SPIA 586a
Machine Learning for Policy Analysis

Brandon Stewart

Princeton

January 27, 2025

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- This Week
  - course structure
  - core ideas of the course
  - preview of seven tools
  - review of regression and basis functions

Where We've Been and Where We're Going...

- This Week
  - course structure
  - core ideas of the course
  - preview of seven tools
  - review of regression and basis functions
- Next Week
  - discovery
  - measurement

Where We've Been and Where We're Going...

- This Week
  - course structure
  - core ideas of the course
  - preview of seven tools
  - review of regression and basis functions
- Next Week
  - discovery
  - measurement
- Long Run
  - target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

Welcome and Introductions

Welcome and Introductions

Welcome and Introductions

I

- . . . am an Associate Professor in Sociology.

Welcome and Introductions

I

- . . . am an Associate Professor in Sociology.
- . . . am trained in political science and statistics

Welcome and Introductions

I

- . . . am an Associate Professor in Sociology.
- . . . am trained in political science and statistics
- . . . do research in methods and statistical text analysis

Overview

Overview

- Goal: bring you up to speed on machine learning language and ground the use of machine learning in good principles of research design.
    - NB: this is more machine learning to do policy analysis than the policy of artificial intelligence

Overview

- Goal: bring you up to speed on machine learning language and ground the use of machine learning in good principles of research design.
  - NB: this is more machine learning to do policy analysis than the policy of artificial intelligence
- This is my first policy class, so content depends a lot on your feedback.

Overview

- Goal: bring you up to speed on machine learning language and ground the use of machine learning in good principles of research design.
  - NB: this is more machine learning to do policy analysis than the policy of artificial intelligence
- This is my first policy class, so content depends a lot on your feedback.
- Syllabus is a useful resource including philosophy of the class.

Prerequisites and Varied Backgrounds

Prerequisites and Varied Backgrounds

- I assume you:
  - can do basic programming in `R`
  - have familiarity with linear regression
  - an interest in learning about machine learning

Prerequisites and Varied Backgrounds

- I assume you:
  - can do basic programming in `R`
  - have familiarity with linear regression
  - an interest in learning about machine learning
- Evaluation of coding will be relatively light so you can scale it up as much as you want.

Outline of Topics

Outline in reverse order:

Outline of Topics

Outline in reverse order:

- Guiding Values:
  why values are essential for responsible use of ML including privacy, fairness and interpretability.

Outline of Topics

Outline in reverse order:

- Guiding Values:
  why values are essential for responsible use of ML including privacy, fairness and interpretability.
- Data Sources:
  why the origins of our data matter and how to use experimental, design and found data.

Outline of Topics

Outline in reverse order:

- Guiding Values:
  why values are essential for responsible use of ML including privacy, fairness and interpretability.
- Data Sources:
  why the origins of our data matter and how to use experimental, design and found data.
- Target Tasks:
  why the task should drive the choice of methods and the distinction between discovery, measurement and three types of prediction.

Outline of Topics

Outline in reverse order:

- Guiding Values:
  why values are essential for responsible use of ML including privacy, fairness and interpretability.
- Data Sources:
  why the origins of our data matter and how to use experimental, design and found data.
- Target Tasks:
  why the task should drive the choice of methods and the distinction between discovery, measurement and three types of prediction.

target tasks $\rightarrow$ data sources $\rightarrow$ guiding values

Ways to Learn

Ways to Learn

- Class

  a mix of lecture and discussion, covering a concept and technique each week.

Ways to Learn

- Class
  a mix of lecture and discussion, covering a concept and technique each week.
- Reading
  complement to class material, will form the basis of discussion

Ways to Learn

- Class
  a mix of lecture and discussion, covering a concept and technique each week.
- Reading
  complement to class material, will form the basis of discussion
- Minor Assignments
  very short activities due the day before class to get your feet wet with coding. Will bridge between material.

Ways to Learn

- Class
  a mix of lecture and discussion, covering a concept and technique each week.
- Reading
  complement to class material, will form the basis of discussion
- Minor Assignments
  very short activities due the day before class to get your feet wet with coding. Will bridge between material.
- Major Assignments
  opportunities to reinforce understanding of material

Ways to Learn

- Class
  a mix of lecture and discussion, covering a concept and technique each week.
- Reading
  complement to class material, will form the basis of discussion
- Minor Assignments
  very short activities due the day before class to get your feet wet with coding. Will bridge between material.
- Major Assignments
  opportunities to reinforce understanding of material
- Proposals
  get and give feedback on possible applications of ML.

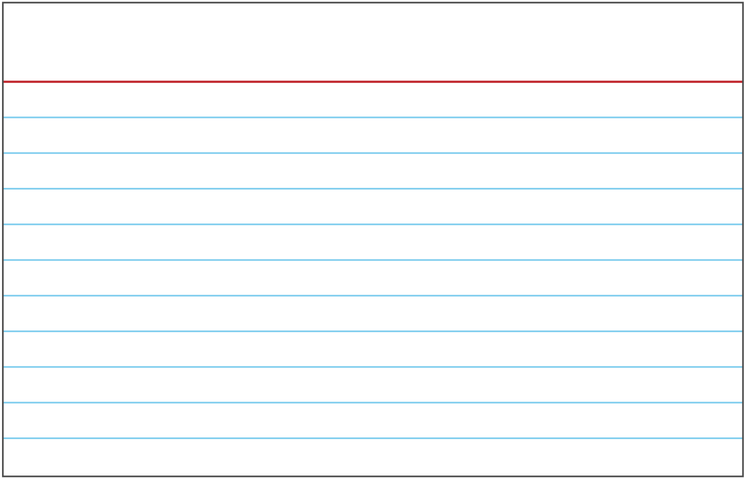
Ways to Learn

- Class
  a mix of lecture and discussion, covering a concept and technique each week.
- Reading
  complement to class material, will form the basis of discussion
- Minor Assignments
  very short activities due the day before class to get your feet wet with coding. Will bridge between material.
- Major Assignments
  opportunities to reinforce understanding of material
- Proposals
  get and give feedback on possible applications of ML.
- Ed
  ask questions of me and your classmates

Ways to Learn

- Class
  a mix of lecture and discussion, covering a concept and technique each week.
- Reading
  complement to class material, will form the basis of discussion
- Minor Assignments
  very short activities due the day before class to get your feet wet with coding. Will bridge between material.
- Major Assignments
  opportunities to reinforce understanding of material
- Proposals
  get and give feedback on possible applications of ML.
- Ed
  ask questions of me and your classmates
- Office Hours
  ask even more questions, but in-person

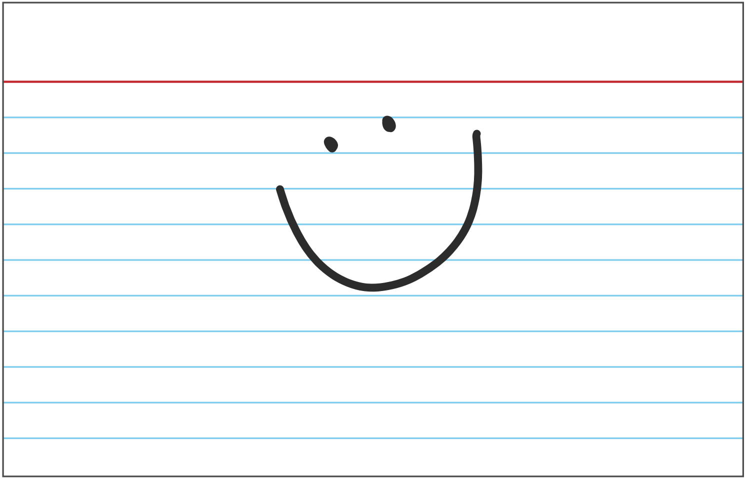
Notecards



Notecards

Can we go over Bayes' Rule again?

Notecards



Overview of Methods Coverage

Overview of Methods Coverage

- Goal: provide you exposure to the intuition and implementation of major machine learning techniques.

Overview of Methods Coverage

- Goal: provide you exposure to the intuition and implementation of major machine learning techniques.
- Stylistically, light math but options to go deeper.

Overview of Methods Coverage

- Goal: provide you exposure to the intuition and implementation of major machine learning techniques.
- Stylistically, light math but options to go deeper.

Computing usually is easy. Figuring out what to compute and what the results mean can be very hard.

-Berk (2020)

What Won't We Cover?

What Won't We Cover?

- anything heavy on math—we are going for breadth of approaches over depth

What Won't We Cover?

- anything heavy on math—we are going for breadth of approaches over depth
- approaches that have less obvious social science uses (e.g. teaching an AI to play Go)

What Won't We Cover?

- anything heavy on math—we are going for breadth of approaches over depth
- approaches that have less obvious social science uses (e.g. teaching an AI to play Go)
- computing uncertainty and confidence intervals—these are much harder for machine learning and getting around these problems takes more math that we wanted to require as prereqs

What Won't We Cover?

- anything heavy on math—we are going for breadth of approaches over depth
- approaches that have less obvious social science uses (e.g. teaching an AI to play Go)
- computing uncertainty and confidence intervals—these are much harder for machine learning and getting around these problems takes more math that we wanted to require as prereqs
- little on interpretation of individual parameters for reasons that will be clearer as the course goes on.

Structure of the Week

Structure of the Week

- Monday Class 9AM-10:30AM: Concepts
- Tuesday 5PM: Minor Assignments Due
- Wednesday Class 9AM-10:30AM: ML and Coding
- Thursday 5PM: Proposal Due (if doing that week)
- Friday 10AM-11AM: office hours
- Friday 5PM: Feedback on Proposals + Major Assignment Deadlines

Readings

Readings

Smattering of articles, but mainly four books:

Readings

Smattering of articles, but mainly four books:

- Salganik, Matthew J. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019. [Salganik] `https://www.bitbybitbook.com/`

Readings

Smattering of articles, but mainly four books:

- Salganik, Matthew J. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019. [Salganik] `https://www.bitbybitbook.com/`
- Grimmer, Justin, Margaret Roberts and Brandon M. Stewart, 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. [GRS] Will provide excerpts—no need to purchase.

Readings

Smattering of articles, but mainly four books:

- Salganik, Matthew J. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019. [Salganik] `https://www.bitbybitbook.com/`
- Grimmer, Justin, Margaret Roberts and Brandon M. Stewart, 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. [GRS] Will provide excerpts—no need to purchase.
- Barocas, Hardt and Narayanan *Fairness and machine learning: Limitations and Opportunities*. Book Manuscript. [BHN] `https://fairmlbook.org/`

Readings

Smattering of articles, but mainly four books:

- Salganik, Matthew J. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019. [Salganik] `https://www.bitbybitbook.com/`
- Grimmer, Justin, Margaret Roberts and Brandon M. Stewart, 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. [GRS] Will provide excerpts—no need to purchase.
- Barocas, Hardt and Narayanan *Fairness and machine learning: Limitations and Opportunities*. Book Manuscript. [BHN] `https://fairmlbook.org/`
- James, Witten, Hastie and Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2021. [JWHT] `https://www.statlearning.com/`

Everything will be provided digitally.

If you need more help. . .

If you need more help...

- post on Ed

If you need more help. . .

- post on Ed
- ask questions during class

If you need more help. . .

- post on Ed
- ask questions during class
- dig into the reading

If you need more help. . .

- post on Ed
- ask questions during class
- dig into the reading
- write a question on your notecard

If you need more help...

- post on Ed
- ask questions during class
- dig into the reading
- write a question on your notecard
- help each other! Everyone has different backgrounds and many folks will have seen these parts before!

If you need more help...

- post on Ed
- ask questions during class
- dig into the reading
- write a question on your notecard
- help each other! Everyone has different backgrounds and many folks will have seen these parts before!

Key Message: if you don't understand something, that's okay. Just ask!

Scarcity and Social Science

Scarcity and Social Science

Scarcity and Social Science

Scarcity and Social Science

- Data collection was very expensive

Scarcity and Social Science

- Data collection was very expensive
- Computation was very limited

Scarcity and Social Science

- Data collection was very expensive
- Computation was very limited
- Thought was comparatively cheaper.

Scarcity and Social Science

- Data collection was very expensive
- Computation was very limited
- Thought was comparatively cheaper.

$\rightsquigarrow$ strong assumptions and high *a priori* theorizing.

Abundance and Social Science

Abundance and Social Science

Abundance and Social Science

LETTER

doi:10.1038/nature11421

A 61-million-person experiment in social influence and political mobilization

Robert M. Bond[1], Christopher J. Fariss[1], Jason J. Jones[2], Adam D. I. Kramer[3], Cameron Marlow[3], Jaime E. Settle[1] & James H. Fowler[1,4]

Human behaviour is thought to spread through face-to-face social networks, but it is difficult to identify social influence effects in observational studies[9–13], and it is unknown whether online social networks operate in the same way[14–19]. Here we report results from a randomized controlled trial of political mobilization messages delivered to 61 million Facebook users during the 2010 US congressional elections. The results show that the messages directly influenced political self-expression, information seeking and real-world voting behaviour of millions of people. Furthermore, the messages not only influenced the users who received them but also the users' friends, and friends of friends. The effect of social transmission on real-world voting was greater than the direct effect of the messages themselves, and nearly all the transmission occurred between 'close friends' who were more likely to have a face-to-face relationship. These results suggest that strong ties are instrumental for spreading both online and real-world behaviour in human social networks.

with all users of at least 18 years of age in the United States who accessed the Facebook website on 2 November 2010, the day of the US congressional elections. Users were randomly assigned to a 'social message' group, an 'informational message' group or a control group. The social message group ($n = 60,055,176$) was shown a statement at the top of their 'News Feed'. This message encouraged the user to vote, provided a link to find local polling places, showed a clickable button reading 'I Voted', showed a counter indicating how many other Facebook users had previously reported voting, and displayed up to six small randomly selected 'profile pictures' of the user's Facebook friends who had already clicked the I Voted button (Fig. 1). The informational message group ($n = 611,044$) was shown the message, poll information, counter and button, but they were not shown any faces of friends. The control group ($n = 613,096$) did not receive any message at the top of their News Feed.

The design of the experiment allowed us to assess the impact that the treatments had on three user actions; clicking the I Voted button,

Abundance and Social Science

Abundance and Social Science

- Data is cheap! Communities passively lead digitized footprints.

Abundance and Social Science

- Data is cheap! Communities passively lead digitized footprints.
- Computational power has exploded (as has the willingness to pay for it).

Abundance and Social Science

- Data is cheap! Communities passively lead digitized footprints.
- Computational power has exploded (as has the willingness to pay for it).
- The cost of thought is constant.

Abundance and Social Science

- Data is cheap! Communities passively lead digitized footprints.
- Computational power has exploded (as has the willingness to pay for it).
- The cost of thought is constant.

$\rightsquigarrow$ data-driven discovery.

Abundance and Social Science

- Data is cheap! Communities passively lead digitized footprints.
- Computational power has exploded (as has the willingness to pay for it).
- The cost of thought is constant.

$\rightsquigarrow$ data-driven discovery.

However! Must be done carefully!

The Perils of Badly Done Data-Driven Discovery

The Perils of Badly Done Data-Driven Discovery

Bennett, Baird, Miller and Wolford. (2009). "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction."

The Perils of Badly Done Data-Driven Discovery

Bennett, Baird, Miller and Wolford. (2009). "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction."

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."

The Perils of Badly Done Data-Driven Discovery

Bennett, Baird, Miller and Wolford. (2009). "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction."

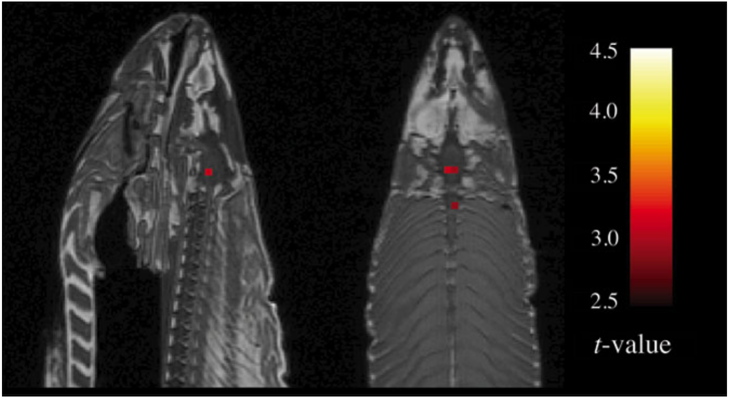
- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."

The Perils of Badly Done Data-Driven Discovery

Bennett, Baird, Miller and Wolford. (2009). "Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction."

- Subject
  "One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning."
- Task
  "The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."
- Design
  "Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes."

Results



"Several active voxels were discovered in a cluster located within the salmon's brain cavity. The size of this cluster was 81 mm^3 with a cluster-level significance of $p = .001$."

Breaking Free of the Scarcity Mindset

Breaking Free of the Scarcity Mindset

Deductive Model

Theory/Model

↓

Hypotheses

↓

Data Collection & Analysis

↓

Results

Iterative Model

Data Collection → Data Analysis → Theory/Model → Data Collection

↓

Hypotheses

↓

Data Collection & Analysis

↓

Results

Roots of Machine Learning in Artificial Intelligence

Roots of Machine Learning in Artificial Intelligence

Original artificial intelligence was rule-based.

Roots of Machine Learning in Artificial Intelligence

Original artificial intelligence was rule-based.

Core insight machine learning—we can train by example if we have enough data.

The Culture of Machine Learning

Machine Learning:

The Culture of Machine Learning

Machine Learning:

- Prioritizes prediction in new datasets

The Culture of Machine Learning

Machine Learning:

- Prioritizes prediction in new datasets
- Often values performance over statistical theory

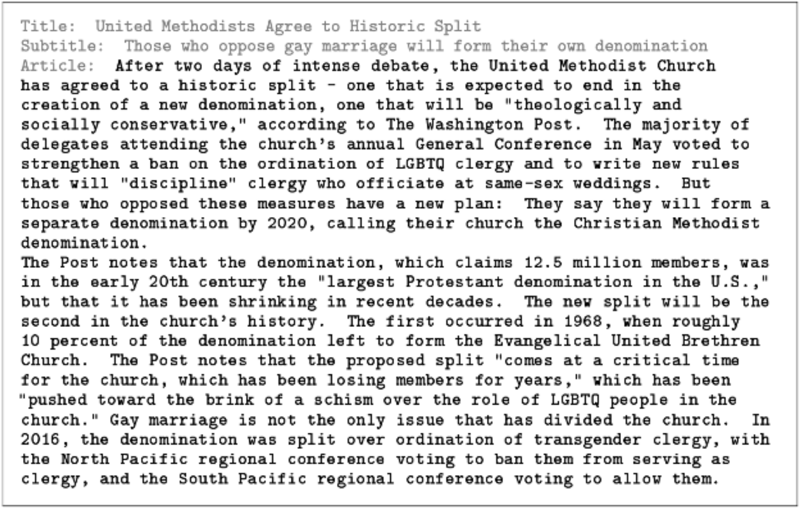
The Culture of Machine Learning

Machine Learning:

- Prioritizes prediction in new datasets
- Often values performance over statistical theory
- Tends to choose well-specified tasks

Amazing Things Are Possible

Amazing Things Are Possible


Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Image Credit: GPT3

Amazing Things Are Possible

Image Credit: Brundage et al 2018, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation"

However. . .


Tony "Abolish (Pol)ICE" Arcieri @bascule · Sep 19
Trying a horrible experiment...
Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?
2.8K
80.3K
197.9K

However. . .

Key Insights

Social Science: the task and the procedure to collect the data are vital for interpretation.

Machine Learning: learn whatever you can from the data. Don't limit yourself at the outset to what a human can imagine.

Overfitting and the Curse of Dimensionality

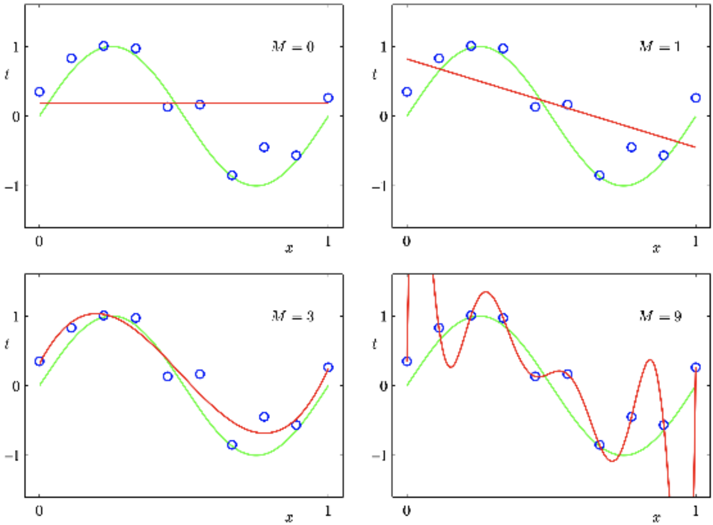


Image Credit: Bishop *Pattern Recognition and Machine Learning*

Overfitting and the Curse of Dimensionality

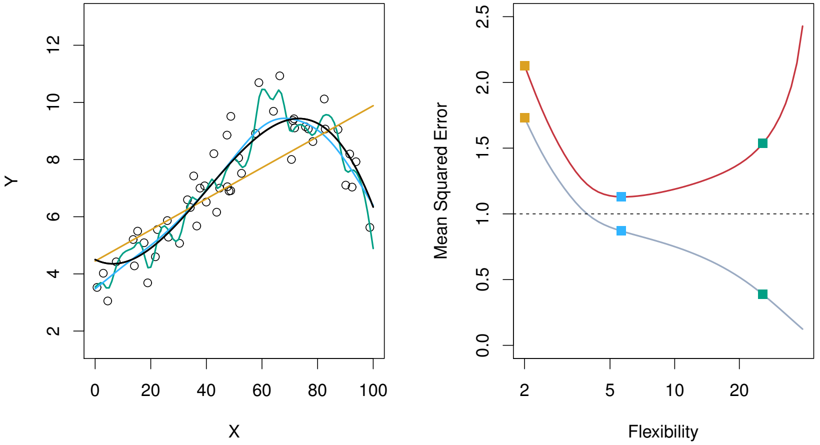


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Overfitting and the Curse of Dimensionality

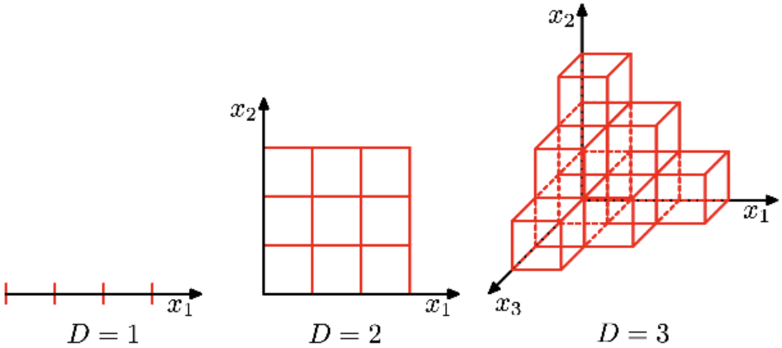
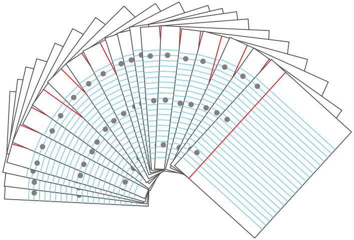


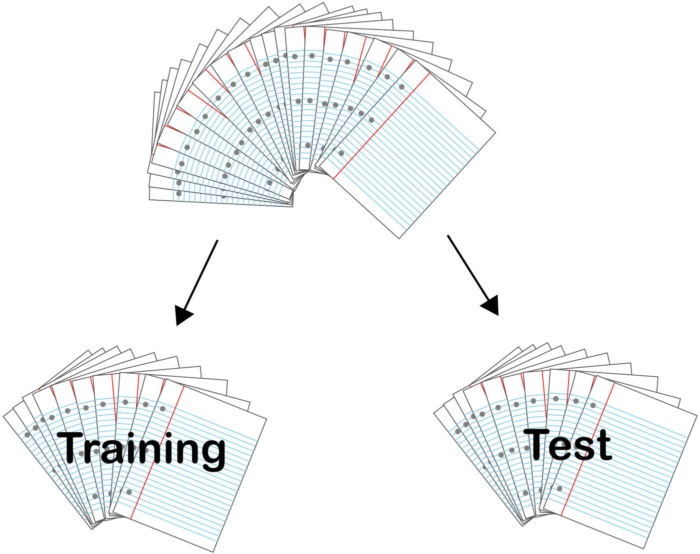
Image Credit: Bishop *Pattern Recognition and Machine Learning*

Tool 1: Sample Splitting

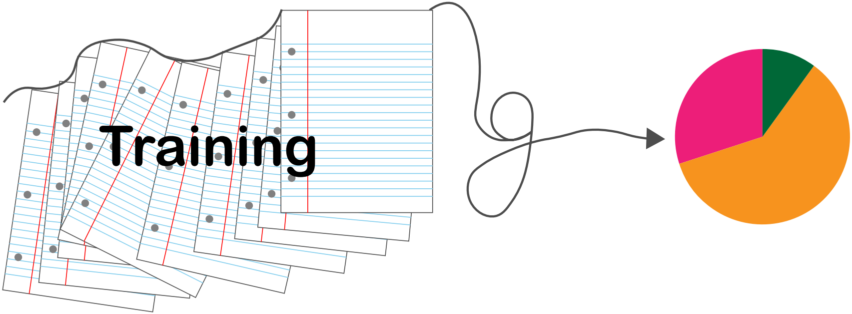
Tool 1: Sample Splitting



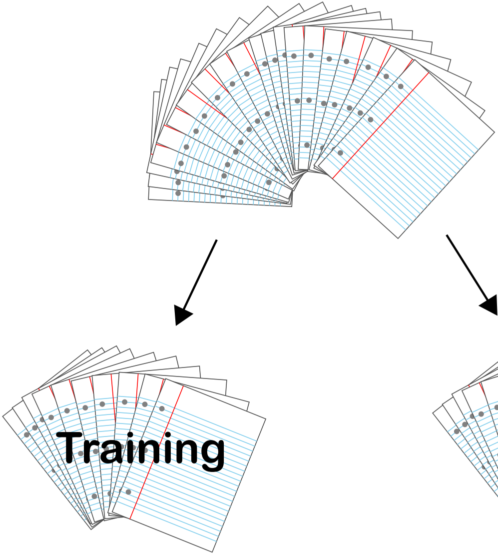
Tool 1: Sample Splitting



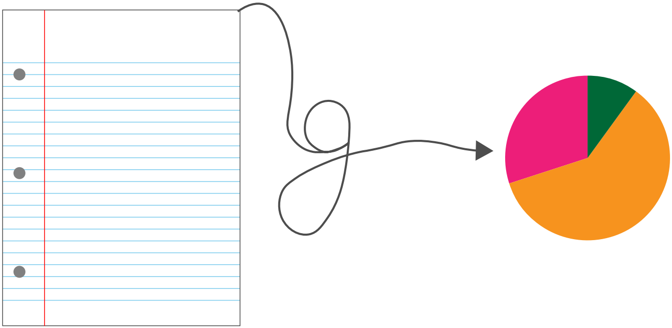
Tool 1: Sample Splitting



Tool 1: Sample Splitting



Tool 1: Sample Splitting



Tool 2: Cross Validation

Tool 2: Cross Validation

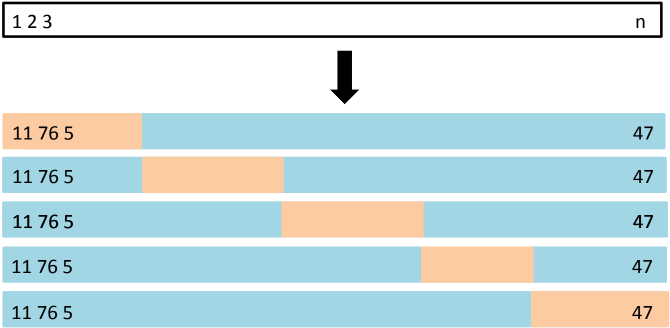


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Tool 3: Regularization

$$\arg\min_{\tilde{\beta}} \sum_i \left(y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j \right)^2 + \lambda \sum_{j=0}^{p} |\beta_j|^q$$

Tool 4: Hyperparameter Search

Tool 4: Hyperparameter Search

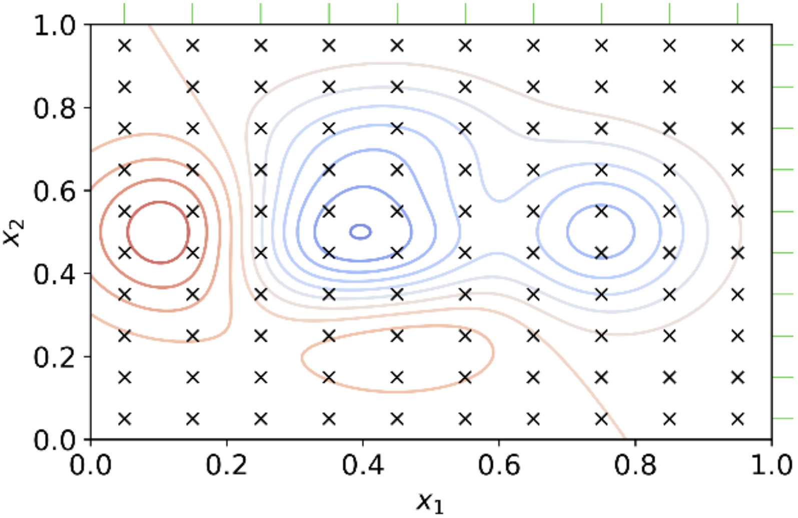


Image Credit: Wikipedia

Tool 5: Automatic Feature Engineering

Tool 5: Automatic Feature Engineering

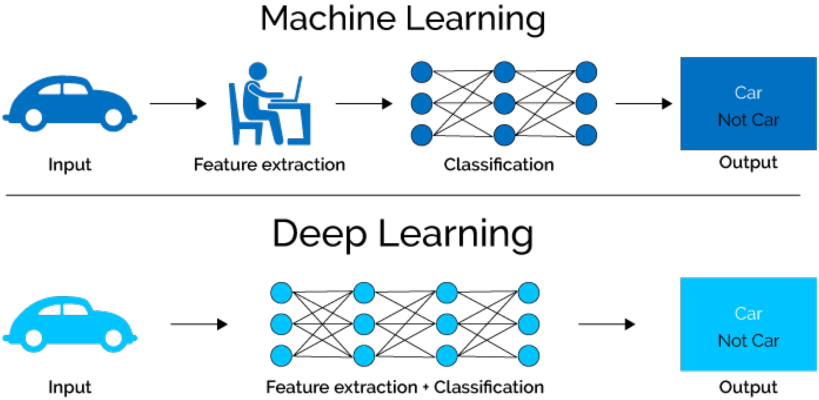


Image Credit: Towards Data Science, Sourish Dey

Tool 6: Reducing Dimensionality

Tool 6: Reducing Dimensionality

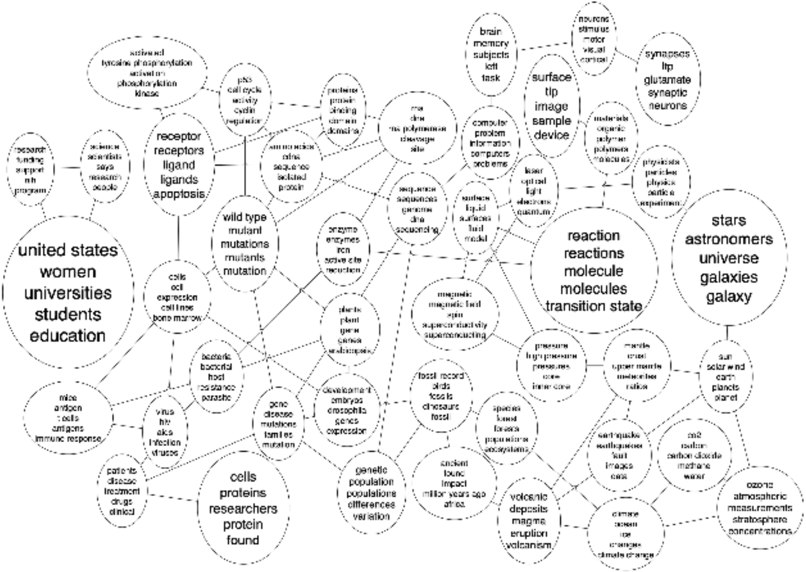
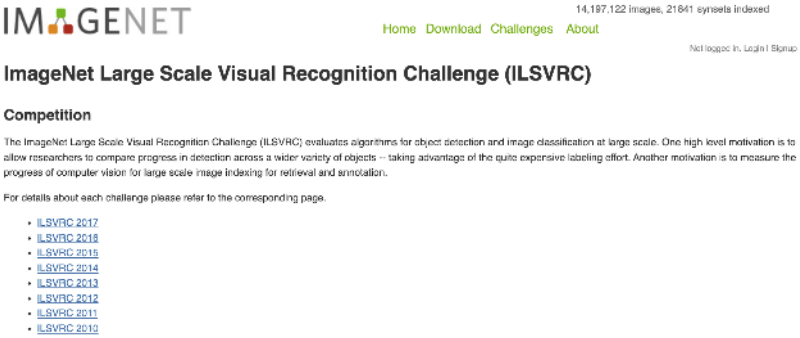


Image Credit: Dave Blei

Tool 7: The Common Task Framework

Tool 7: The Common Task Framework



IMAGENET

14,197,122 images, 21841 synsets indexed

Home Download Challenges About

Not logged in. Login | Signup

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Competition

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale. One high level motivation is to allow researchers to compare progress in detection across a wider variety of objects -- taking advantage of the quite expensive labeling effort. Another motivation is to measure the progress of computer vision for large scale image indexing for retrieval and annotation.

For details about each challenge please refer to the corresponding page.

- ILSVRC 2017
- ILSVRC 2016
- ILSVRC 2015
- ILSVRC 2014
- ILSVRC 2013
- ILSVRC 2012
- ILSVRC 2011
- ILSVRC 2010

Image Credit: ImageNet

A Note About Field Generalizations

- Generalizing about other fields is always a bit precarious. Machine Learning isn't one thing!

A Note About Field Generalizations

- Generalizing about other fields is always a bit precarious. Machine Learning isn't one thing!
- Any statements about 'social scientists' think like this or 'machine learning researchers' think like that should be taken as statements of 'many in this community think in this way.'

A Note About Field Generalizations

- Generalizing about other fields is always a bit precarious. Machine Learning isn't one thing!
- Any statements about 'social scientists' think like this or 'machine learning researchers' think like that should be taken as statements of 'many in this community think in this way.'
- Keep in mind that people often provide a charitable view of their own field and a critical view of others (but I'm going to try to highlight the best of all fields!).

Case Study: ChatGPT

THE DAILY PRINCETONIAN

Sunday, January 29
Previous Issues

MASTHEAD | ADVERTISE | CONTACT US NEWS OPINION SPORTS FEATURES THE PROSPECT MULTIMEDIA PODCAST CARTOON DATA PUZZLES HUMOR PROJECT

Follow us on Instagram

Try our latest crossword

NEWS

University declines to ban ChatGPT, releases faculty guidance for its usage

Language Modeling

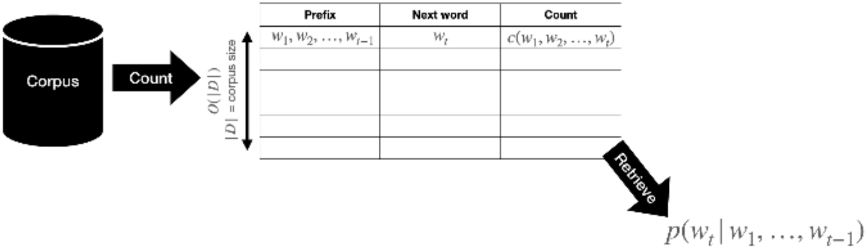


Image Credit: Kyunghyun Cho

Language Modeling with Neural Network Retrieval

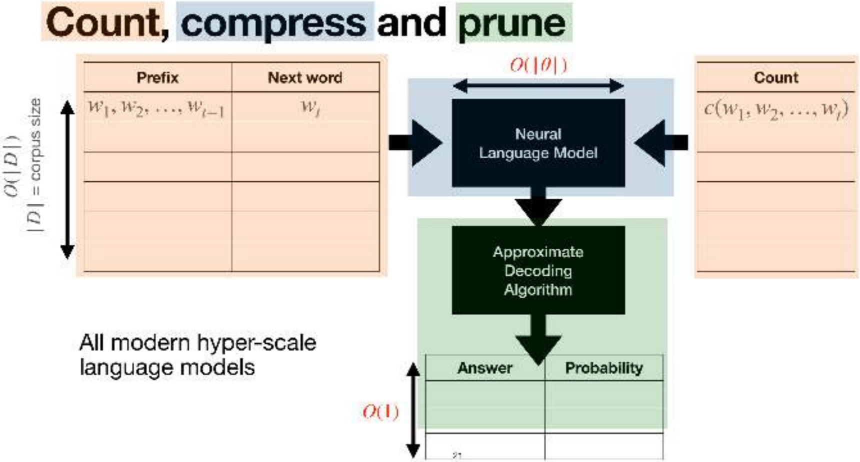


Image Credit: Kyunghyun Cho

Methods: Where We've Been and Where We're Going...

Methods: Where We've Been and Where We're Going...

- This Week
  - review of linear regression
  - introduction to basis functions

Methods: Where We've Been and Where We're Going...

- This Week
  - review of linear regression
  - introduction to basis functions
- Next Week
  - logistic regression
  - regularization

Methods: Where We've Been and Where We're Going...

- This Week
  - review of linear regression
  - introduction to basis functions
- Next Week
  - logistic regression
  - regularization
- Preview of Future Weeks
  - trees $\rightarrow$ boosting $\rightarrow$ causal identification frameworks $\rightarrow$ neural networks

Linear Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

Linear Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

Three Pieces:

- Loss Function

Linear Regression

$$\hat{\beta} = \arg\min_{\tilde{\beta}} \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

Three Pieces:

- Loss Function
- Prediction Function

Linear Regression

$$\hat{\beta} = \underset{\tilde{\beta}}{\arg\min} \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

Three Pieces:

- Loss Function
- Prediction Function
- Optimization

1) Loss Function

$$\sum_i (y_i - \hat{y}_i)^2$$

- large errors are magnified

1) Loss Function

$$\sum_i (y_i - \hat{y}_i)^2$$

- large errors are magnified
- very sensitive to outliers and leverage points

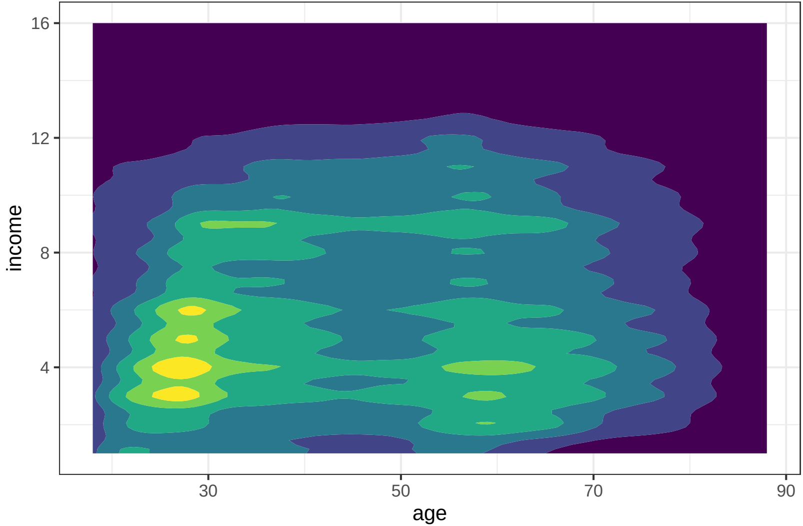
1) Loss Function

$$\sum_i (y_i - \hat{y}_i)^2$$

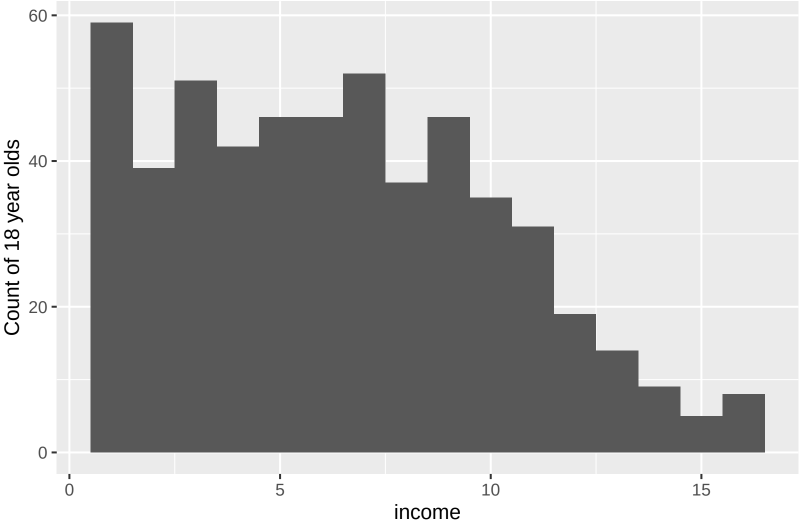
- large errors are magnified
- very sensitive to outliers and leverage points
- targets the conditional expectation function (CEF) $E[Y|X]$

CEF as the Best Prediction

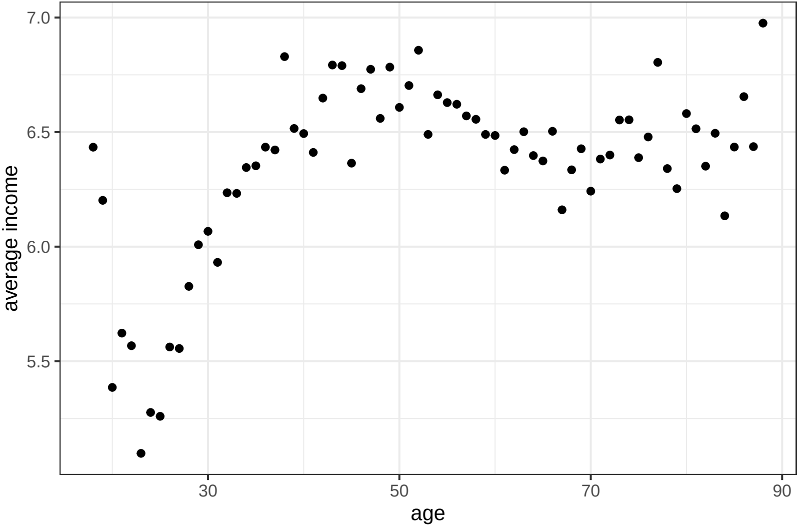
CEF as the Best Prediction



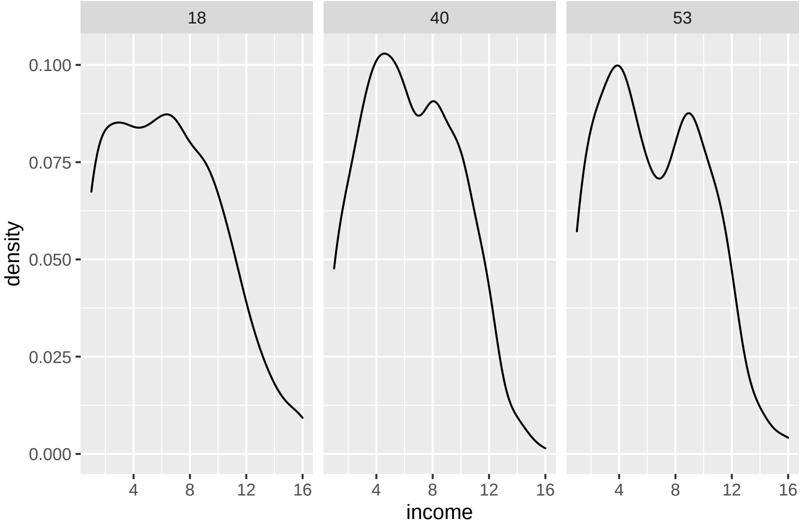
CEF as the Best Prediction



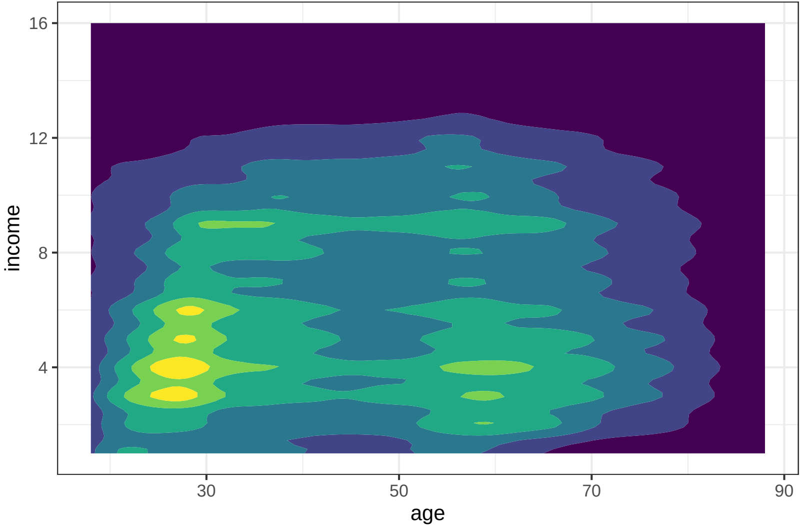
CEF as the Best Prediction



CEF as the Best Prediction



CEF as the Best Prediction



2) Prediction Function

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

2) Prediction Function

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

- linear function of $X \rightsquigarrow$ simple prediction and efficient.

2) Prediction Function

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

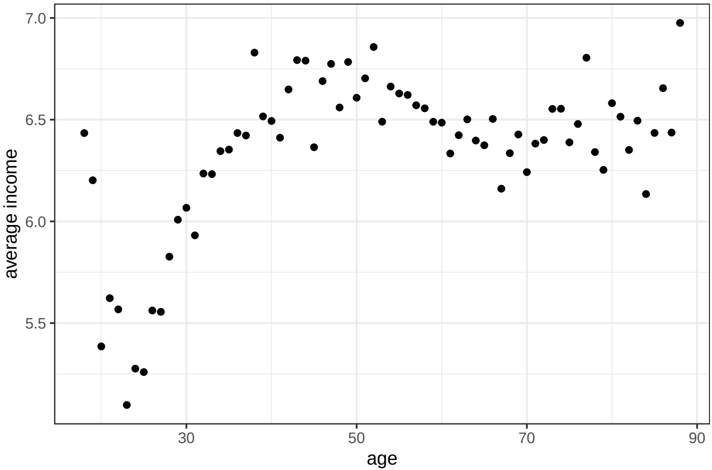
- linear function of $X \rightsquigarrow$ simple prediction and efficient.
- if true CEF is non-linear, we will find the best linear predictor (BLP).

2) Prediction Function

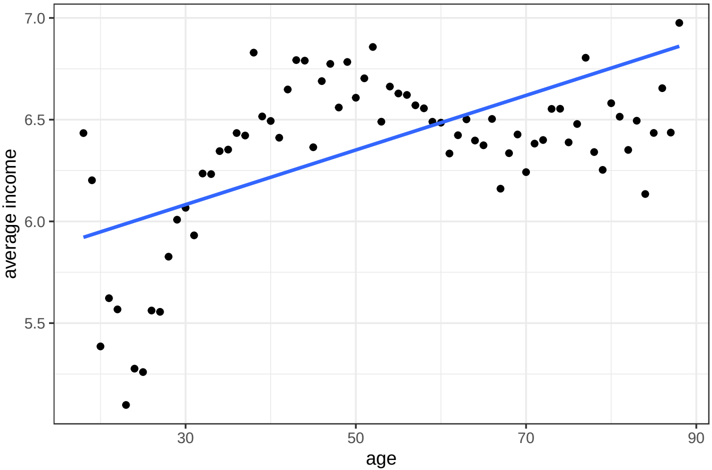
$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j$$

- linear function of $X \rightsquigarrow$ simple prediction and efficient.
- if true CEF is non-linear, we will find the best linear predictor (BLP).
- with multiple X variables the prediction surface is a plane.

Linear Approximations



Linear Approximations



Linear Approximations

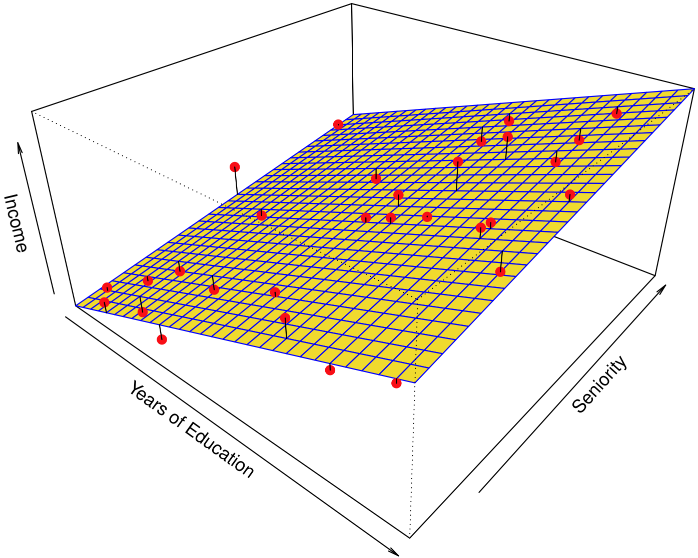


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Breaking Down Sources of Error

Breaking Down Sources of Error

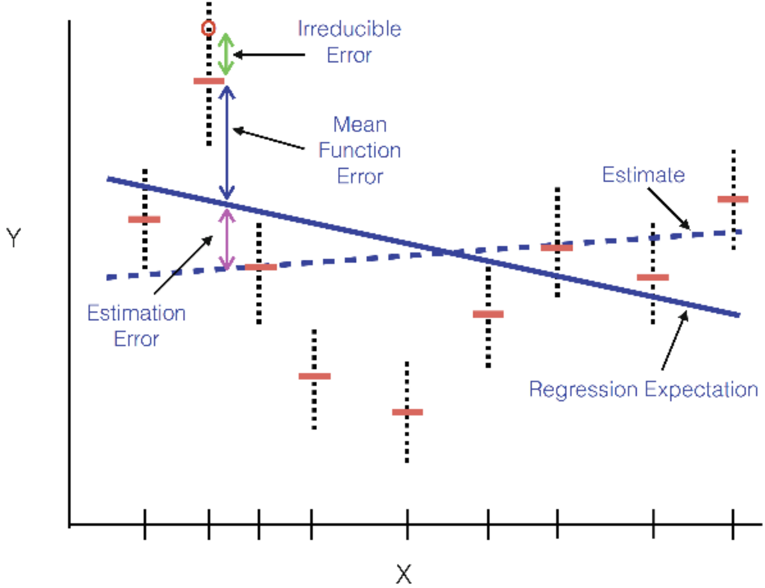


Image Credit: Berk (2020) *Statistical learning from a regression perspective*

BLP Approximations Depend on the Marginal Distribution of X

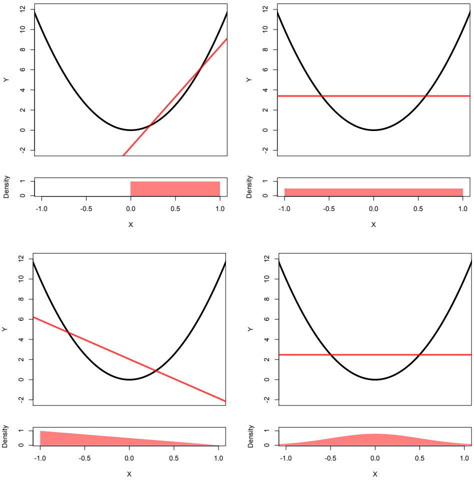


Figure 2.2.2: Plotting the CEF and the BLP over Different Distributions of X

Image Credit: Aronow and Miller (2019) *Foundations of Agnostic Statistics*

3) Optimization

$$\hat{\beta} = \underset{\tilde{\beta}}{\arg\min} \sum_i \left(y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j \right)^2$$

3) Optimization

$$\hat{\beta} = \underset{\tilde{\beta}}{\arg\min} \sum_i \left(y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j \right)^2$$

- closed-form solution $(X'X)^{-1}X'y$

3) Optimization

$$\hat{\beta} = \underset{\tilde{\beta}}{\arg\min} \sum_i \left(y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} X_{i,j} \tilde{\beta}_j \right)^2$$

- closed-form solution $(X'X)^{-1}X'y$
- solvable by gradient descent

3) Optimization

$$\hat{\beta} = \operatorname*{arg\,min}_{\tilde{\beta}} \sum_i \left(y_i - \tilde{\beta}_0 - \sum_{j=1}^{p} X_{i,j}\tilde{\beta}_j \right)^2$$

- closed-form solution $(X'X)^{-1}X'y$
- solvable by gradient descent

Overall the takeaway is that this optimization problem is very easy for $p < n$ and has infinite solutions if not.

Optimization Surface

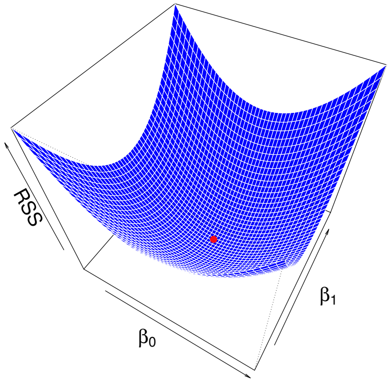


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

By Contrast. . .

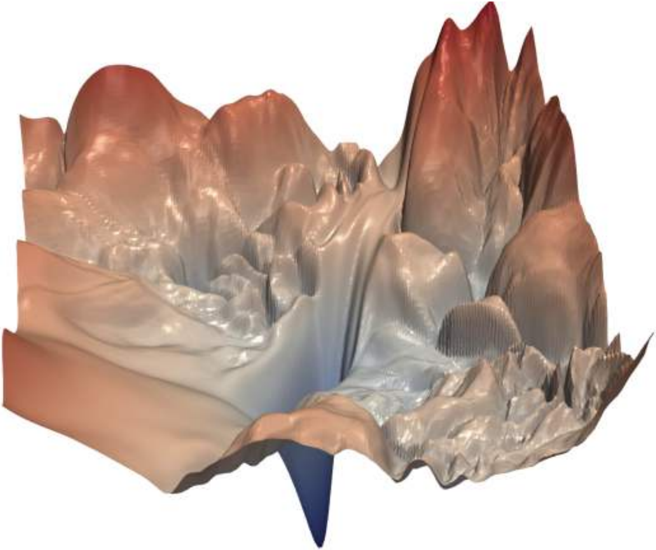


Image Credit: Li et al. 2018 "Visualizing the Loss Landscape of Neural Networks"

Implementation

Implementation

- `lm()` computes linear regression

Implementation

- `lm()` computes linear regression
- `biglm()` in the `biglm` package can handle really giant regressions

Implementation

- `lm()` computes linear regression
- `biglm()` in the `biglm` package can handle really giant regressions
- `predict()` will generate predictions.

Non-Linearity

- Linearity is highly restrictive.

Non-Linearity

- Linearity is highly restrictive.
- Imagine we want to predict salary based on years of education and seniority.

Non-Linearity

- Linearity is highly restrictive.
- Imagine we want to predict salary based on years of education and seniority.
- Linearity requires that:
  - each marginal year of education produces the same increase in income

Non-Linearity

- Linearity is highly restrictive.
- Imagine we want to predict salary based on years of education and seniority.
- Linearity requires that:
  - each marginal year of education produces the same increase in income
  - years of education and seniority can't be interactive

Non-Linearity

- Linearity is highly restrictive.
- Imagine we want to predict salary based on years of education and seniority.
- Linearity requires that:
  - each marginal year of education produces the same increase in income
  - years of education and seniority can't be interactive
- Essentially all gains in modern machine learning come from either allowing non-linearity, allowing more variables, or both.

Polynomials

- You may have already seen ways to add non-linearity in a single variable through polynomials.

$$\hat{y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

Polynomials

- You may have already seen ways to add non-linearity in a single variable through polynomials.

$$\hat{y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

- This allows us to fit a broader range of functions than just a line.

Polynomials

- You may have already seen ways to add non-linearity in a single variable through polynomials.

$$\hat{y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

- This allows us to fit a broader range of functions than just a line.
- The order-2 polynomial here let's us form any parabola

Polynomials

- You may have already seen ways to add non-linearity in a single variable through polynomials.

$$\hat{y}_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$$

- This allows us to fit a broader range of functions than just a line.
- The order-2 polynomial here let's us form any parabola
- Higher order polynomials let us fit increasingly complex functions.

Example Polynomials

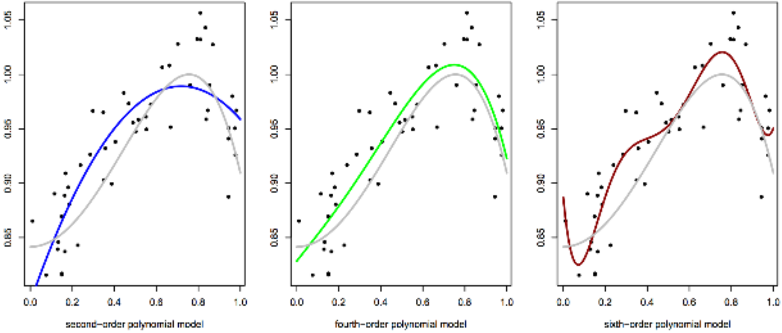


Image Credit: Radford Neal

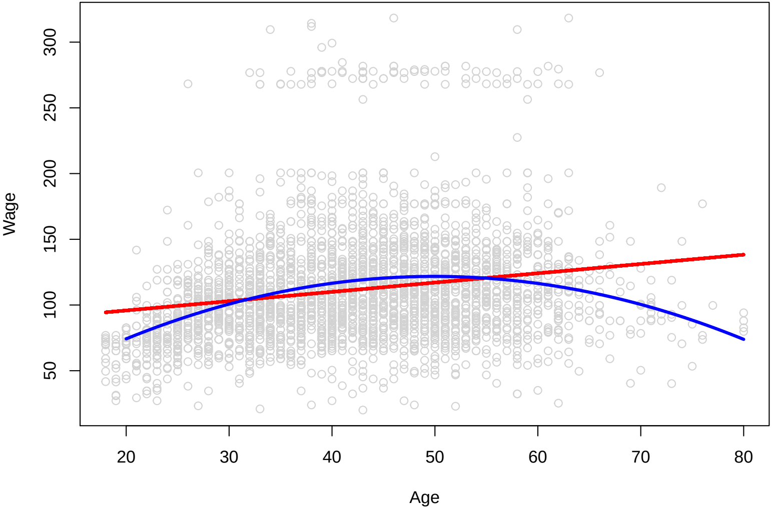
Implementation in R

This is super easy in R!

```
library(ISLR)
m1 <- lm(wage~age,data=Wage)
m2 <- lm(wage~poly(age,degree=2), data=Wage)
```

NB: `poly(x, raw=FALSE)` vs. `poly(x,raw=TRUE)`

Polynomial



Linear Basis Function Models

Linear Basis Function Models

- Using polynomials this way is an instance of a more general idea: a linear basis function model.

Linear Basis Function Models

- Using polynomials this way is an instance of a more general idea: a linear basis function model.
- The idea is to do a linear regression of y on $\phi_1(X), \phi_2(X), \dots, \phi_M(X)$ where ϕ_j are basis functions.

Linear Basis Function Models

- Using polynomials this way is an instance of a more general idea: a linear basis function model.
- The idea is to do a linear regression of y on $\phi_1(X), \phi_2(X), \ldots, \phi_M(X)$ where ϕ_j are basis functions.
- The prediction function is now:

$$\hat{y} = \tilde{\beta}_0 + \sum_{m=1}^{M} \tilde{\beta}_m \phi_m(X)$$

Linear Basis Function Models

- Using polynomials this way is an instance of a more general idea: a linear basis function model.
- The idea is to do a linear regression of y on $\phi_1(X), \phi_2(X), \ldots, \phi_M(X)$ where ϕ_j are basis functions.
- The prediction function is now:

$$\hat{y} = \tilde{\beta}_0 + \sum_{m=1}^{M} \tilde{\beta}_m \phi_m(X)$$

- The polynomial defined $\phi_m(X) = X^m$.

Linear Basis Function Models

- Using polynomials this way is an instance of a more general idea: a linear basis function model.
- The idea is to do a linear regression of y on $\phi_1(X), \phi_2(X), \ldots, \phi_M(X)$ where ϕ_j are basis functions.
- The prediction function is now:

$$\hat{y} = \tilde{\beta}_0 + \sum_{m=1}^{M} \tilde{\beta}_m \phi_m(X)$$

- The polynomial defined $\phi_m(X) = X^m$.
- Today we will consider fixed basis functions that are set by the analyst in advance of the analysis.

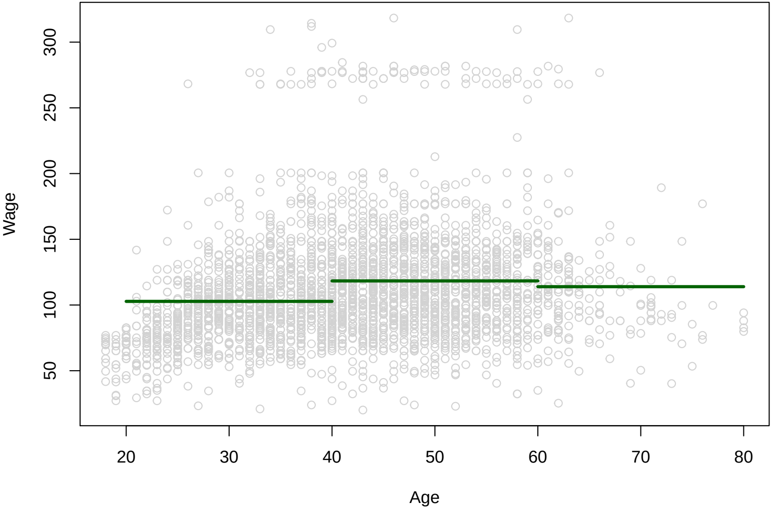
Linear Basis Function Models

- Using polynomials this way is an instance of a more general idea: a linear basis function model.
- The idea is to do a linear regression of y on $\phi_1(X), \phi_2(X), \ldots, \phi_M(X)$ where ϕ_j are basis functions.
- The prediction function is now:

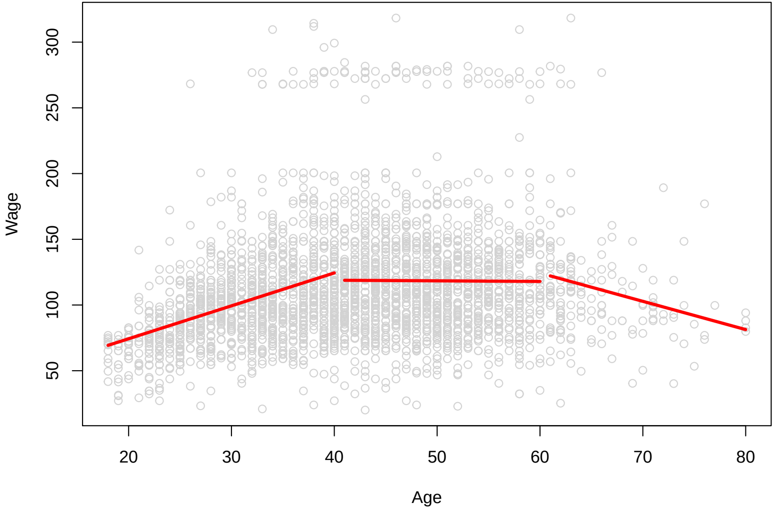
$$\hat{y} = \tilde{\beta}_0 + \sum_{m=1}^{M} \tilde{\beta}_m \phi_m(X)$$

- The polynomial defined $\phi_m(X) = X^m$.
- Today we will consider fixed basis functions that are set by the analyst in advance of the analysis.
- Many techniques in machine learning can be expressed as linear models with adaptive basis functions (i.e. learned from the data during the optimization procedure).

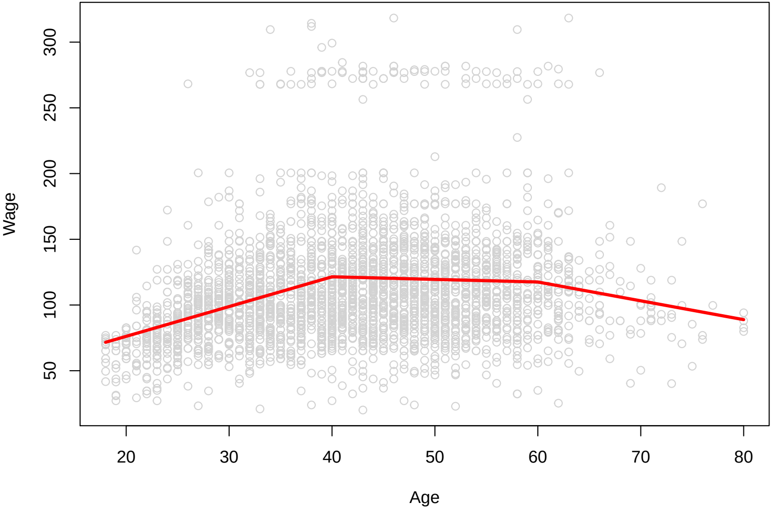
Piecewise Constant



Piecewise Linear



Piecewise Linear (with Continuity Constraint)



Splines

- Splines are flexible basis functions that extend the continuous piecewise approach.

Splines

- Splines are flexible basis functions that extend the continuous piecewise approach.
- They will allow us to combine the polynomial strategy with the piecewise strategy!

Splines

- Splines are flexible basis functions that extend the continuous piecewise approach.
- They will allow us to combine the polynomial strategy with the piecewise strategy!
- They extend the notion of continuity to be continuity in some number of derivatives of the function.

Splines

- Splines are flexible basis functions that extend the continuous piecewise approach.
- They will allow us to combine the polynomial strategy with the piecewise strategy!
- They extend the notion of continuity to be continuity in some number of derivatives of the function.

Definition (d-Degree Spline)

A d-degree spline is a piecewise polynomial with degree d and continuity in all derivates less than d.

Splines

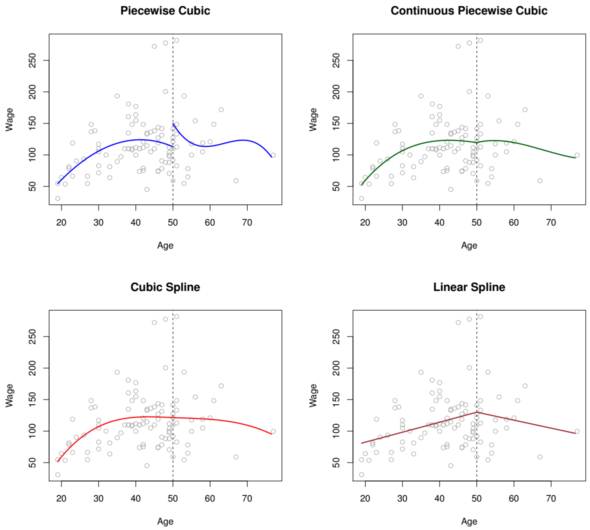


Image Credit: James et al. 2013 *An Introduction to Statistical Learning, with applications in R.*

Splines: Origin of the Name

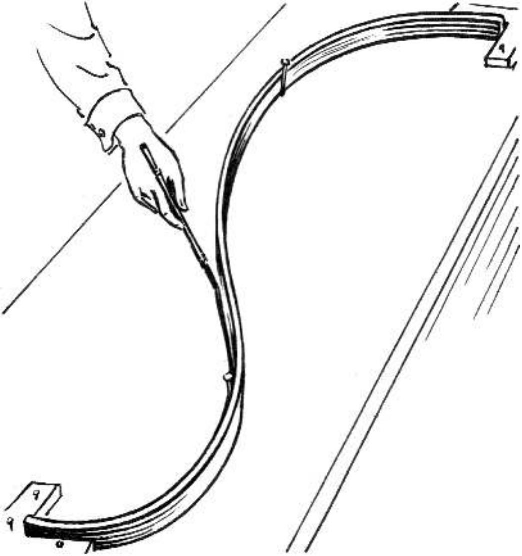


Image Credit: Pearson Scott Foresman via Wikipedia

Things to Know About Splines

Things to Know About Splines

- The points of discontinuity are called the knots of the spline.

Things to Know About Splines

- The points of discontinuity are called the knots of the spline.
- The more knots you have the more flexible the function can be.

Things to Know About Splines

- The points of discontinuity are called the knots of the spline.
- The more knots you have the more flexible the function can be.
- By convention knots are typically placed uniformly over the quantiles of the data.

Things to Know About Splines

- The points of discontinuity are called the knots of the spline.
- The more knots you have the more flexible the function can be.
- By convention knots are typically placed uniformly over the quantiles of the data.
- Most people use cubic splines because the knots tend to look very smooth to humans (because you can't see discontinuities in the third derivative easily!).

Things to Know About Splines

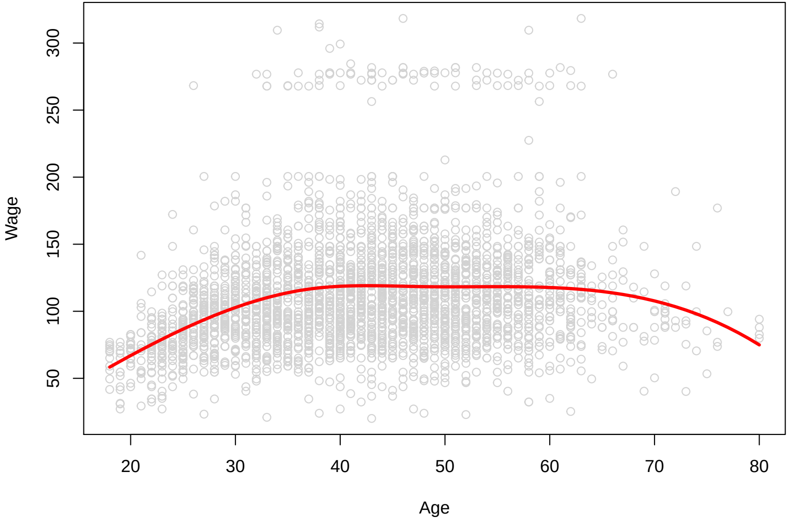
- The points of discontinuity are called the knots of the spline.
- The more knots you have the more flexible the function can be.
- By convention knots are typically placed uniformly over the quantiles of the data.
- Most people use cubic splines because the knots tend to look very smooth to humans (because you can't see discontinuities in the third derivative easily!).
- In `R` you often build the spline basis functions using degrees of freedom which are (`degree of polynomial + number of knots`) and it will place the knots automatically at evenly spaced quantiles.

Implementation in R

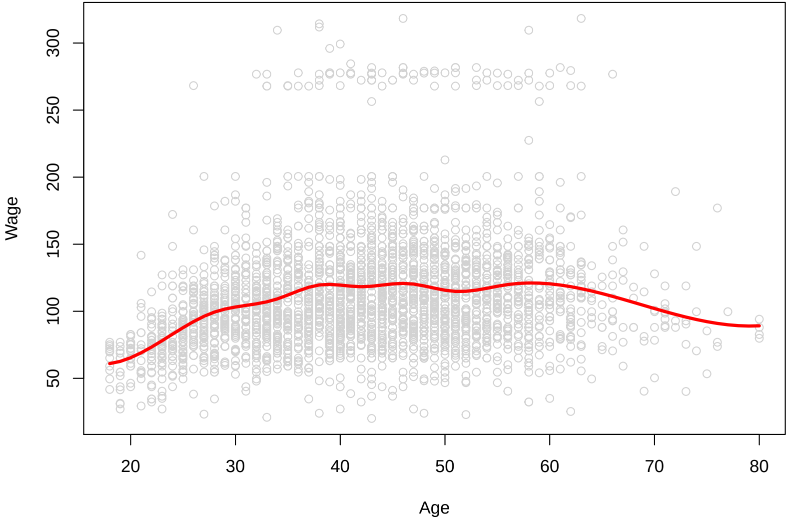
This is super easy in R!

```
library(ISLR)
library(splines)
m3 <- lm(wage~bs(age, df=5),data=Wage)
```

Cubic Spline with 5 Degrees of Freedom



Cubic Spline with 10 Degrees of Freedom



Limitations

Limitations

- This really only works well for one variable so it doesn't help much with interactions between the variables.

Limitations

- This really only works well for one variable so it doesn't help much with interactions between the variables.
- Thin-plate splines can generalize this idea to two or three variables, but it all stops working after that due to the curse of dimensionality.

Limitations

- This really only works well for one variable so it doesn't help much with interactions between the variables.
- Thin-plate splines can generalize this idea to two or three variables, but it all stops working after that due to the curse of dimensionality.
- Using more knots means the spline is more flexible but it also can be more likely to overfit (more on how to address this next week!).

Why Did We Do This?

Why Did We Do This?

- Linear basis function models are a way to get the ease of a linear model with the flexibility of non-linear modeling!

Why Did We Do This?

- Linear basis function models are a way to get the ease of a linear model with the flexibility of non-linear modeling!
- Many types of machine learning we will discuss later (e.g. random forests, neural networks) are just ways of adaptively creating the basis function to fit the data.

Why Did We Do This?

- Linear basis function models are a way to get the ease of a linear model with the flexibility of non-linear modeling!
- Many types of machine learning we will discuss later (e.g. random forests, neural networks) are just ways of adaptively creating the basis function to fit the data.
- These models can seem really complex, but it is useful to know that we can just see them as a linear model where the 'predictors' have been transformed in the same way we did with polynomials and splines.

This Week in Review

- Introduction to the Course
- What is Machine Learning?
- 7 Tools of Machine Learning
- Methods: Regression and Basis Functions

This Week in Review

- Introduction to the Course
- What is Machine Learning?
- 7 Tools of Machine Learning
- Methods: Regression and Basis Functions

Reading Preview for Discovery and Measurement:

Salganik Chapter 1: Introduction

GRS Chapter 2 Social Science Research and Text Analysis

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366, no. 6464: 447-453.

Selections from JWHT