DATA 300
Statistical Machine Learning

Fall 2022

Chapter 2: Intro to Statistical Learning

Agenda (Chapter 2 in ISLR)

- **Supervised learning vs unsupervised learning**
- The goal of supervised learning
- Model assessment in regression

Statistical Learning

What is the relationship between *years of education* and *income*?

Statistical Learning

What is the relationship between *years of education* and *income*?

e.g., Income = 5k * years of education + unaccounted error

Statistical learning is the process of finding an appropriate functional form to represent the relationship among concepts (variables).

Refreshing on definitions

- A ***unit*** or ***object*** is an item we observe. When the unit is a person, we refer to the unit as a ***subject***.

- An ***observation*** is a piece of information or characteristic recorded for each unit.

- A characteristic that can vary from unit to unit is called a ***variable***.

- In most datasets, every row is often an observation, and every column is often a variable.

Refreshing on definitions

- **Predictors** (independent variable, feature) are the variables used to predict a response.
- **Response** (dependent variable) is the variable being predicted.

Refreshing on definitions

- **Predictors** (independent variable, feature) are the variables used to predict a response. E.g., years of education
- **Response** (dependent variable) is the variable being predicted. E.g., income

Income = 5k * years of education + unaccounted error

Refreshing on definitions

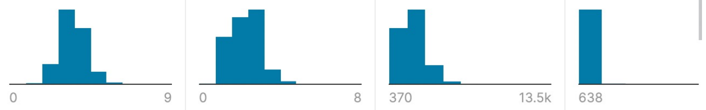
- **Predictors** (independent variables, features) are the variables used to predict a response.
- **Response** (dependent variable) is the variable being predicted.

Identifying predictors and the response requires domain expertise, in other words, the relationship needs to make practical sense in the domain.

Refreshing on definitions

- **Predictors** (independent variables, features) are the variables used to predict a response.
- **Response** (dependent variable) is the variable being predicted.

Sometimes your data analysis task might not need a **Response** variable from the dataset, e.g.,



# bedrooms	# bathrooms	# sqft_living	# sqft_lot
3.0	1.5	1340	7912
5.0	2.5	3650	9050
3.0	2.0	1930	11947
3.0	2.25	2000	8030
4.0	2.5	1940	10500
2.0	1.0	880	6380
2.0	2.0	1350	2560

What are the houses that are similar in terms of these four aspects?

Refreshing on definitions

- **Predictors** (independent variables, features) are the variables used to predict a response.
- **Response** (dependent variable) is the variable being predicted.

Sometimes your data analysis task might not need a **Response** variable from the dataset:

- It calls for unsupervised learning models if there is **no** response (major focus in DATA 180).
- Otherwise, the models are called supervised learning models (major focus in DATA 300).

Types of supervised statistical learning

<u>Classification</u> refers to the type of supervised learning models with a binary response variable, for example:

- Is this email a spam or not?
- Is this patient diagnosed with cancer or not?
- Is this picture a cat or not?

<u>Regression</u> refers to the type of supervised learning models with a non-binary response variable, for example:

- Credit card balance of customers.
- Students’ grade from a class.

Agenda (Chapter 2 in ISLR)

- Supervised learning vs unsupervised learning
- **The goal of supervised learning**
- Model assessment in regression

Supervised statistical learning models

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in,$$

where there should be:

- only one response variable y,
- one or multiple predictors **X**.
- f(**X**) stands for some function of **X**.
- ∈ (epsilon) is the error term, standing for the part of the response that can not be explained by **X.**

Supervised statistical learning models

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in,$$

Examples of this functional relationship?

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 1: Why do we need to estimate this function $f(X)$?

- Prediction
  - Knowing this function is the only way to **approximate** the response y whenever we have information on the predictor X.

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 1: Why do we need to estimate this function $f(X)$?

- Prediction
  - Knowing this function is the only way to **approximate** the response y whenever we have information on the predictor X.
  - Why can we only **approximate** (instead of calculating) the response y?

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 1: Why do we need to estimate this function $f(X)$?

- Prediction
  - Knowing this function is the only way to **approximate** the response y whenever we have information on the predictor X.
  - Why can we only approximate the response y?

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} , \end{aligned}$$

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 1: Why do we need to estimate this function $f(X)$?

- Prediction
  - Knowing this function is the only way to **approximate** the response y whenever we have information on the predictor X.
  - Why can we only approximate the response y?

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}, \end{aligned}$$

  - The goal of statistical learning is to find a function to minimize the reducible error.

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 1: Why do we need to estimate this function $f(X)$?

- Prediction
- Inference
  - Sometimes we care about the exact form of this function $f(X)$, as the parameters might help us understand the relationship between X and y.

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 2: How do we estimate this function $f(X)$?

- Step 2.1: What is the form of f(X)?

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 2: How do we estimate this function $f(X)$?

- Step 2.1: What is the form of f(X)?
  - Parametric: make an assumption of the form
  - Non-parametric: does not make assumptions of the form

The goal of supervised learning

Generally speaking, a supervised learning model assumes that there is the following relationship between the predictors **X** and the response y:

$$y = f(X) + \in$$

Step 2: How do we estimate this function $f(X)$?

- Step 2.1: What is the form (equation) of f(X)?
  - Parametric: make an assumption of the form
  - ~~Non-parametric: does not make assumptions of the form (not the focus of this class)~~
- Step 2.2: Estimate the parameters in the assumed form.

Exercise

Think about the difference of focus between the following two tasks:

- Predict price of Apple’s stock in the next month, and
- Analyze what are the factors that have been affecting the stock price for Apple so far.

Exercise

Think about the difference of focus between the following two tasks:

- Predict price of Apple's stock in the next month, and
- Analyze what are the factors that have been affecting the stock price for Apple so far.

Why succeeding one task does not mean you can succeed in the other?

Parameter estimation: the trade-off between accuracy and model interpretability

There are always two types of tasks in supervised machine learning:

- Prediction (to predict the response **y for out-of-sample** units)
- Interpretation (to explain the relationship between **X** and **y using the sample**)

Next, we measure the **quality of a model** with these two tasks in mind.

Agenda (Chapter 2 in ISLR)

- Supervised learning vs unsupervised learning
- The goal of supervised learning
- **Model assessment in regression**

Exercise - binary

Assuming the response variable is whether a customer used a coupon in its transaction or not. y = 1 means yes, y = 0 means no. think of a few ways to measure the performance of the following model:

	True coupon usage	Model predicted usage
Customer 1	1	1
Customer 2	0	1
Customer 3	1	0
Customer 4	1	1
Customer 5	0	1

Exercise – non-binary

Assuming the response variable is customers’ monthly expenditure, think of a few ways to measure the performance of the following model:

	True expenditure	Model predicted expenditure
Customer 1	$100	$60
Customer 2	$120	$200
Customer 3	$40	$50
Customer 4	$10	$0
Customer 5	$80	$100

Assessing model accuracy: quality of fit

There are a lot of different ways to measure the quality of fit of a model, but they are all about comparing the *true response variable y* and the *predicted response variable* $\hat{y}$.

Assessing model accuracy: quality of fit

There are a lot of different ways to measure the quality of fit of a model, but they are all about comparing the *true response variable y* and the *predicted response variable* $\hat{y}$.

In classification, this is measured by accuracy and accuracy-related measures (will discuss later in the semester).

Assessing model accuracy: quality of fit

There are a lot of different ways to measure the quality of fit of a model, but they are all about comparing the *true response variable y* and the *predicted response variable* $\hat{y}$.

In regression, this is often measured by different Mean _____ Error:

- Mean Squared Error: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$
- Mean Absolute Error
- …

Assessing model accuracy: quality of fit

There are a lot of different ways to measure the quality of fit of a model, but they are all about comparing the *true response variable y* and the *predicted response variable* $\hat{y}$.

In regression, this is often measured by different Mean _____ Error:

- **Mean Squared Error:** $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$
- Mean Absolute Error
- …

Exercise – minimize MSE

To minimize MSE, we are trying to solve the following objective function:

$$\min \ E\left(y_0 - \hat{f}(x_0)\right)^2 :$$

Expand the function above.

Assessing model accuracy: bias-variance trade-off

MSE can be decomposed to

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

In other words, there are three components in MSE:

- The variance of the model
- The bias of the model.
- Irreducible variance that cannot be controlled by the model.

Assessing model accuracy: bias-variance trade-off

MSE can be decomposed to

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

In other words, there are three components in MSE:

- The variance of the model
  - The amount of change in the model when we change the training set.
- The bias of the model
  - The error introduced by using a model to approximate a real-life problem.
- Irreducible variance that cannot be controlled by the model.

Exercise

MSE can be decomposed to

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

In other words, there are three components in MSE:

- The variance of the model
  - The amount of change in the model when we change the training set.
- The bias of the model
  - The error introduced by using a model to approximate a real-life problem.
- Irreducible variance that cannot be controlled by the model.

Think practically: if a model is simple (linear regression), what tends to happen for variance and bias?

What about a more complicated model?

Assessing model accuracy: bias-variance trade-off

MSE can be decomposed to

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

In other words, there are three components in MSE:

- The variance of the model
  - The amount of change in the model when we change the training set.
  - Tend to be low if the model is simple and less flexible.
- The bias of the model
  - The error introduced by using a model to approximate a real-life problem.
  - Tend to be low if the model is flexible and complicated.
- Irreducible variance that cannot be controlled by the model.

Assessing model accuracy: bias-variance trade-off

MSE can be decomposed to

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

In other words, there are three components in MSE:

- The variance of the model
  - The amount of change in the model when we change the training set.
  - Tend to be low if the model is simple and less flexible.
- The bias of the model
  - The error introduced by using a model to approximate a real-life problem.
  - Tend to be low if the model is flexible and complicated.
- Irreducible variance that cannot be controlled by the model.

Hence, it is challenging to find a model that can reduce variance and bias at the same time.