldots$.
- In the simplest case, we observe the values of a quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.
- We assume there is a relationship between these observed values:

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

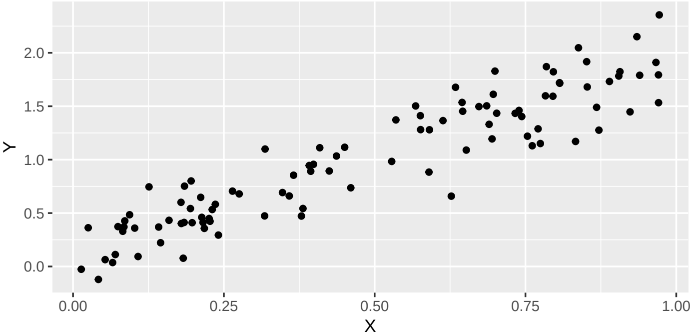
- Here, ϵ represents a random or unobserved error term

The overarching goal of stat learning is to estimate f, given data on X and Y.

An Example



An Example




```
X = runif(100, 0,1 )
E = rnorm(100, 0, .25)
Y = 2*X + E

df<-data.frame(X,Y)
```

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

> *Suppose for each Reed faculty, we have year undergrad degree was awarded X and want to predict age Y.*
>
> *We wish to create a model f that takes in X as input and outputs our best guess $\hat{Y}$ for Y.*

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

Suppose for each Reed faculty, we have year undergrad degree was awarded X and want to predict age Y.

We wish to create a model f that takes in X as input and outputs our best guess $\hat{Y}$ for Y.

- Note that even if we have a perfect estimate for f in $Y = f(X) + \epsilon$, the predicted value $\hat{Y} = f(X)$ of Y may not equal Y, since Y also depends on ϵ

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

Suppose for each Reed faculty, we have year undergrad degree was awarded X and want to predict age Y.

We wish to create a model f that takes in X as input and outputs our best guess $\hat{Y}$ for Y.

- Note that even if we have a perfect estimate for f in $Y = f(X) + \epsilon$, the predicted value $\hat{Y} = f(X)$ of Y may not equal Y, since Y also depends on ϵ
- Thus, there are two sources of error in our model:

1. Reducible error, in the form of our estimate $\hat{f}$ for f.

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

Suppose for each Reed faculty, we have year undergrad degree was awarded X and want to predict age Y.

We wish to create a model f that takes in X as input and outputs our best guess $\hat{Y}$ for Y.

- Note that even if we have a perfect estimate for f in $Y = f(X) + \epsilon$, the predicted value $\hat{Y} = f(X)$ of Y may not equal Y, since Y also depends on ϵ
- Thus, there are two sources of error in our model:

1. Reducible error, in the form of our estimate $\hat{f}$ for f.
2. Irreducible error, in the form of ϵ

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

Suppose for each Reed faculty, we have year undergrad degree was awarded X and want to predict age Y.

We wish to create a model f that takes in X as input and outputs our best guess $\hat{Y}$ for Y.

- Note that even if we have a perfect estimate for f in $Y = f(X) + \epsilon$, the predicted value $\hat{Y} = f(X)$ of Y may not equal Y, since Y also depends on ϵ
- Thus, there are two sources of error in our model:

1. Reducible error, in the form of our estimate $\hat{f}$ for f.
2. Irreducible error, in the form of ϵ

We study techniques to minimize error of the first type

Inference

In other settings, we are more interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response.

Inference

In other settings, we are more interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response.

1. Which predictors are likely associated with response?

Inference

In other settings, we are more interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response.

1. Which predictors are likely associated with response?
2. What is the degree and strength of the relationship between signficant predictors and the response?

Inference

In other settings, we are more interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response.

1. Which predictors are likely associated with response?
2. What is the degree and strength of the relationship between signficant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Logistic? Something more complicated?)

Inference

In other settings, we are more interested in the relationship between each predictor $X_1, \dots, X_p$ and the response.

1. Which predictors are likely associated with response?
2. What is the degree and strength of the relationship between signficant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Logistic? Something more complicated?)

Ex:

A data set contains information on a professor's age, gender, tenure-status, ethnicity, and department. Which of these predictors are associated with course evaluation scores, and how?

Inference

In other settings, we are more interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response.

1. Which predictors are likely associated with response?
2. What is the degree and strength of the relationship between signficant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Logistic? Something more complicated?)

Ex:

A data set contains information on a professor's age, gender, tenure-status, ethnicity, and department. Which of these predictors are associated with course evaluation scores, and how?

Here, we are trying to **infer** information about the factors which contribute to course eval score.

Section 3

Methods of Stat Learning

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.

- The linear model is a common choice for the shape of f:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.

- The linear model is a common choice for the shape of f:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

2. After a model has been chosen, we implement a procedure for estimating the **parameters** of the model that minimizes the reducible error.

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.
   - The linear model is a common choice for the shape of f:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

2. After a model has been chosen, we implement a procedure for estimating the **parameters** of the model that minimizes the reducible error.
   - In the case of the linear model, we estimate the values of $\beta_0, \ldots, \beta_p$ using the *method of least squares*.

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions

- In doing so, non-parametric models avoid the problem of mis-characterizing the relationship between predictors and response

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions

- In doing so, non-parametric models avoid the problem of mis-characterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions

- In doing so, non-parametric models avoid the problem of mis-characterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables
- Non-parametric models often require orders of magnitute more data to make accurate predictions, compared to parametric models

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions

- In doing so, non-parametric models avoid the problem of mis-characterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables
- Non-parametric models often require orders of magnitute more data to make accurate predictions, compared to parametric models
- Some examples of non-parametric models include: Spline Regression, Support Vector Machines, and Neural Networks

Problem Types

Most statistical learning **techniques** fall into one of two categories:

Problem Types

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables

Problem Types

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable

Problem Types

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable

Statistical learning **problems** also fall into a pair of categories:

Problem Types

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable

Statistical learning **problems** also fall into a pair of categories:

1. Regression problems, wherein we measure the magnitude of a **quantitative** response variable

Problem Types

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable

Statistical learning **problems** also fall into a pair of categories:

1. Regression problems, wherein we measure the magnitude of a **quantitative** response variable
2. Classification problems, wherein we sort a **qualitative** response variable into several discrete classes.

Section 4

Guess My Age

The Task

1. Open a new .Rmd file in RStudio and import the data set from Monday's class, available on the course webpage:

https://reed-stat-learning-fall-2020.github.io/data/how_old.csv

2. Explore the data using ggplot
3. Mutate the data set using `dplyr` verbs to assess each groups accuracy. Which group seemed to have the most accurate predictions?
4. Which faculty member's age predictions seemed to be the most (and least) variables?