SOSC 4300/5500: Prediction

Han Zhang

Sep 15, 2020

Outline

Logistics

- Grouping?

Logistics

- Grouping?
- Git basics:

Logistics

- Grouping?
- Git basics:
    - Are everyone able to commit/push at least once?

Logistics

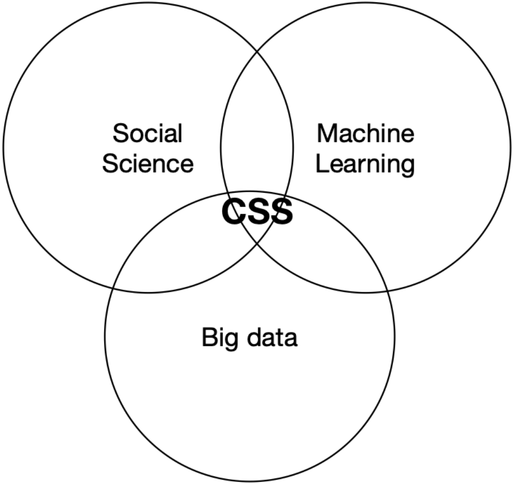
- Grouping?
- Git basics:
  - Are everyone able to commit/push at least once?
- We will be moving to mix-mode teaching from the fourth week

Logistics

- Grouping?
- Git basics:
  - Are everyone able to commit/push at least once?
- We will be moving to mix-mode teaching from the fourth week
  - Wait for further notice

Computational Social Science (CSS)

- Next we focus predictive models, using on machine learning



Prediction vs. Explanation

- Prediction vs explanation
- [In class activities]: Can you give other examples? Type it in chatbox!

Prediction vs. Explanation

- Prediction vs explanation
  - Prediction: Whether Trump of Clinton will win the election?

- [In class activities]: Can you give other examples? Type it in chatbox!

Prediction vs. Explanation

- Prediction vs explanation
  - Prediction: Whether Trump of Clinton will win the election?
  - Explanation: Why Trump won?
- [In class activities]: Can you give other examples? Type it in chatbox!

Prediction vs. Explanation: the ideal case

- Strongest example: classical physics, such as Newton's Law of Motion

Prediction vs. Explanation: the ideal case

- Strongest example: classical physics, such as Newton's Law of Motion
- Predictive: we can precisely predict location of planets in solar system

Prediction vs. Explanation: the ideal case

- Strongest example: classical physics, such as Newton's Law of Motion
- Predictive: we can precisely predict location of planets in solar system
- And explanative: we have a theory on why it's the case

Prediction vs Explanation in Social Sciences

- Social worlds are too complicated to sumarize using several equations

Prediction vs Explanation in Social Sciences

- Social worlds are too complicated to sumarize using several equations
  - We do not have a powerful formula such as $F = ma$

Prediction vs Explanation in Social Sciences

- Social worlds are too complicated to sumarize using several equations
  - We do not have a powerful formula such as $F = ma$
- Current social science research focus dominantly on explanation

Prediction vs Explanation in Social Sciences

- Social worlds are too complicated to sumarize using several equations
  - We do not have a powerful formula such as $F = ma$
- Current social science research focus dominantly on explanation
  - Testing a theory that looks like “A leads to B”

Prediction vs Explanation in Social Sciences

- Social worlds are too complicated to sumarize using several equations
  - We do not have a powerful formula such as $F = ma$
- Current social science research focus dominantly on explanation
  - Testing a theory that looks like "A leads to B"
- But not asking "whether a given theory can predict some outcome of interest"

Failure of theory

- Are our theory really useful?

Failure of theory

- Are our theory really useful?
- Timur Kuran, *Now out of Never: The Element of Surprise in the East European Revolution of 1989*, World Politics **44** (1991), no. 1, 7–48

Failure of theory

- Are our theory really useful?
- Timur Kuran, *Now out of Never: The Element of Surprise in the East European Revolution of 1989*, World Politics **44** (1991), no. 1, 7–48
- In 1987, the American academy of Arts and Sciences invited a dozen of specialist, including several living in Eastern Europe, to prepare interpretive essays on East European developments. . .

Failure of theory

- Are our theory really useful?
- Timur Kuran, *Now out of Never: The Element of Surprise in the East European Revolution of 1989*, World Politics **44** (1991), no. 1, 7–48
- In 1987, the American academy of Arts and Sciences invited a dozen of specialist, including several living in Eastern Europe, to prepare interpretive essays on East European developments. . .
- This was publised in the journal *Daedalus*

Failure of theory

- Are our theory really useful?
- Timur Kuran, *Now out of Never: The Element of Surprise in the East European Revolution of 1989*, World Politics **44** (1991), no. 1, 7–48
- In 1987, the American academy of Arts and Sciences invited a dozen of specialist, including several living in Eastern Europe, to prepare interpretive essays on East European developments. . .
- This was publised in the journal *Daedalus*
- “None forsaw what was to happen”.

Failure of theory

- Rational choice theory

Failure of theory

- Rational choice theory
    - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride
  - So it predicts the lack of revolution

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride
  - So it predicts the lack of revolution
- Structural theory: revolution occurs when the state becomes weaker

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride
  - So it predicts the lack of revolution
- Structural theory: revolution occurs when the state becomes weaker
  - Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, Cambridge University Press, 1979

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride
  - So it predicts the lack of revolution
- Structural theory: revolution occurs when the state becomes weaker
  - Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, Cambridge University Press, 1979
  - Partially gives a prediction

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride
  - So it predicts the lack of revolution
- Structural theory: revolution occurs when the state becomes weaker
  - Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, Cambridge University Press, 1979
  - Partially gives a prediction
  - But there are many countries with weak state power but no revolution

Failure of theory

- Rational choice theory
  - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
  - People has incentive to free ride
  - So it predicts the lack of revolution
- Structural theory: revolution occurs when the state becomes weaker
  - Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, Cambridge University Press, 1979
  - Partially gives a prediction
  - But there are many countries with weak state power but no revolution
  - Eastern European countries were certainly not the countries with the weakest state power then

Failure of theory

- Rational choice theory
    - Mancur Olson, *The Logic of Collective Action*, Harvard University Press, 1965
    - People has incentive to free ride
    - So it predicts the lack of revolution
- Structural theory: revolution occurs when the state becomes weaker
    - Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia and China*, Cambridge University Press, 1979
    - Partially gives a prediction
    - But there are many countries with weak state power but no revolution
    - Eastern European countries were certainly not the countries with the weakest state power then
- Both cannot precisely predict the occurence of revolution

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014
- Background: influenza (flu) tracking system in CDC

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014
- Background: influenza (flu) tracking system in CDC
  - Patients visit doctors -> doctors make diagnosis -> report to CDC

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014
- Background: influenza (flu) tracking system in CDC
  - Patients visit doctors -> doctors make diagnosis -> report to CDC
  - Accurate, but with a lag of weeks

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014
- Background: influenza (flu) tracking system in CDC
  - Patients visit doctors -> doctors make diagnosis -> report to CDC
  - Accurate, but with a lag of weeks
- Using Google Searches to track fluence in real time

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014
- Background: influenza (flu) tracking system in CDC
  - Patients visit doctors -> doctors make diagnosis -> report to CDC
  - Accurate, but with a lag of weeks
- Using Google Searches to track fluence in real time
  - Intuition: people will search flu-related words, such as “flu symptoms”

Example of Prediction: Google Search and Flu

- Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2009), no. 7232, 1012–1014
- Background: influenza (flu) tracking system in CDC
  - Patients visit doctors -> doctors make diagnosis -> report to CDC
  - Accurate, but with a lag of weeks
- Using Google Searches to track fluence in real time
  - Intuition: people will search flu-related words, such as "flu symptoms"
  - And the trends of these searches predict ups and downs of flu cases

Google Flu Trends: Details

- 45 search queries related to Influenza-like illness (ILI)

Google Flu Trends: Details

- 45 search queries related to Influenza-like illness (ILI)
- $Q(t)$: ILI-related query fraction at time t, out of all searches in a geographic region

Google Flu Trends: Details

- 45 search queries related to Influenza-like illness (ILI)
- $Q(t)$: ILI-related query fraction at time t, out of all searches in a geographic region
- $I(t)$: Number of ILI physician visit at time t

Google Flu Trends: Details

- 45 search queries related to Influenza-like illness (ILI)
- $Q(t)$: ILI-related query fraction at time t, out of all searches in a geographic region
- $I(t)$: Number of ILI physician visit at time t
- Model: simple linear regression

Google Flu Trends: Details

- 45 search queries related to Influenza-like illness (ILI)
- $Q(t)$: ILI-related query fraction at time t, out of all searches in a geographic region
- $I(t)$: Number of ILI physician visit at time t
- Model: simple linear regression
- $logit(I(t)) = \beta logit(Q(t)) + \epsilon$

Google Flu Trends: Details

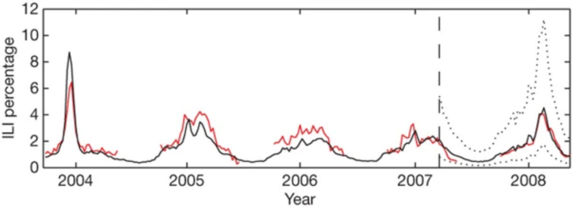
- 45 search queries related to Influenza-like illness (ILI)
- $Q(t)$: ILI-related query fraction at time t, out of all searches in a geographic region
- $I(t)$: Number of ILI physician visit at time t
- Model: simple linear regression
- $logit(I(t)) = \beta logit(Q(t)) + \epsilon$
- The model was fit using data from 2003 to 2007

Google Flu Trends: Details

- 45 search queries related to Influenza-like illness (ILI)
- $Q(t)$: ILI-related query fraction at time t, out of all searches in a geographic region
- $I(t)$: Number of ILI physician visit at time t
- Model: simple linear regression
- $logit(I(t)) = \beta logit(Q(t)) + \epsilon$
- The model was fit using data from 2003 to 2007
- And make predictions for 2008

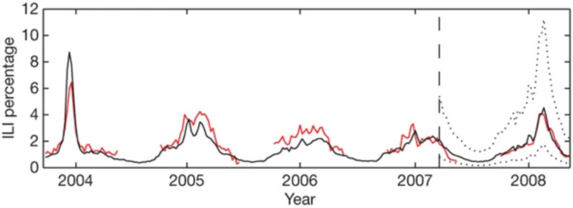
Google Flu Trends: Results

- Red is Google Search; black is CDC's count



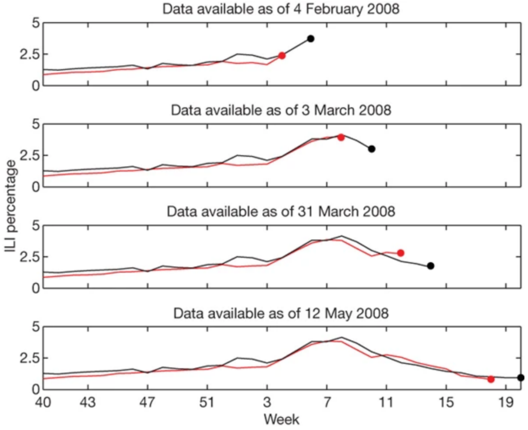
Google Flu Trends: Results

- Red is Google Search; black is CDC's count
- Correlation in 2008 is 0.95



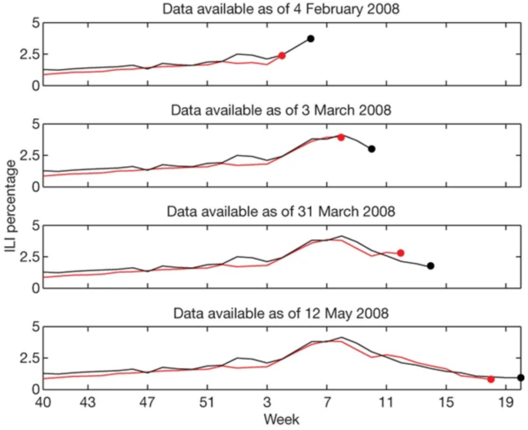
Google Flu Trends: nowcasting

- Nowcasting: predict what will happen in the near future/now



Google Flu Trends: nowcasting

- Nowcasting: predict what will happen in the near future/now
- A weaker and more realizable version of forecasting



Google Flu Trends: discussions

- https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_
- [In Class Activities]

Google Flu Trends: discussions

- https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_
- [In Class Activities]
  - What else you think Google's search trend can predict?

Google Flu Trends: discussions

- https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_
- [In Class Activities]
  - What else you think Google's search trend can predict?
    - https://trends.google.com/trends/explore?q=covid&geo=US

Google Flu Trends: discussions

- https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_
- [In Class Activities]
  - What else you think Google's search trend can predict?
    - https://trends.google.com/trends/explore?q=covid&geo=US
  - What do you think are the potential problems of using search queries to predict influenza counts?

Google Flu Trends: Critique 1

- Challenge 1: we can actually use old methods and old data to predict flu

Google Flu Trends: Critique 1

- Challenge 1: we can actually use old methods and old data to predict flu
  - Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, *Predicting consumer behavior with Web search*, Proceedings of the National Academy of Sciences **107** (2010), no. 41, 17486–17490

Google Flu Trends: Critique 1

- Challenge 1: we can actually use old methods and old data to predict flu
    - Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, *Predicting consumer behavior with Web search*, Proceedings of the National Academy of Sciences **107** (2010), no. 41, 17486–17490
    - $I(t) = \alpha + \beta_1 I(t-2) + \beta_1 I(t-3) + \epsilon$

Google Flu Trends: Critique 1

- Challenge 1: we can actually use old methods and old data to predict flu
  - Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, *Predicting consumer behavior with Web search*, Proceedings of the National Academy of Sciences **107** (2010), no. 41, 17486–17490
  - $I(t) = \alpha + \beta_1 I(t-2) + \beta_1 I(t-3) + \epsilon$
  - The above autoregressive model achieves similar performances

Google Flu Trends: Critique 1

- Challenge 1: we can actually use old methods and old data to predict flu
  - Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, *Predicting consumer behavior with Web search*, Proceedings of the National Academy of Sciences **107** (2010), no. 41, 17486–17490
  - $I(t) = \alpha + \beta_1 I(t-2) + \beta_1 I(t-3) + \epsilon$
  - The above autoregressive model achieves similar performances
    - But no need to collect big data! Existing statistics from the CDC is enough

Google Flu Trends: Critique 1

- Challenge 1: we can actually use old methods and old data to predict flu
    - Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts, *Predicting consumer behavior with Web search*, Proceedings of the National Academy of Sciences **107** (2010), no. 41, 17486–17490
    - $I(t) = \alpha + \beta_1 I(t-2) + \beta_1 I(t-3) + \epsilon$
    - The above autoregressive model achieves similar performances
        - But no need to collect big data! Existing statistics from the CDC is enough
    - "search data are comparable in utility to alternative information soruces, but not necessarily superior"

Google Flu Trends: Critique 2

- Drifting

Google Flu Trends: Critique 2

- Drifting
  - Users may change their search behaviors during pandamic period, leading to overrestimation

Google Flu Trends: Critique 2

- Drifting
  - Users may change their search behaviors during pandamic period, leading to overrestimation
  - Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi, *Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic*, PLOS ONE **6** (2011), no. 8, e23610

Google Flu Trends: Critique 2

- Drifting
  - Users may change their search behaviors during pandamic period, leading to overrestimation
  - Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi, *Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic*, PLOS ONE **6** (2011), no. 8, e23610
- Algorithm confounding!

Google Flu Trends: Critique 2

- Drifting
  - Users may change their search behaviors during pandamic period, leading to overrestimation
  - Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi, *Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic*, PLOS ONE **6** (2011), no. 8, e23610
- Algorithm confounding!
  - Google began to suggest related search words

Google Flu Trends: Critique 2

- Drifting
  - Users may change their search behaviors during pandamic period, leading to overrestimation
  - Samantha Cook, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi, *Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic*, PLOS ONE **6** (2011), no. 8, e23610
- Algorithm confounding!
  - Google began to suggest related search words
  - David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, *The Parable of Google Flu: Traps in Big Data Analysis*, Science **343** (2014), no. 6176, 1203–1205

Google Flu Trends: Aftermath

- There are tons of media report titled “Google’s Flu Project Shows the Failings of Big Data”

Google Flu Trends: Aftermath

- There are tons of media report titled "Google's Flu Project Shows the Failings of Big Data"
    - `https://time.com/23782/google-flu-trends-big-data-problems/`

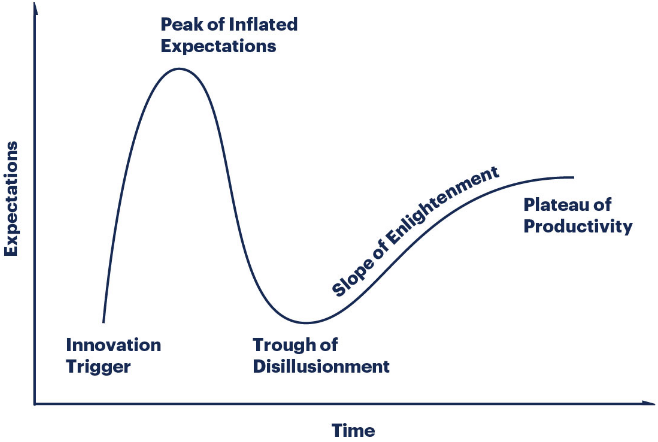
Google Flu Trends: Aftermath

- There are tons of media report titled "Google's Flu Project Shows the Failings of Big Data"
  - `https://time.com/23782/google-flu-trends-big-data-problems/`
- And Google stopped publishing estimate of ILI counts after 2015

Google Flu Trends: Aftermath

- There are tons of media report titled “Google's Flu Project Shows the Failings of Big Data”
  - https://time.com/23782/google-flu-trends-big-data-problems/
- And Google stopped publishing estimate of ILI counts after 2015
  - https://ai.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html

Hype Cycle of Using Big Data for Prediction



COVID-19

- Google began to release search queries related to COVID-19

COVID-19

- Google began to release search queries related to COVID-19
  - https://github.com/google-research/open-covid-19-data/tree/master/data/exports/search_trends_symptoms_dataset

COVID-19

- Google began to release search queries related to COVID-19
  - `https://github.com/google-research/open-covid-19-data/tree/master/data/exports/search_trends_symptoms_dataset`
- Many recent studies (just google "Google Search Predicts COVID")

COVID-19

- Google began to release search queries related to COVID-19
  - https://github.com/google-research/open-covid-19-data/tree/master/data/exports/search_trends_symptoms_dataset
- Many recent studies (just google "Google Search Predicts COVID")
  - https://www.cghjournal.org/article/S1542-3565(20)30922-8/fulltext

COVID-19

- Google began to release search queries related to COVID-19
  - https://github.com/google-research/open-covid-19-data/tree/master/data/exports/search_trends_symptoms_dataset
- Many recent studies (just google “Google Search Predicts COVID”)
  - https://www.cghjournal.org/article/S1542-3565(20)30922-8/fulltext
  - Loss of taste and loss of appetite correlated most strongly with the rise in COVID-19 (with a four-week lead)

Short summary

- Prediction is different from explanation

Short summary

- Prediction is different from explanation
- Current social sciences focus too much on explanation, but theory is often not good at predicction

Short summary

- Prediction is different from explanation
- Current social sciences focus too much on explanation, but theory is often not good at predicction
- Predction can be useful for real-world problems

Short summary

- Prediction is different from explanation
- Current social sciences focus too much on explanation, but theory is often not good at predicction
- Predction can be useful for real-world problems
- But bear a critical mind

Prediction and Explanation using statistics

- Leo Breiman, *Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)*, Statistical Science **16** (2001), no. 3, 199–231. MR1874152

Social scientists/statisticans	Computer scientists/data scientists
Data modeling	Algorithmic modeling
Traditional regression models	Machine learning
$Y = \color{red}{\beta} X$	$\color{red}{Y} = \beta X$
Explanation	Prediction

Traditional data modeling approach

- Breiman, 2001:

  Assume that the data are generated by the following model

Traditional data modeling approach

- Breiman, 2001:

  Assume that the data are generated by the following model

- But why? We know the assumptions are mostly wrong

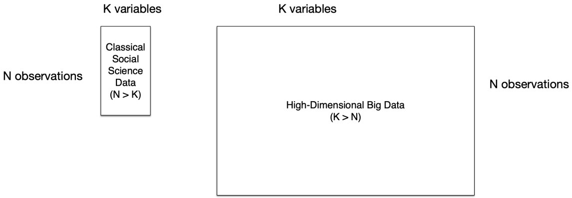
Traditional data modeling approach

- Breiman, 2001:

  Assume that the data are generated by the following model

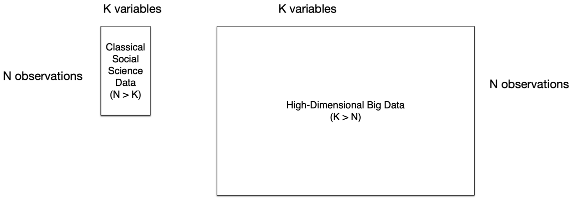
- But why? We know the assumptions are mostly wrong
- Why not use more powerful models (machine learning), if the goal is prediction?

Why Machine Learning? I



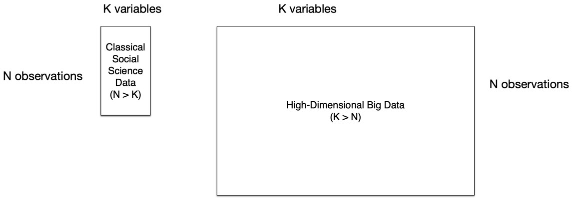
- Big data are not only big (large N)

Why Machine Learning? I



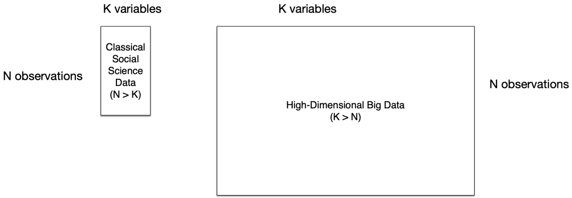
- Big data are not only big (large N)
- They are also high-dimensional ($K > N$)

Why Machine Learning? I



- Big data are not only big (large N)
- They are also high-dimensional ($K > N$)
- This is known as curse of dimensionality

Why Machine Learning? I



- Big data are not only big (large N)
- They are also high-dimensional ($K > N$)
- This is known as curse of dimensionality
- And traditional regression models familiar to social scientists do not work very well with high-dimensional data

Why Machine Learning? II

- When $K > N$, standard regression models do not have unique solutions

Picture 1

$$\begin{aligned} y = &\ 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} \\ &- 2.1x_{17} + 3.2x_{27}, \end{aligned}$$

Picture 2

$$\begin{aligned} y = &\ -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} \\ &+ 17.5x_{21} + 0.2x_{22}, \end{aligned}$$

Picture 3

$$\begin{aligned} y = &\ -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 \\ &+ 3.4x_{11} + 7.2x_{28}. \end{aligned}$$

Why Machine Learning? II

- When $K > N$, standard regression models do not have unique solutions
- Breiman's example: Rashomon and the multiplicity of good models

Picture 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},$$

Picture 2

$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},$$

Picture 3

$$y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.$$

Why Machine Learning? II

- When $K > N$, standard regression models do not have unique solutions
- Breiman's example: Rashomon and the multiplicity of good models
- The below three models may all good solutions

Picture 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},$$

Picture 2

$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},$$

Picture 3

$$y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.$$

Goal of machine learning

- Observed: a set of input X and output Y in training data

Goal of machine learning

- Observed: a set of input X and output Y in training data
- Goal: find an algorithm $f(\cdot)$, "such that for future X in a test set, $f(X)$ will be a good predictior for Y" (Breiman, 2001)

Goal of machine learning

- Observed: a set of input X and output Y in training data
- Goal: find an algorithm $f(\cdot)$, "such that for future X in a test set, $f(X)$ will be a good predictior for Y" (Breiman, 2001)
    - We often do not care what f exactly looks like, as long as it can improve predictive performance

Goal of machine learning

- Observed: a set of input X and output Y in training data
- Goal: find an algorithm $f(\cdot)$, "such that for future X in a test set, $f(X)$ will be a good predictior for Y" (Breiman, 2001)
  - We often do not care what f exactly looks like, as long as it can improve predictive performance
  - To say that one algorithm f is better than another algorithm g, we need to evaluate their predictive performances

Two types of machine learning

- Predicting continuous outcomes is often called regression tasks

Two types of machine learning

- Predicting continuous outcomes is often called regression tasks
  - Yes linear regressions are a type of machine learning, the simpliest one

Two types of machine learning

- Predicting continuous outcomes is often called regression tasks
  - Yes linear regressions are a type of machine learning, the simplest one
- Predicting categorical outcomes is called classification tasks

Two types of machine learning

- Predicting continuous outcomes is often called regression tasks
  - Yes linear regressions are a type of machine learning, the simpliest one
- Predicting categorical outcomes is called classification tasks
  - Slightly differnet notation: logistic regression is treated as a classification task in machine learning community

Next steps

- From week 4, we will see how machine learning techniques are used in text data

Baseline of prediction

- Recall the Google Flu Trends example

Baseline of prediction

- Recall the Google Flu Trends example
- Using search queries for prediction is just as good as a simplest autoregressive model

Baseline of prediction

- Recall the Google Flu Trends example
- Using search queries for prediction is just as good as a simplest autoregressive model
- Takeaway: to say a predictive model f is good, we need to quantitatively measure it's performances, against some baseline prediction

Baseline of prediction

- Recall the Google Flu Trends example
- Using search queries for prediction is just as good as a simplest autoregressive model
- Takeaway: to say a predictive model f is good, we need to quantitatively measure it's performances, against some baseline prediction
  - Baseline prediction is something you can achieve very easily with existing data

Baseline of prediction

- Recall the Google Flu Trends example
- Using search queries for prediction is just as good as a simplest autoregressive model
- Takeaway: to say a predictive model f is good, we need to quantitatively measure it's performances, against some baseline prediction
  - Baseline prediction is something you can achieve very easily with existing data
  - E.g., random guesses

Baseline of prediction

- Recall the Google Flu Trends example
- Using search queries for prediction is just as good as a simplest autoregressive model
- Takeaway: to say a predictive model f is good, we need to quantitatively measure it's performances, against some baseline prediction
  - Baseline prediction is something you can achieve very easily with existing data
  - E.g., random guesses
  - So that you are not spending time/resource and find that you are only slightly better than a very simple method

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:
- R^2: what you get from regression models; focusing on variances predictable from your X

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:
- R^2: what you get from regression models; focusing on variances predictable from your X
  - The larger the R^2, the better the model

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:
- R^2: what you get from regression models; focusing on variances predictable from your X
  - The larger the R^2, the better the model
- *MSE* (mean squared error): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:
- R^2: what you get from regression models; focusing on variances predictable from your X
  - The larger the R^2, the better the model
- *MSE* (mean squared error): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - The smaller the MSE, the better the model

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:
- R^2: what you get from regression models; focusing on variances predictable from your X
  - The larger the R^2, the better the model
- MSE (mean squared error): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - The smaller the MSE, the better the model
  - Sometimes we also use RMSE ($\sqrt{MSE}$)

Prediction evaluations for continuous outcomes

- From now on, I use $\hat{Y} = f(X)$ as the predicted value of Y
- For continuous outcomes:
- R^2: what you get from regression models; focusing on variances predictable from your X
    - The larger the R^2, the better the model
- MSE (mean squared error): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
    - The smaller the MSE, the better the model
    - Sometimes we also use RMSE ($\sqrt{MSE}$)
- MAE (mean absolute error): $\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$

Prediction evaluations for categorical outcomes

- For categorical outcomes, evaluation is more complex

Prediction evaluations for categorical outcomes

- For categorical outcomes, evaluation is more complex
- Let us work with the simplest example of binary outcomes

Prediction evaluations for categorical outcomes

- For categorical outcomes, evaluation is more complex
- Let us work with the simplest example of binary outcomes
- Say we has an algorithm predicting COVID infection (positive $= 1$ vs. negative $= 0$)

Prediction evaluations for categorical outcomes

- For categorical outcomes, evaluation is more complex
- Let us work with the simpliest example of binary outcomes
- Say we has an algorithm predicting COVID infection (positive $= 1$ vs. negative $= 0$)
- We found that 99% of our predictions are correct. Yeah!

Prediction evaluations for categorical outcomes

- For categorical outcomes, evaluation is more complex
- Let us work with the simplest example of binary outcomes
- Say we has an algorithm predicting COVID infection (positive $= 1$ vs. negative $= 0$)
- We found that 99% of our predictions are correct. Yeah!
- But wait, is that good enough?

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)
- What's the accuracy for this trivial prediction?

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)
- What's the accuracy for this trivial prediction?
- There are 10,000 students/employees at HKUST

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)
- What's the accuracy for this trivial prediction?
- There are 10,000 students/employees at HKUST
- 10 infected cases

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)
- What's the accuracy for this trivial prediction?
- There are 10,000 students/employees at HKUST
- 10 infected cases
- So the error rate is 10/10000, and accuracy = 1 - 10/10000 = 99.9%

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)
- What's the accuracy for this trivial prediction?
- There are 10,000 students/employees at HKUST
- 10 infected cases
- So the error rate is 10/10000, and accuracy = 1 - 10/10000 = 99.9%
- If class is imbalanced, it is very easy to achieve a high accuracy by predicting the majority class all the time

Prediction evaluations for categorical outcomes

- In fact, for any classification task, one of the simpliest baseline is to predict every data point as belonging to the majority class
- Here, we know most people are not affected
- So the trivial prediction here just predict that every one is negative (0)
- What's the accuracy for this trivial prediction?
- There are 10,000 students/employees at HKUST
- 10 infected cases
- So the error rate is 10/10000, and accuracy = 1 - 10/10000 = 99.9%
- If class is imbalanced, it is very easy to achieve a high accuracy by predicting the majority class all the time
  - But it's not useful at all!

Prediction evaluations for categorical outcomes

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

- It's better to use confusion matrix

Prediction evaluations for categorical outcomes

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

- It's better to use confusion matrix
- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

Prediction evaluations for categorical outcomes

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

- It's better to use confusion matrix
- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- precision $= \frac{TP}{TP+FP}$

Prediction evaluations for categorical outcomes

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

- It's better to use confusion matrix
- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- precision $= \frac{TP}{TP+FP}$
  - Interpretation: what proportion of predicted positives are actual positive?

Prediction evaluations for categorical outcomes

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

- It's better to use confusion matrix
- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- precision $= \frac{TP}{TP+FP}$
  - Interpretation: what proportion of predicted positives are actual positive?
- recall $= \frac{TP}{TP+FN}$

Prediction evaluations for categorical outcomes

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/negative	False Negative (FN)	True Negative (TN)

- It's better to use confusion matrix
- accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- precision $= \frac{TP}{TP+FP}$
  - Interpretation: what proportion of predicted positives are actual positive?
- recall $= \frac{TP}{TP+FN}$
  - interpretation: what proportion true positives are identified by predictions?

Case 1: high precision/ low recall

Actual

Prediction		1/positive	0/negative
	1/positive	True Positive (n = 5)	False Positive (n =0)
	0/negative	False Negative (n = 5)	True Negative (n = 9990)

- Accuracy: $\frac{9+9986}{10000} = 99.95\%$

Case 1: high precision/ low recall

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 5)	False Positive (n =0)
	0/negative	False Negative (n = 5)	True Negative (n = 9990)

- Accuracy: $\frac{9+9986}{10000} = 99.95\%$
- Precision: $\frac{5}{5+0} = 100\%$

Case 1: high precision/ low recall

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 5)	False Positive (n =0)
	0/negative	False Negative (n = 5)	True Negative (n = 9990)

- Accuracy: $\frac{9+9986}{10000} = 99.95\%$
- Precision: $\frac{5}{5+0} = 100\%$
- Recall: $\frac{5}{5+5} = 50\%$

Case 1: high precision/ low recall

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 5)	False Positive (n =0)
	0/negative	False Negative (n = 5)	True Negative (n = 9990)

- Accuracy: $\frac{9+9986}{10000} = 99.95\%$
- Precision: $\frac{5}{5+0} = 100\%$
- Recall: $\frac{5}{5+5} = 50\%$
- So every predicted infected case is indeed infected

Case 1: high precision/ low recall

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 5)	False Positive (n =0)
	0/negative	False Negative (n = 5)	True Negative (n = 9990)

- Accuracy: $\frac{9+9986}{10000} = 99.95\%$
- Precision: $\frac{5}{5+0} = 100\%$
- Recall: $\frac{5}{5+5} = 50\%$
- So every predicted infected case is indeed infected
- But we missed 50% of actual infected cases

Case 2: high recall/low precision

		Actual: 1/positive	Actual: 0/negative
Prediction	1/positive	True Positive (n = 9)	False Positive (n = 4)
Prediction	0/negative	False Negative (n = 1)	True Negative (n = 9986)

- We lower the threshold to be considered as infection case

Case 2: high recall/low precision

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 9)	False Positive (n = 4)
	0/negative	False Negative (n = 1)	True Negative (n = 9986)

- We lower the threshold to be considered as infection case
- Accuracy: $\frac{9+9986}{10000} = 99.95\%$; the same

Case 2: high recall/low precision

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 9)	False Positive (n = 4)
	0/negative	False Negative (n = 1)	True Negative (n = 9986)

- We lower the threshold to be considered as infection case
- Accuracy: $\frac{9+9986}{10000} = 99.95\%$; the same
- Precision: $\frac{9}{9+4} = 69.23\%$

Case 2: high recall/low precision

		Actual: 1/positive	Actual: 0/negative
Prediction	1/positive	True Positive (n = 9)	False Positive (n = 4)
Prediction	0/negative	False Negative (n = 1)	True Negative (n = 9986)

- We lower the threshold to be considered as infection case
- Accuracy: $\frac{9+9986}{10000} = 99.95\%$; the same
- Precision: $\frac{9}{9+4} = 69.23\%$
- Recall: $\frac{9}{9+1} = 90\%$

Case 2: high recall/low precision

		Actual: 1/positive	Actual: 0/negative
Prediction	1/positive	True Positive (n = 9)	False Positive (n = 4)
Prediction	0/negative	False Negative (n = 1)	True Negative (n = 9986)

- We lower the threshold to be considered as infection case
- Accuracy: $\frac{9+9986}{10000} = 99.95\%$; the same
- Precision: $\frac{9}{9+4} = 69.23\%$
- Recall: $\frac{9}{9+1} = 90\%$
- Our prediction captures 90% of actual infected cases

Case 2: high recall/low precision

Actual

		1/positive	0/negative
Prediction	1/positive	True Positive (n = 9)	False Positive (n = 4)
	0/negative	False Negative (n = 1)	True Negative (n = 9986)

- We lower the threshold to be considered as infection case
- Accuracy: $\frac{9+9986}{10000} = 99.95\%$; the same
- Precision: $\frac{9}{9+4} = 69.23\%$
- Recall: $\frac{9}{9+1} = 90\%$
- Our prediction captures 90% of actual infected cases
- But less than 70% predicted cases are actually infected; false alarm

Trivial prediction: Majority Class

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 0)	False Positive (n = 0)
	0/negative	False Negative (n = 10)	True Negative (n = 9900)

- Predict the majority class (no one is affected)

Trivial prediction: Majority Class

		Actual: 1/positive	Actual: 0/negative
Prediction	1/positive	True Positive (n = 0)	False Positive (n = 0)
	0/negative	False Negative (n = 10)	True Negative (n = 9900)

- Predict the majority class (no one is affected)
- Accuracy: $\frac{9+9986}{10000}$ =99.9%; slightly worse

Trivial prediction: Majority Class

		Actual: 1/positive	Actual: 0/negative
Prediction	1/positive	True Positive (n = 0)	False Positive (n = 0)
	0/negative	False Negative (n = 10)	True Negative (n = 9900)

- Predict the majority class (no one is affected)
- Accuracy: $\frac{9+9986}{10000}$ =99.9%; slightly worse
- Better measures should tell us that this is a trivial prediction

Trivial prediction: Majority Class

		Actual: 1/positive	Actual: 0/negative
Prediction	1/positive	True Positive (n = 0)	False Positive (n = 0)
	0/negative	False Negative (n = 10)	True Negative (n = 9900)

- Predict the majority class (no one is affected)
- Accuracy: $\frac{9+9986}{10000}$ =99.9%; slightly worse
- Better measures should tell us that this is a trivial prediction
- Precision: not defined

Trivial prediction: Majority Class

Actual

		1/positive	0/negative
Prediction	1/positive	True Positive (n = 0)	False Positive (n = 0)
	0/negative	False Negative (n = 10)	True Negative (n = 9900)

- Predict the majority class (no one is affected)
- Accuracy: $\frac{9+9986}{10000}$ =99.9%; slightly worse
- Better measures should tell us that this is a trivial prediction
- Precision: not defined
- Recall: $\frac{9}{9+1} = 0\%$

Trivial prediction: Majority Class

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (n = 0)	False Positive (n = 0)
	0/negative	False Negative (n = 10)	True Negative (n = 9900)

- Predict the majority class (no one is affected)
- Accuracy: $\frac{9+9986}{10000}$ =99.9%; slightly worse
- Better measures should tell us that this is a trivial prediction
- Precision: not defined
- Recall: $\frac{9}{9+1} = 0\%$
- Even though accuracy is high, precision/recall is not satisfactory

Precision-recall trade-off

- In evaluting perdiction performances for categorical outcome, do not use accuracy

Precision-recall trade-off

- In evaluting perdiction performances for categorical outcome, do not use accuracy
- Instead, use precision and recall

Precision-recall trade-off

- In evaluting perdiction performances for categorical outcome, do not use accuracy
- Instead, use precision and recall
- Depending on tasks, we may emphasize one or the other
- [in class activities]: can you think of examples we should focus one or the other?

Precision-recall trade-off

- In evaluting perdiction performances for categorical outcome, do not use accuracy
- Instead, use precision and recall
- Depending on tasks, we may emphasize one or the other
- [in class activities]: can you think of examples we should focus one or the other?
- Ideally, we want both precision and recall to be high

Precision-recall trade-off

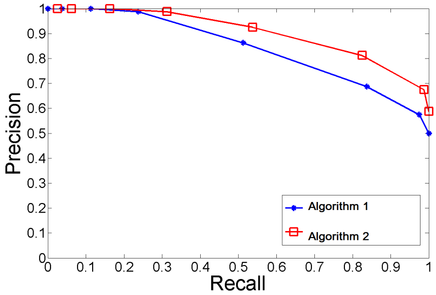
- In evaluting perdiction performances for categorical outcome, do not use accuracy
- Instead, use precision and recall
- Depending on tasks, we may emphasize one or the other
- [in class activities]: can you think of examples we should focus one or the other?
- Ideally, we want both precision and recall to be high
- In reality, it's often that one comes at the cost of another

Precision-recall trade-off

- In evaluting perdiction performances for categorical outcome, do not use accuracy
- Instead, use precision and recall
- Depending on tasks, we may emphasize one or the other
- [in class activities]: can you think of examples we should focus one or the other?
- Ideally, we want both precision and recall to be high
- In reality, it's often that one comes at the cost of another
- F-1 score: balancing the two $2 * \frac{precision * recall}{precision + recall}$

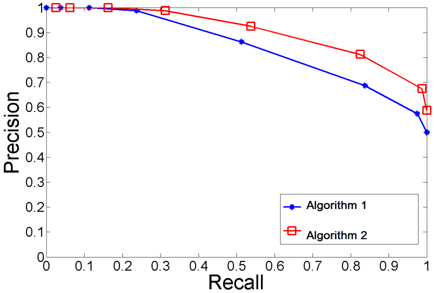
Precision-recall curve

- We can find a trade-off by using precision-recall curve



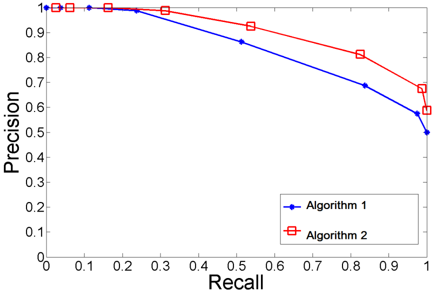
Precision-recall curve

- We can find a trade-off by using precision-recall curve
- $f(X)$ generate predicted probability of $Y = 1$



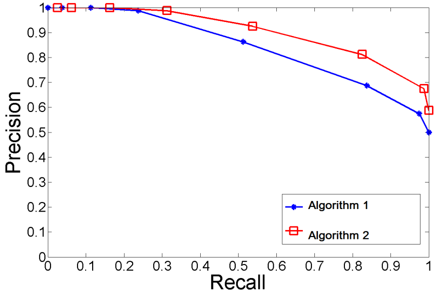
Precision-recall curve

- We can find a trade-off by using precision-recall curve
- $f(X)$ generate predicted probability of $Y = 1$
- If predicted probability is larger than a threshold ϕ, $\hat{Y} = 1$; otherwise $\hat{Y} = 0$



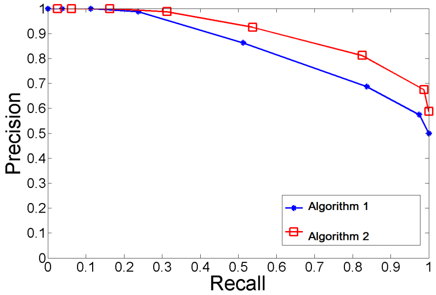
Precision-recall curve

- We can find a trade-off by using precision-recall curve
- $f(X)$ generate predicted probability of $Y = 1$
- If predicted probability is larger than a threshold ϕ, $\hat{Y} = 1$; otherwise $\hat{Y} = 0$
- Large threshold ϕ -> high precision



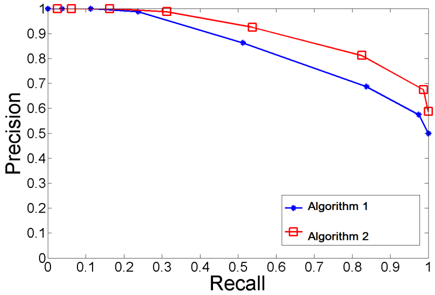
Precision-recall curve

- We can find a trade-off by using precision-recall curve
- $f(X)$ generate predicted probability of $Y = 1$
- If predicted probability is larger than a threshold ϕ, $\hat{Y} = 1$; otherwise $\hat{Y} = 0$
- Large threshold ϕ -> high precision
- Small threshold ϕ -> high recall



Precision-recall curve

- We can find a trade-off by using precision-recall curve
- $f(X)$ generate predicted probability of $Y = 1$
- If predicted probability is larger than a threshold ϕ, $\hat{Y} = 1$; otherwise $\hat{Y} = 0$
- Large threshold ϕ -> high precision
- Small threshold ϕ -> high recall
- Algorithm 2 is better than 1



ROC curve

- Another common measure is called ROC curve

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

ROC curve

- Another common measure is called ROC curve

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

- True positive rate (i.e., recall): $\frac{TP}{TP+FN}$

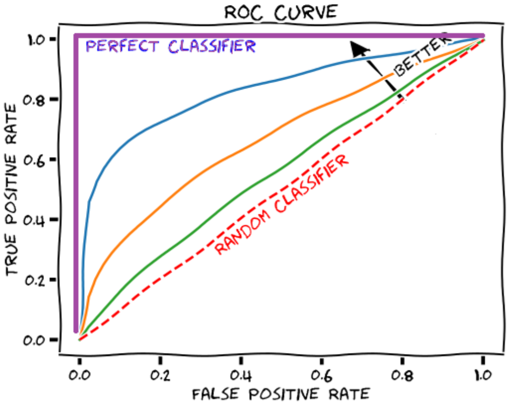
ROC curve

- Another common measure is called ROC curve

		Actual	
		1/positive	0/negative
Prediction	1/positive	True Positive (TP)	False Positive (FP)
	0/ negative	False Negative (FN)	True Negative (TN)

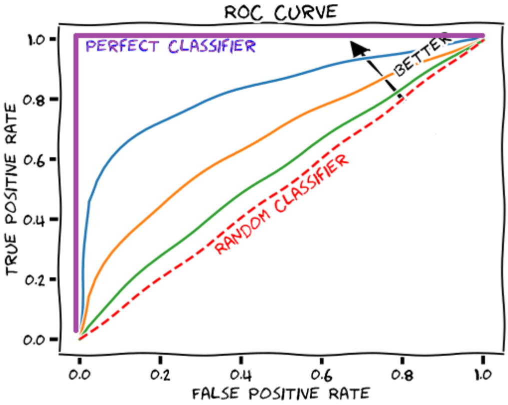
- True positive rate (i.e., recall): $\frac{TP}{TP+FN}$
- False positive rate: $\frac{FP}{FP+TN}$

ROC curve



- AUC: area under the curve

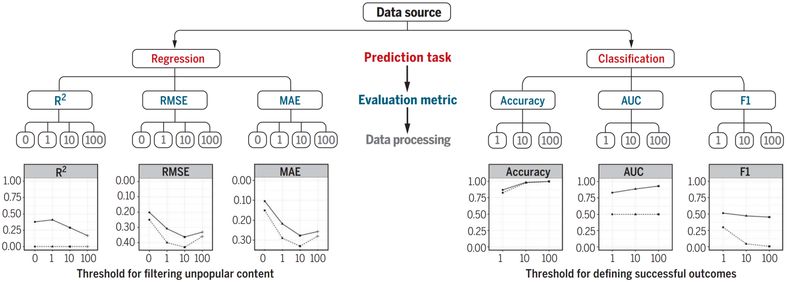
ROC curve



- AUC: area under the curve
- Bigger AUC -> better prediction performance

Summary of evaluation characteristics

- Jake M. Hofman, Amit Sharma, and Duncan J. Watts, *Prediction and explanation in social systems*, Science **355** (2017), no. 6324, 486–488



Limits of prediction

- Jake M. Hofman, Amit Sharma, and Duncan J. Watts, *Prediction and explanation in social systems*, Science **355** (2017), no. 6324, 486–488

Limits of prediction

- Jake M. Hofman, Amit Sharma, and Duncan J. Watts, *Prediction and explanation in social systems*, Science **355** (2017), no. 6324, 486–488
- Some problems are hard to predict; others are easy

Limits of prediction

- Jake M. Hofman, Amit Sharma, and Duncan J. Watts, *Prediction and explanation in social systems*, Science **355** (2017), no. 6324, 486–488
- Some problems are hard to predict; others are easy
- Revolution may be very hard to predict

Limits of prediction

- Jake M. Hofman, Amit Sharma, and Duncan J. Watts, *Prediction and explanation in social systems*, Science **355** (2017), no. 6324, 486–488
- Some problems are hard to predict; others are easy
- Revolution may be very hard to predict
- Trends of influenza counts may be easier

Summary

- Critical evaluation of prediction problem vs. explanation problem

Summary

- Critical evaluation of prediction problem vs. explanation problem
- Knowledge:

Summary

- Critical evaluation of prediction problem vs. explanation problem
- Knowledge:
    - What is high-dimensional data?

Summary

- Critical evaluation of prediction problem vs. explanation problem
- Knowledge:
  - What is high-dimensional data?
  - What is train-split

Summary

- Critical evaluation of prediction problem vs. explanation problem
- Knowledge:
  - What is high-dimensional data?
  - What is train-split
  - Evaluations

Next week

- Survey and Big Data