Foundations of Statistical Learning

Prof Wells

STA 295: Stat Learning

January 30th, 2024

Outline

In today's class, we will. . .

Outline

In today's class, we will. . .

- Discuss the goals of statistical learning algorithms
- Survey some of the most common methods for statistical learning
- Practice coding in R

Section 1

Foundations of Stat Learning

Review

- Stat learning is collection of tools to understand data

Review

- Stat learning is collection of tools to understand data
- Tools are often divided into *supervised* and *unsupervised* methods.

Review

- Stat learning is collection of tools to understand data
- Tools are often divided into *supervised* and *unsupervised* methods.
    - **Supervised**: involve building models to predict the value of one or more output variables
    - **Unsupervised**: seek to learn about relationships and structure within data
- Supervising learning is divided into two tasks, depending on output variable's type:

Review

- Stat learning is collection of tools to understand data
- Tools are often divided into *supervised* and *unsupervised* methods.
  - **Supervised**: involve building models to predict the value of one or more output variables
  - **Unsupervised**: seek to learn about relationships and structure within data
- Supervising learning is divided into two tasks, depending on output variable's type:
  - **Regression**: Predicting (estimating) numeric value of quantitative variables
  - **Classification**: Predicting (classifying) qualitative level of categorical variables

Supervised Learning

- Supervised methods study of the relationships between **predictor variables** $X_1, \ldots, X_p$ for a population, and one or more **response variables** $Y_1, Y_2, \ldots$.

Supervised Learning

- Supervised methods study of the relationships between **predictor variables** $X_1, \ldots, X_p$ for a population, and one or more **response variables** $Y_1, Y_2, \ldots$.
  - **Ex:** For a newly published book, given its genre (X_1) , page count (X_2), whether author has previously published (X_3), and author visibility (X_4), predict how many copies (Y_1) of the book will be sold 1 year after publication, as well as number of consecutive weeks (Y_2) book will sell at least 100 units.

Supervised Learning

- Supervised methods study of the relationships between **predictor variables** $X_1, \ldots, X_p$ for a population, and one or more **response variables** $Y_1, Y_2, \ldots$.
  - **Ex:** For a newly published book, given its genre (X_1) , page count (X_2), whether author has previously published (X_3), and author visibility (X_4), predict how many copies (Y_1) of the book will be sold 1 year after publication, as well as number of consecutive weeks (Y_2) book will sell at least 100 units.
- In the simplest regression tasks, we observe the values of *one quantitative* response Y, as well as p many predictors $X_1, \ldots, X_p$ (these may be quantitative or categorical)

Supervised Learning

- Supervised methods study of the relationships between **predictor variables** $X_1, \ldots, X_p$ for a population, and one or more **response variables** $Y_1, Y_2, \ldots$.
  - **Ex:** For a newly published book, given its genre (X_1) , page count (X_2), whether author has previously published (X_3), and author visibility (X_4), predict how many copies (Y_1) of the book will be sold 1 year after publication, as well as number of consecutive weeks (Y_2) book will sell at least 100 units.
- In the simplest regression tasks, we observe the values of *one quantitative* response Y, as well as p many predictors $X_1, \ldots, X_p$ (these may be quantitative or categorical)
- We assume there is a certain relationship between response and predictors:

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

Supervised Learning

- Supervised methods study of the relationships between **predictor variables** $X_1, \ldots, X_p$ for a population, and one or more **response variables** $Y_1, Y_2, \ldots$.
  - **Ex:** For a newly published book, given its genre (X_1) , page count (X_2), whether author has previously published (X_3), and author visibility (X_4), predict how many copies (Y_1) of the book will be sold 1 year after publication, as well as number of consecutive weeks (Y_2) book will sell at least 100 units.
- In the simplest regression tasks, we observe the values of *one quantitative* response Y, as well as p many predictors $X_1, \ldots, X_p$ (these may be quantitative or categorical)
- We assume there is a certain relationship between response and predictors:

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

  - The function f is called the **model** or **regression function** and the random variable ϵ is the **error** term

Supervised Learning

- Supervised methods study of the relationships between **predictor variables** $X_1, \ldots, X_p$ for a population, and one or more **response variables** $Y_1, Y_2, \ldots$.
    - **Ex:** For a newly published book, given its genre (X_1) , page count (X_2), whether author has previously published (X_3), and author visibility (X_4), predict how many copies (Y_1) of the book will be sold 1 year after publication, as well as number of consecutive weeks (Y_2) book will sell at least 100 units.
- In the simplest regression tasks, we observe the values of *one quantitative* response Y, as well as p many predictors $X_1, \ldots, X_p$ (these may be quantitative or categorical)
- We assume there is a certain relationship between response and predictors:

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

    - The function f is called the **model** or **regression function** and the random variable ϵ is the **error** term
    - The function f represents our best estimate of the value of Y given X, or the expected value of Y given X

Estimating f

- In practice, we will never know the **true** formula for f.

Estimating f

- In practice, we will never know the **true** formula for f.
  - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.

Estimating f

- In practice, we will never know the **true** formula for f.
  - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.
  - Our estimate for the the true model f is called the **fitted** model $\hat{f}$.

Estimating f

- In practice, we will never know the **true** formula for f.
    - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.
    - Our estimate for the the true model f is called the **fitted** model $\hat{f}$.
- How we estimate f will depend on our research goals.

Estimating f

- In practice, we will never know the **true** formula for f.
    - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.
    - Our estimate for the the true model f is called the **fitted** model $\hat{f}$.
- How we estimate f will depend on our research goals.

1. Make **predictions** about the values of Y using $X_1, \ldots, X_p$

Estimating f

- In practice, we will never know the **true** formula for f.
  - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.
  - Our estimate for the the true model f is called the **fitted** model $\hat{f}$.
- How we estimate f will depend on our research goals.

1. Make **predictions** about the values of Y using $X_1, \ldots, X_p$
   - Very interested in finding $\hat{f}$ that makes accurate predictions for Y
   - Less interested in the learning the true form of f
   - STA 295: Statistical Learning

Estimating f

- In practice, we will never know the **true** formula for f.
  - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.
  - Our estimate for the the true model f is called the **fitted** model $\hat{f}$.
- How we estimate f will depend on our research goals.

1. Make **predictions** about the values of Y using $X_1, \ldots, X_p$
   - Very interested in finding $\hat{f}$ that makes accurate predictions for Y
   - Less interested in the learning the true form of f
   - STA 295: Statistical Learning
2. Make **inferences** about relationship between Y and $X_1, \ldots, X_p$

Estimating f

- In practice, we will never know the **true** formula for f.
  - Goal of stat learning is to estimate f, given sample data for $X_1, \ldots, X_p$ and Y.
  - Our estimate for the the true model f is called the **fitted** model $\hat{f}$.
- How we estimate f will depend on our research goals.

1. Make **predictions** about the values of Y using $X_1, \ldots, X_p$
   - Very interested in finding $\hat{f}$ that makes accurate predictions for Y
   - Less interested in the learning the true form of f
   - STA 295: Statistical Learning
2. Make **inferences** about relationship between Y and $X_1, \ldots, X_p$
   - Very interested in learning the true form of f
   - Less interested in finding $\hat{f}$ that makes accurate predictions for Y
   - STA 310: Statistical Modeling

An Example

Consider two quantitative variables X and Y

An Example

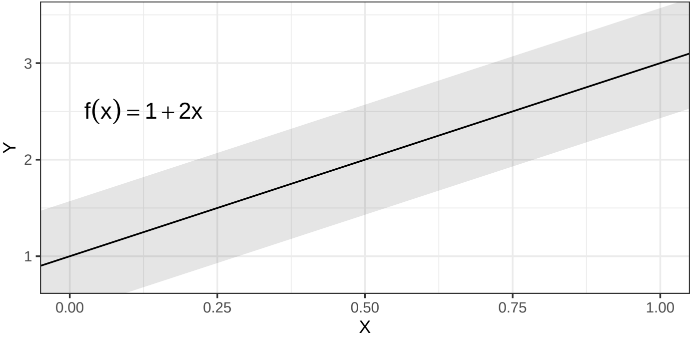
Consider two quantitative variables X and Y

- Suppose, in truth, $Y = 1 + 2X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma = 0.25)$.

An Example

Consider two quantitative variables X and Y

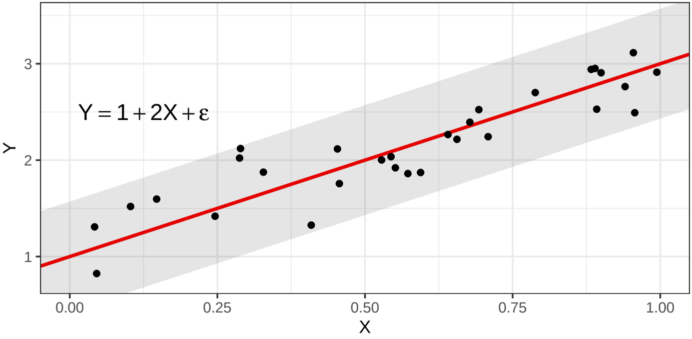
- Suppose, in truth, $Y = 1 + 2X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma = 0.25)$.
- The true model is $f(x) = 1 + 2x$.



An Example

Consider two quantitative variables X and Y

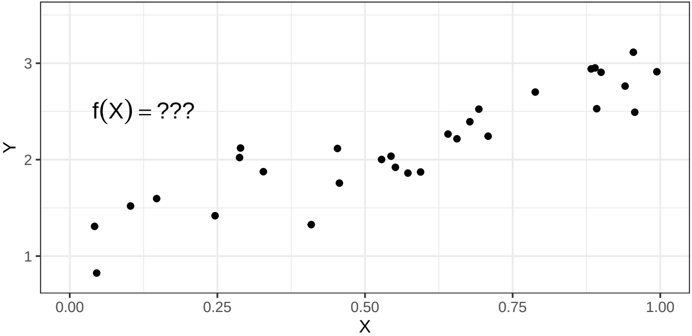
- Suppose, in truth, $Y = 1 + 2X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma = 0.25)$.
- But data Y will not always lie on this line:



An Example

Consider two quantitative variables X and Y

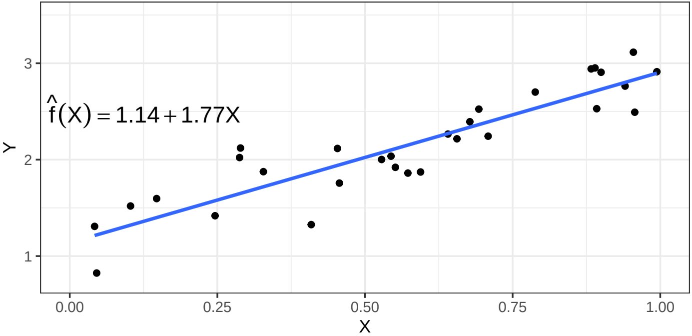
- In reality, we won't know the true model.
- We only have the observed data



An Example

Consider two quantitative variables X and Y

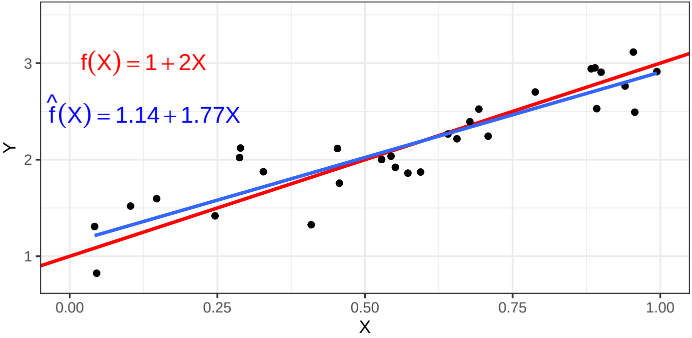
- Instead, we create an estimate $\hat{f}$ based on data
- Here, we use least squares regression to estimate f



An Example

Consider two quantitative variables X and Y

- Our estimated model is $\hat{f}(x) = 1.14 + 1.77x$
- Which is close to the true model of $f(x) = 1 + 2x$



Estimating f for Prediction

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

Estimating f for Prediction

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

- Suppose, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.

Estimating f for Prediction

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

- Suppose, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.
- Ultimately, I want to estimate the final grade Y for each student.

Estimating f for Prediction

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

- Suppose, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.
- Ultimately, I want to estimate the final grade Y for each student.
    - Why is it useful to predict Y?

Estimating f for Prediction

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

- Suppose, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.
- Ultimately, I want to estimate the final grade Y for each student.
  - Why is it useful to predict Y?
- To do so, we theorize a model f that takes in X_1, X_2, X_3 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, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.
- Ultimately, I want to estimate the final grade Y for each student.
    - Why is it useful to predict Y?
- To do so, we theorize a model f that takes in X_1, X_2, X_3 as input and outputs our best guess $\hat{Y}$ for Y.
    - What is one such possible model f?

Estimating f for Prediction

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

- Suppose, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.
- Ultimately, I want to estimate the final grade Y for each student.
  - Why is it useful to predict Y?
- To do so, we theorize a model f that takes in X_1, X_2, X_3 as input and outputs our best guess $\hat{Y}$ for Y.
  - What is one such possible model f?
- But 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, after 8 weeks of the semester, I have collected the following grading information from each student: a midterm exam score X_1, a homework average for the first 8 weeks X_2, and a note indicating whether the student has ever been absent X_3.
- Ultimately, I want to estimate the final grade Y for each student.
  - Why is it useful to predict Y?
- To do so, we theorize a model f that takes in X_1, X_2, X_3 as input and outputs our best guess $\hat{Y}$ for Y.
  - What is one such possible model f?
- But 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.
  - What are some sources of error ϵ in the previous model?

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

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

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

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

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

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

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

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

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

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

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

What steps can be taken to improve reducible error?

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

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

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

What steps can be taken to improve reducible error?

What about irreducible error?

Inference

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

Inference

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

1. Which predictors are likely to be associated with response?

Inference

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

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

Inference

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

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

Inference

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

1. Which predictors are likely to be associated with response?
2. What is the degree and strength of the relationship between significant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Exponential? 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 many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

1. Which predictors are likely to be associated with response?
2. What is the degree and strength of the relationship between significant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Exponential? 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.

Parametric Methods

Parametric methods for estimating f involve two steps:

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

Parametric Methods

Parametric methods for estimating f involve two steps:

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

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

$$\begin{aligned} f(X_1) =& \beta_0 + \beta_1 X_1 \quad \text{simple linear} \\ f(X_1, \ldots, X_p) =& \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{multilinear} \end{aligned}$$

Parametric Methods

Parametric methods for estimating f involve two steps:

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

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

$$f(X_1) = \beta_0 + \beta_1 X_1 \quad \text{simple linear}$$
$$f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{multilinear}$$

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 functional form or shape of f.

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

$$f(X_1) = \beta_0 + \beta_1 X_1 \quad \text{simple linear}$$
$$f(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{multilinear}$$

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*.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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 mischaracterizing 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 mischaracterizing 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 mischaracterizing 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
  - How is this possible?

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 mischaracterizing 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
  - How is this possible?
- Non-parametric models often require orders of magnitude 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 mischaracterizing 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
  - How is this possible?
- Non-parametric models often require orders of magnitude more data to make accurate predictions, compared to parametric models
- Some examples of non-parametric models include: K Nearest Neighbors, Spline Regression, Support Vector Machines, and Neural Networks