NBA 4920/6921 Lecture 25

Final Review

Murat Unal

Johnson Graduate School of Management

12/07/2021

Agenda

Machine learning model

Type of models
- Parametric models
- Non-parametric models

Exploratory Data Analysis (EDA)

Inference
- Linear regression

Supervised learning

The goal is to build a model that captures the relationship between Y and X using a function f

$$Y = f(X)$$

$\rightsquigarrow$ Regression (linear, logistic, trees)

Machine learning model

Goal: Build a model to understand `Sales` as a function of advertisement spent in different media.

Output/Target/Response/Dependent Variable:
$Y =$ `Sales`

Input/Feature/Predictor/Explanatory/Independent Variable:
$X =$ (`TV`,`Radio`,`Newspaper`)

Machine learning model

The relationship between output Y and p inputs, $X = (X_1, \cdots, X_p)$, can be written as

$$Y = f(X) + \epsilon$$

f is an unknown function we want to learn/estimate

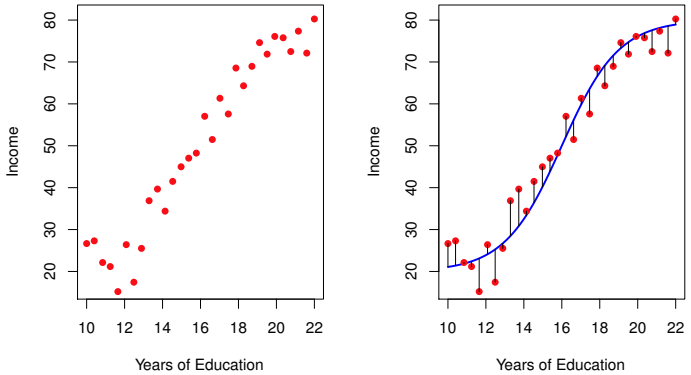
It represents the **systematic** information that X provides about Y

ϵ is a mean-zero error term that is independent of the inputs

It represents the **noise/randomness/unobservables** that can not be explained using X

Machine learning model

The blue curve is the true underlying relationship we want to learn



Source: ISL

What can we use $\hat{f}$ for?

$$\texttt{Sales} = \hat{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Using the observed data we learn/estimate f and obtain $\hat{f}$ for two main purposes.

What can we use $\hat{f}$ for?

$$\texttt{Sales} = \hat{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Using the observed data we learn/estimate f and obtain $\hat{f}$ for two main purposes.

1. **Inference**: Is higher advertising expenditure associated with higher sales? Which media contributes more?

What can we use $\hat{f}$ for?

$$\texttt{Sales} = \hat{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Using the observed data we learn/estimate f and obtain $\hat{f}$ for two main purposes.

1. **Inference**: Is higher advertising expenditure associated with higher sales? Which media contributes more?

   Q: Can we make causal claims? Does advertising increase sales?

What can we use $\hat{f}$ for?

$$\texttt{Sales} = \hat{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Using the observed data we learn/estimate f and obtain $\hat{f}$ for two main purposes.

1. **Inference**: Is higher advertising expenditure associated with higher sales? Which media contributes more?

   Q: Can we make causal claims? Does advertising increase sales?

   A: With observational studies *usually* we can not make causal claims.

   Association $\neq$ Causation.

   Econometrics is the field that studies methods for causal inference in observational settings.

What can we use $\hat{f}$ for?

$$\texttt{Sales} = \hat{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Using the observed data we learn/estimate f and obtain $\hat{f}$ for two main purposes.

2 **Prediction**: Predict sales from advertising expenditure.

How do we estimate f

Assume we have observed a set of n different data points, which are called the **training data**.

We use these observations to train a statistical learning method how to estimate the unknown function f

i.e. we want to find a function $\hat{f}$ s.t. $Y \approx \hat{f}(X)$

Most methods for this task can be characterized as either: **parametric** or **non-parametric**

Parametric models

First assumes a functional form of f then uses the training data to train/fit the model.

$\rightsquigarrow$ The **linear model**: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

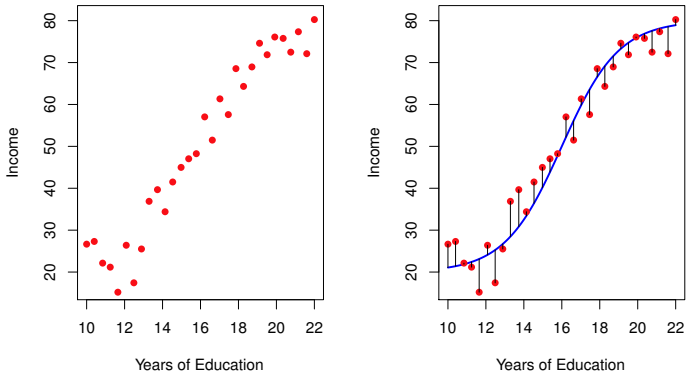
$\rightsquigarrow$ We can estimate the parameters $\beta_0, \beta_1, \beta_2, \cdots . \beta_p$ using **ordinary least squares** (OLS)

Pro: Easy to estimate

Con: Less flexible. Can be a poor approximation for the true unknown form of f

Parametric models

What do you think would happen if we estimated the true blue curve with a liner parametric model, such as OLS?



Source: ISL

Non-parametric models

Do not make explicit assumptions about the functional form of f

$\rightsquigarrow$ regression trees, random forests

Pro: Increased flexibility. Can be a good approximation for the true unknown form of f

Con: Far more observations is required in order to obtain an accurate estimate for f

Exploratory Data Analysis (EDA)

EDA is an important part of any data analysis. Use EDA to:

1. Generate questions about your data
2. Search for answers by visualizing, transforming, and/or modeling your data
3. Use what you learn to refine your questions and/or generate new questions

Exploratory Data Analysis (EDA)

Key concepts:

1. Summary statistics
2. Plots for visulaization
3. Variation
4. Covariation

Linear regression

Key concepts:

1. RSS, TSS, R^2
2. Hypothesis testing, null hypothesis (H_0), alternative hypothesis (H_A)
3. F-test, t-test, confidence interval
4. Covariation

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017)
An Introduction to Statistical Learning
Springer.
https://www.statlearning.com/

Andriy Burkov (2021)
The Hundred-Page Machine Learning Book
http://themlbook.com

Ed Rubin (2020)
Economics 524 (424): Prediction and Machine-Learning in Econometrics
Univ, of Oregon.