Big Data and Machine Learning

Juncheng Jiang & Zhilin Zhou

The 11th Zhongyuan Lectures
School of Public Finance and Taxation
Zhongnan University of Economics and Law

May 23, 2024

Foreword

- We are also new to this topic, especially some points in machine learning.
- All the groups (including ours) did not actively choose this topic. Thus, we decided to try to challenge this task.
- We integrated and restated the relevant concepts and principles, thanks to Professor Raj Chetty's courses, Professor Chen's videos, the Tsinghua University seminar and other references.
- We are also particularly grateful for the guidance provided by the teachers in our training camp.
- We try our best to make the presentation interesting and not be a recitation of slides.
- All errors are ours.

Outline

What is Big Data?

What is Big Data?

We often hear about **data** and **big data**, but here are two questions worthy of thinking about:

How to explain what is **data**?

How to understand the meaning of ***BIG*** in the word of **"big data"**?

Data & Big data

DATA(in economics)

- Def: information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer
- Types: **Cross-Sectional Data, Time Series Data, Pooled Cross Sections, Panel Data**

BIG DATA

- Def: extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions

More about *Big Data*

- Perhaps everyone has heard some aspects of big data, especially **its famous features: "5V" characteristics**.

More about *Big Data*

- Perhaps everyone has heard some aspects of big data, especially **its famous features: "5V" characteristics**.
  - Volume（大量）:
    Large data volume and the measurement unit is above P (kT)

More about *Big Data*

- Perhaps everyone has heard some aspects of big data, especially **its famous features: "5V" characteristics**.
  - Volume（大量）:
    Large data volume and the measurement unit is above P (kT)
  - Velocity （高速）:
    Fast data growth rate & Fast data processing speed with HPC (High Performance Computing)

More about *Big Data*

- Perhaps everyone has heard some aspects of big data, especially **its famous features: "5V" characteristics**.
  - Volume（大量）:
    Large data volume and the measurement unit is above P (kT)
  - Velocity （高速）:
    Fast data growth rate & Fast data processing speed with HPC (High Performance Computing)
  - Variety （多样）:
    Diversity of data types and sources (I will talk about in the next slide)

More about *Big Data*

- Perhaps everyone has heard some aspects of big data, especially **its famous features: "5V" characteristics**.
  - Volume（大量）:
    Large data volume and the measurement unit is above P (kT)
  - Velocity （高速）:
    Fast data growth rate & Fast data processing speed with HPC (High Performance Computing)
  - Variety （多样）:
    Diversity of data types and sources (I will talk about in the next slide)
  - Value （低价值密度）:
    Sounds abnormal but makes sense! How to analyze and identify the significant value of the data is just the direction of machine learning and artificial intelligence efforts

More about *Big Data*

- Perhaps everyone has heard some aspects of big data, especially **its famous features: "5V" characteristics**.
  - Volume（大量）:
    Large data volume and the measurement unit is above P (kT)
  - Velocity （高速）:
    Fast data growth rate & Fast data processing speed with HPC (High Performance Computing)
  - Variety （多样）:
    Diversity of data types and sources (I will talk about in the next slide)
  - Value （低价值密度）:
    Sounds abnormal but makes sense! How to analyze and identify the significant value of the data is just the direction of machine learning and artificial intelligence efforts
  - Veracity （真实）:
    This refers to the quality of big data. Its content is closely related to the real world. Reality does not necessarily represent accuracy, but it is not false data. This is also the foundation of data analysis.

Big data

Let's talk about the classification mentioned before:

Structured

The storage and arrangement of data are very regular, with certain rules.

Semi-Structured

It is a form of structured data that has a format but no fixed data model. It contains relevant tags to separate semantic elements and layer records and fields. eg: XML/Html

Unstructured

Without a fixed structure. Various documents, images, videos, audio, etc.

Meta Data

Metadata is structured reference data that helps to sort and identify attributes of the information it describes.

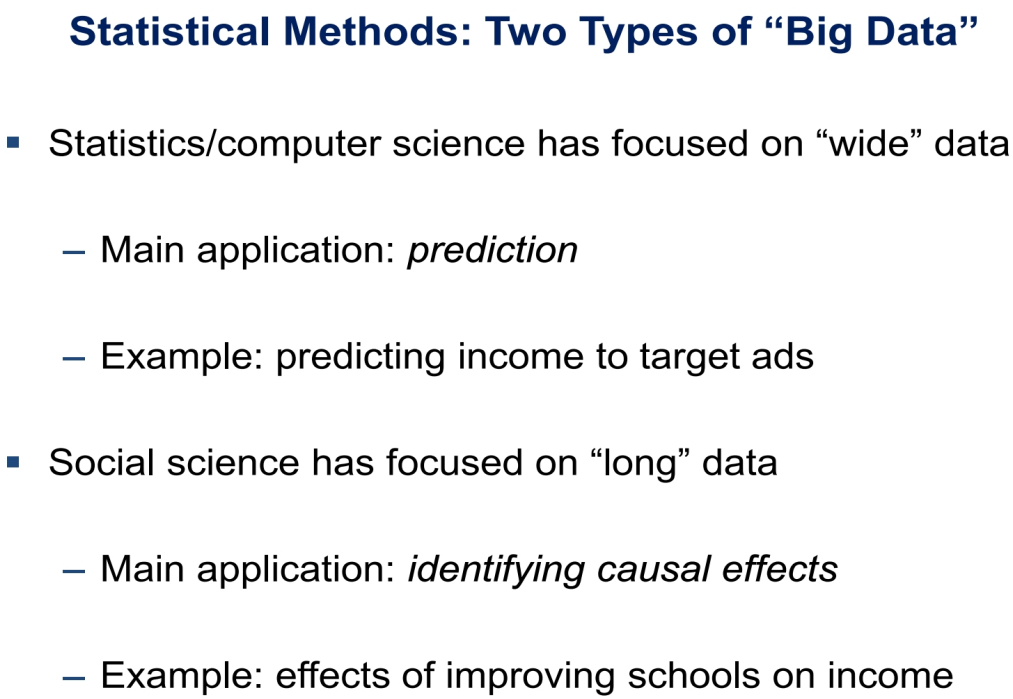
Extension

Raj Chetty offered a course for Harvard students in 2019 called "**Using Big Data to Solve Economic and Social Problems**"

In his first lesson, he gave a simpler classification of Big Data like this:

– "Long" data: many observations relative to variables (e.g. tax records)

– "Wide" data: few observations relative to variables (e.g. Amazon clicks, newspapers)



Thinking more

In the slides of Chetty(2019), he wrote down the words "Social Science in the Age of Big Data", and what is its meaning?

Social Science in the Age of Big Data

- Recent availability of "big data" has accelerated this trend
  - Large datasets are starting to transform social science, as they have transformed business
- Examples:
  - Government data: tax records, Medicare
  - Corporate data: Google, Uber, retailer data
  - Unstructured data: Twitter, newspapers

Thinking more

The next question is"Why is Big Data Transforming Social Science?"

He gave us four reasons:

- 1. Greater reliability than surveys
- 2. Ability to measure new variables (e.g., emotions)
- 3. Universal coverage $\Rightarrow$ can "zoom in" to subgroups
- 4. Large samples $\Rightarrow$ can approximate scientific experiments

Reviewing some examples of Statistical Methods

On this page, I still copied Chetty's. The original title is what you will learn in his course.

But I think that it also tells the basic methods we usually use in economics.

- 1.Descriptive Data Analysis: correlation, regression, survival analysis

Reviewing some examples of Statistical Methods

On this page, I still copied Chetty's. The original title is what you will learn in his course.

But I think that it also tells the basic methods we usually use in economics.

- 1.Descriptive Data Analysis: correlation, regression, survival analysis
- 2.Experiments: randomization, non-compliance

Reviewing some examples of Statistical Methods

On this page, I still copied Chetty's. The original title is what you will learn in his course.

But I think that it also tells the basic methods we usually use in economics.

- 1.Descriptive Data Analysis: correlation, regression, survival analysis
- 2.Experiments: randomization, non-compliance
- 3.Quasi-Experiments: regression discontinuity, difference-in-differences

Reviewing some examples of Statistical Methods

On this page, I still copied Chetty's. The original title is what you will learn in his course.

But I think that it also tells the basic methods we usually use in economics.

- 1.Descriptive Data Analysis: correlation, regression, survival analysis
- 2.Experiments: randomization, non-compliance
- 3.Quasi-Experiments: regression discontinuity, difference-in-differences
- 4.Machine Learning: prediction, overfitting, cross-validation

Reviewing some examples of Statistical Methods

On this page, I still copied Chetty's. The original title is what you will learn in his course.

But I think that it also tells the basic methods we usually use in economics.

- 1.Descriptive Data Analysis: correlation, regression, survival analysis
- 2.Experiments: randomization, non-compliance
- 3.Quasi-Experiments: regression discontinuity, difference-in-differences
- 4.Machine Learning: prediction, overfitting, cross-validation
- 5.Stata (or other) statistical programming language

What is Machine Learning?

Introduction to Machine Learning

Machine learning is a subfield of artificial intelligence (AI) that uses algorithms trained on data sets to create self-learning models that are capable of predicting outcomes and classifying information without human intervention. Machine learning is used today for a wide range of commercial purposes, including suggesting products to consumers based on their past purchases, predicting stock market fluctuations, and translating text from one language to another.

In common usage, the terms "machine learning" and "artificial intelligence" are often used interchangeably with one another due to the prevalence of machine learning for AI purposes in the world today. But, the two terms are meaningfully distinct.
While AI refers to the general attempt to create machines capable of human-like cognitive abilities, machine learning specifically refers to the use of algorithms and data sets to do so.

Simplification

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. –Arthur Samuel

Due to the use of many statistical methods, statisticians also refer to it as **Statistical learning**.

Essentially, it originated from the field of artificial intelligence in computer science.

e.g.: We can tell the spam in our email box, how to make machines do the same?

Hard Coding VS Learning

There is an explanation mentioned in Chen's book (2021).

- The famous case in machine learning is the *spam filtering*, which means how to automatically filter spam without misjudging ham.

Hard Coding VS Learning

There is an explanation mentioned in Chen's book (2021).

- The famous case in machine learning is the *spam filtering*, which means how to automatically filter spam without misjudging ham.
  - Traditionally, we humans share knowledge about spam with computers and program the rules in code.
    This method is called hard coding, and you can imagine that it doesn't work well.

Hard Coding VS Learning

There is an explanation mentioned in Chen's book (2021).

- The famous case in machine learning is the *spam filtering*, which means how to automatically filter spam without misjudging ham.
  - Traditionally, we humans share knowledge about spam with computers and program the rules in code.
    This method is called hard coding, and you can imagine that it doesn't work well.
  - A breakthrough idea is to introduce learning, which means that we don't directly tell computers about spam, but let them make their judgments by learning a large amount of data.

Hard Coding VS Learning

There is an explanation mentioned in Chen's book (2021).

- The famous case in machine learning is the *spam filtering*, which means how to automatically filter spam without misjudging ham.
  - Traditionally, we humans share knowledge about spam with computers and program the rules in code.
    This method is called hard coding, and you can imagine that it doesn't work well.
  - A breakthrough idea is to introduce learning, which means that we don't directly tell computers about spam, but let them make their judgments by learning a large amount of data.
- In the next slide, I will give you a brief introduction to the principle of the approach

The way to achieve

- First, we need to give computers a large number of emails which are labelled "spam" or "ham" before by humans.

The way to achieve

- First, we need to give computers a large number of emails which are labelled "spam" or "ham" before by humans.
- According to these data, the computer can calculate the frequency of different words appearing among spam and normal emails.

The way to achieve

- First, we need to give computers a large number of emails which are labelled "spam" or "ham" before by humans.
- According to these data, the computer can calculate the frequency of different words appearing among spam and normal emails.
- For example, the spam often contains the words of *"factored invoice"*. The computer can take into account other factors and calculate the conditional probability of whether the email which has a "factored invoice" is spam according to the Bayes Rule.

The way to achieve

- First, we need to give computers a large number of emails which are labelled "spam" or "ham" before by humans.
- According to these data, the computer can calculate the frequency of different words appearing among spam and normal emails.
- For example, the spam often contains the words of *"factored invoice"*. The computer can take into account other factors and calculate the conditional probability of whether the email which has a "factored invoice" is spam according to the Bayes Rule.
- If the probability is over a certain threshold (e.g. 0.9), the computer will classify this email as spam and this method is called *Bayes spam filtering*.

The way to achieve

- First, we need to give computers a large number of emails which are labelled "spam" or "ham" before by humans.
- According to these data, the computer can calculate the frequency of different words appearing among spam and normal emails.
- For example, the spam often contains the words of *"factored invoice"*. The computer can take into account other factors and calculate the conditional probability of whether the email which has a "factored invoice" is spam according to the Bayes Rule.
- If the probability is over a certain threshold (e.g. 0.9), the computer will classify this email as spam and this method is called *Bayes spam filtering*.
- The process we talking about above is to make the computer have a certain ability by learning big data. Thus, the way is so-called **"Machine Learning"** and the Bayes spam filtering is one of the "learning machine" or "Learner".

Classification

The example shows us a specific method of **Machine Learning** and we must classify it globally.

- Supervised
- Semi-supervised
- Unsupervised
- Others (like reinforcement learning)

Generally, you can easily find that the keyword of classification above is "SUPERVISED", which means " 有监督的" in Chinese.

Classification

Simply, the "supervised" depends on whether you set a certain goal for learning.
Here is the detailed explanation
link:https://www.coursera.org/articles/what-is-machine-learning

1. Supervised machine learning

In supervised machine learning, algorithms are trained on *labeled* data sets that include tags describing each piece of data. In other words, the algorithms are fed data that includes an "answer key" describing how the data should be interpreted. For example, an algorithm may be fed images of flowers that include tags for each flower type so that it will be able to identify the flower better again when fed a new photograph.

Supervised machine learning is often used to create machine learning models used for prediction and classification purposes.

2. Unsupervised machine learning

Unsupervised machine learning uses *unlabeled* data sets to train algorithms. In this process, the algorithm is fed data that doesn't include tags, which requires it to uncover patterns on its own without any outside guidance. For instance, an algorithm may be fed a large amount of unlabeled user data culled from a social media site in order to identify behavioral trends on the platform.

Unsupervised machine learning is often used by researchers and data scientists to identify patterns within large, unlabeled data sets quickly and efficiently.

Classification

3. Semi-supervised machine learning

Semi-supervised machine learning uses both unlabeled and labeled data sets to train algorithms. Generally, during semi-supervised machine learning, algorithms are first fed a small amount of labeled data to help direct their development and then fed much larger quantities of unlabeled data to complete the model. For example, an algorithm may be fed a smaller quantity of labeled speech data and then trained on a much larger set of unlabeled speech data in order to create a machine learning model capable of speech recognition.

Semi-supervised machine learning is often employed to train algorithms for classification and prediction purposes in the event that large volumes of labeled data is unavailable.

4. Reinforcement learning

Reinforcement learning uses trial and error to train algorithms and create models. During the training process, algorithms operate in specific environments and then are provided with feedback following each outcome. Much like how a child learns, the algorithm slowly begins to acquire an understanding of its environment and begins to optimize actions to achieve particular outcomes. For instance, an algorithm may be optimized by playing successive games of chess, which allows it to learn from its past successes and failures playing each game.

Reinforcement learning is often used to create algorithms that must effectively make sequences of decisions or actions to achieve their aims, such as playing a game or summarizing an entire text.

Extension

- Some of you may have hear about these terms before:
  - Artificial Intelligence
  - Machine Learning
  - Deep Learning

  So, what is their relation ?

  The relationship between **AI>ML>DL**
- And the terms are a little different from what we commonly use:
  - $\mathbf{X}_i$: we usually call it "independent variables(自变量)", "regressors, explanatory variables(解释变量)", "covariates(协变量)", or "control variables(控制变量)" in Statistics and Econometrics.
    In Machine Learning, it is named as "features(特征)", "feature vector(特征向量)", "predictors(预测变量)" or "attributes(属性)".
  - $\mathbf{Y}_i$: Correspondingly, it is "dependent variable (因变量、被解释变量)" or "outcome variable (结果变量)".
    In the latter, it is called "response(响应变量)" or "target(目标)".
  - $\mathbf{i}$: In statistics we call it "observation" or "data point", but in machine learning it is named as "example(样例)" or "instance(示例)"

Other concepts

- In ML, the sample data is always used to train the computer to obtain the learning ability. So it is also called "training data", "training sample" or "training set".

Other concepts

- In ML, the sample data is always used to train the computer to obtain the learning ability. So it is also called "training data", "training sample" or "training set".
- The data is usually classified into two groups.
  The main part of the data is used to train the computer, so we call it "training data". The rest of the minority is used as "test data", also called "validation data" or "hold-on data".

Other concepts

- In ML, the sample data is always used to train the computer to obtain the learning ability. So it is also called "training data", "training sample" or "training set".
- The data is usually classified into two groups.
  The main part of the data is used to train the computer, so we call it "training data". The rest of the minority is used as "test data", also called "validation data" or "hold-on data".

- The test data is only used to verify the effectiveness of machine learning, to avoid overfitting, which means that although the fitting effect within the sample is good, the extrapolation prediction effect is poor.

How many algorithms do we have?

Algorithms

To use *machine learning*, we at least need three elements: Data, Task, Model.

So, I'd like to list some algorithms mainly used. (Each title has a hyperlink for further learning)

Decision Tree Learning Algorithm

A tree-like structure consists of internal nodes and leaf nodes, where internal nodes represent a dimension (feature) and leaf nodes represent a classification.

Many "if...else"

Random Forest Algorithm

Combining multiple decision trees together, each dataset is randomly selected with some features selected as inputs.

The random forest algorithm is a bagging algorithm that uses decision trees as estimators.

Algorithms

Standard Linear Regression Algorithm

Given some random sample points (x_1, y_1), (x_2, y_2).

Find a hyperplane (a straight line for a single variable and a plane for two variables) to fit these sample points.

Just like OLS in two dimension.

Lasso Regression Algorithm

The complete name of LASSO is *the least absolute shrinkage and selection operator* algorithm, which is a regularization cost function method for solving multicollinearity.

This method is a compression estimation. It obtains a more refined model by constructing a penalty function, which compresses some regression coefficients.(Robert Tibshirani, 1996)

Machine Learning & Econometrics

Comparison

The Differences between Econometrics and Machine Learning

- Econometrics: Analysis, Estimation, Hypothesis Testing & a little bit Prediction, actually researchers hardly use econometrics methods to prediction real econ variables.

Comparison

The Differences between Econometrics and Machine Learning

- Econometrics: Analysis, Estimation, Hypothesis Testing & a little bit Prediction, actually researchers hardly use econometrics methods to prediction real econ variables.
- Machine Learning: Prediction.

Comparison

The Differences between Econometrics and Machine Learning

- Econometrics: Analysis, Estimation, Hypothesis Testing & a little bit Prediction, actually researchers hardly use econometrics methods to prediction real econ variables.
- Machine Learning: Prediction.
- Econometrics focus on the effect of explanation, so there are a lot of testing methods.

Comparison

The Differences between Econometrics and Machine Learning

- Econometrics: Analysis, Estimation, Hypothesis Testing & a little bit Prediction, actually researchers hardly use econometrics methods to prediction real econ variables.
- Machine Learning: Prediction.
- Econometrics focus on the effect of explanation, so there are a lot of testing methods.
- Machine Learning only focus on the characteristics of data itself.

Comparison

The Differences between Econometrics and Machine Learning

- Econometrics: Analysis, Estimation, Hypothesis Testing & a little bit Prediction, actually researchers hardly use econometrics methods to prediction real econ variables.
- Machine Learning: Prediction.
- Econometrics focus on the effect of explanation, so there are a lot of testing methods.
- Machine Learning only focus on the characteristics of data itself.
- Econometrics shows the intuition of econ and society.

Comparison

The Differences between Econometrics and Machine Learning

- Econometrics: Analysis, Estimation, Hypothesis Testing & a little bit Prediction, actually researchers hardly use econometrics methods to prediction real econ variables.
- Machine Learning: Prediction.
- Econometrics focus on the effect of explanation, so there are a lot of testing methods.
- Machine Learning only focus on the characteristics of data itself.
- Econometrics shows the intuition of econ and society.
- Machine Learning is more like "Black Box", it only treats on the data, but often has no actual meaning.

Details

The Differences of Linear Regression between these two.

- Over-Fitting and Under-Fitting: High Bias and High Variance is a a widespread problem in econometrics.
- Regularization could solve this problem with a Neural Network Algorithm. Through thousands of simulations, it finds the best polynomial model to describe the data.
- For example: For any normal econometrics model, , it needs the Glejser Test to find the heteroscedasticity, and give the correct regression model for WLS. Usually, we use White Test to check the regression result. But before the WLS, we could use machine learning method to optimize the model, so that we could get larger R Squares.
- Another example: A Cross-Validation Method in R

An Application in R: 10-Fold Cross-Validation

Method

Cross-validation is a statistical method used to **estimate the skill of machine learning models**. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

The general procedure is as follows:

- Shuffle the dataset randomly.
- Split the dataset into k groups. (We often let k=10)
- For each unique group:
  1. Take the group as a test data set;
  2. Take the remaining groups as a training data set;
  3. Fit a model on the training set and evaluate it on the test set;
  4. Retain the evaluation score and discard the model.
- Summarize the skill of the model using the sample of model evaluation scores.
- We focus on the NMSE (Normalized Mean Square Error)

Code

```
#Cross Validation Examples
rm(list=ls(all=TRUE))  #清空所有对象，类似Stata中的clear。
setwd ("D:\\")  #设置路径，类似Stata中的cd。
#构建test data set和training data set
CV=function(n,Z=10,seed=8848){
      #定义函数CV，其中n为样本量，Z为分组数，利用随机种子seed进行随机选择。
  z=rep(1:Z,ceiling(n/Z))[1:n]
        #使用重复函数创建向量Z，其中元素按从1到Z的顺序循环重复，ceiling选项确保Z的长度大于等于n，最
  set.seed(seed)  #设置随机种子，确保每次运行时生成的随机数相同。
  z=sample(z,n)  #对向量Z进行随机的重新排列。
  mm=list()  #创建一个空列表mm，用于存储分组后的样本索引。其中mm[[i]]为第i个分组。
  for (i in 1:Z) mm[[i]]=(1:n)[z==i];return(mm)
        #使用for循环从1遍历至Z，将Z中等于i的索引储存在mm的第i个位置，最后函数返回列表mm。
}
#载入数据
w=read.csv("zhongyuan_lectures.csv")  #定义w为数据框。
n=nrow(w);Z=10;mm=CV(n,Z);D=1  #使用CV函数将数据分为10组，并将每组的索引存储在了mm中。
#计算均方误差MSE
MSE=rep(0,Z)  #创建长为Z的向量MSE并令所有元素为0以储存结果。
for(i in 1:Z){  #使用for循环从1遍历至Z。
  m=mm[[i]]  #将第i组的索引储存在m中。
  M=mean((w[m,D]-mean(w[m,D]))^2)  #求方差M。
  a=lm(y~.,w[-m,])  #构造线性回归，[-m,]为所有的training data set。
  MSE[i]=mean((w[m,D]-predict(a,w[m,]))^2)/M  #先求对拟合值的方差，除以方差M以得到均方误差。
}
NMSE=mean(MSE,na.rm=TRUE)  #对10组每组的均方误差求平均数，忽略缺失值。
print(NMSE)  #输出最后的归一化均方误差NMSE。
```


Thank
you