Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 2: Survey of Methods

Jan 19, 2016


Linear regression
Deep learning
Decision trees
Ordinal regression
Probabilistic graphical models
Random forests
Logistic regression
Networks
Support vector machines
Topic models
Survival models
K-means clustering
Neural networks
Hierarchical clustering
Perceptron

Classification

A mapping *h* from input data x (drawn from instance space $\mathcal{X}$) to a label (or labels) y from some enumerable output space $\mathcal{Y}$

$\mathcal{X}$ = set of all skyscrapers
$\mathcal{Y}$ = {art deco, neo-gothic, modern}

x = the empire state building
y = art deco

Classification

h(x) = y
h(empire state building) = art deco

Classification

Let $h(x)$ be the "true" mapping. We never know it. How do we find the best $\hat{h}(x)$ to approximate it?

One option: rule based

if x has "sunburst motif":
    $\hat{h}(x)$ = art deco

Classification

Supervised learning

Given training data in the form of <x, y> pairs, learn $\hat{h}(x)$

task	$\mathcal{X}$	$\mathcal{Y}$
spam classification	email	{spam, not spam}
authorship attribution	text	{jk rowling, james joyce, …}
genre classification	song	{hip-hop, classical, pop, …}
image tagging	image	{B&W, color, ocean, fun, …}

Methods differ in form of $\hat{h}(x)$ learned

Model differences

- Binary classification: $|\mathcal{Y}| = 2$
  [one out of 2 labels applies to a given x]
- Multiclass classification: $|\mathcal{Y}| > 2$
  [one out of N labels applies to a given x]
- Multilabel classification: $|y| > 1$
  [multiple labels apply to a given x]

Regression

A mapping from input data x (drawn from instance space $\mathcal{X}$) to a point y in $\mathbb{R}$

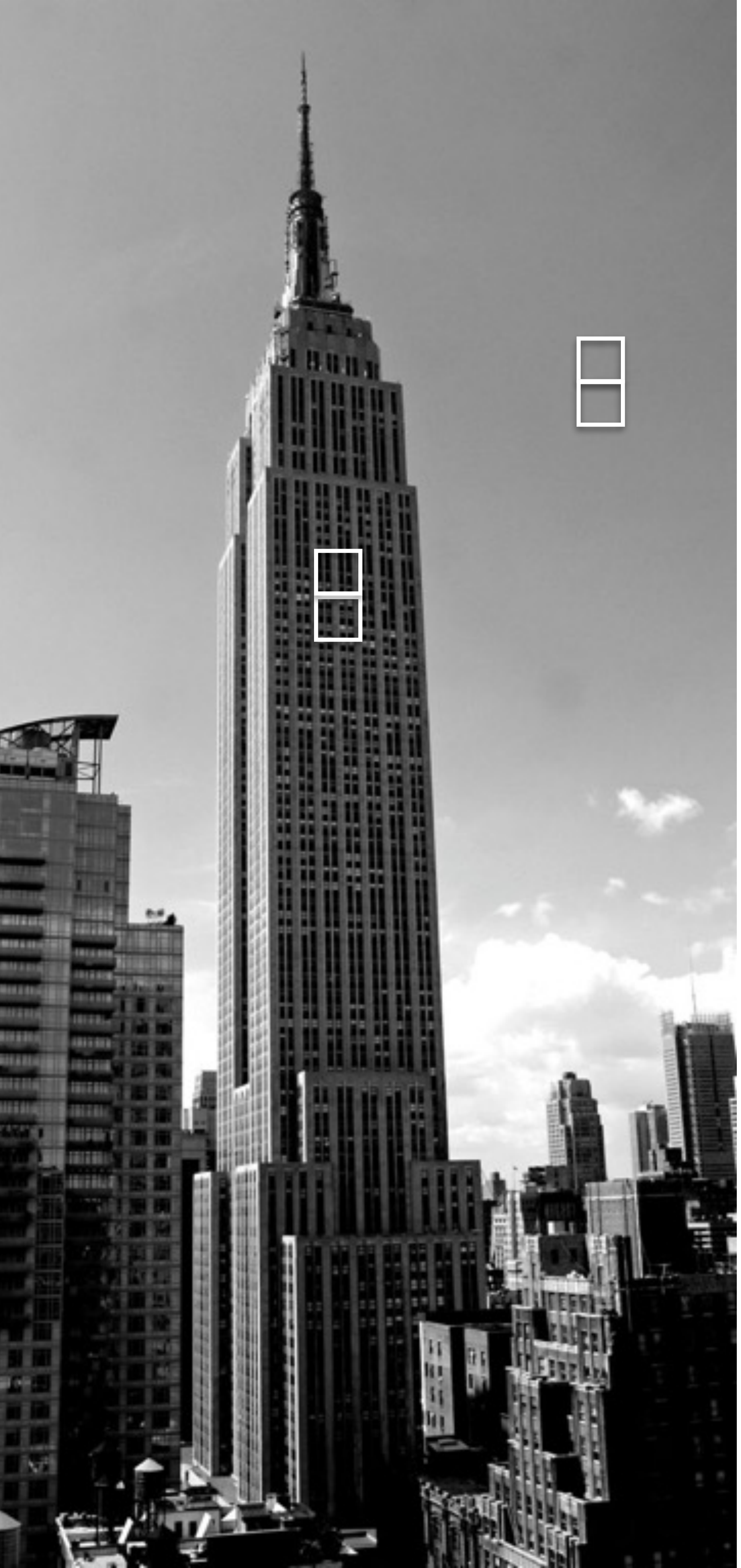
($\mathbb{R}$ = the set of real numbers)

x = the empire state building
y = 17444.6"


Linear regression
Deep learning
Decision trees
Ordinal regression
Probabilistic graphical models
Random forests
Networks
Support vector machines (regression)
Survival models
Neural networks
Perceptron

Big differences

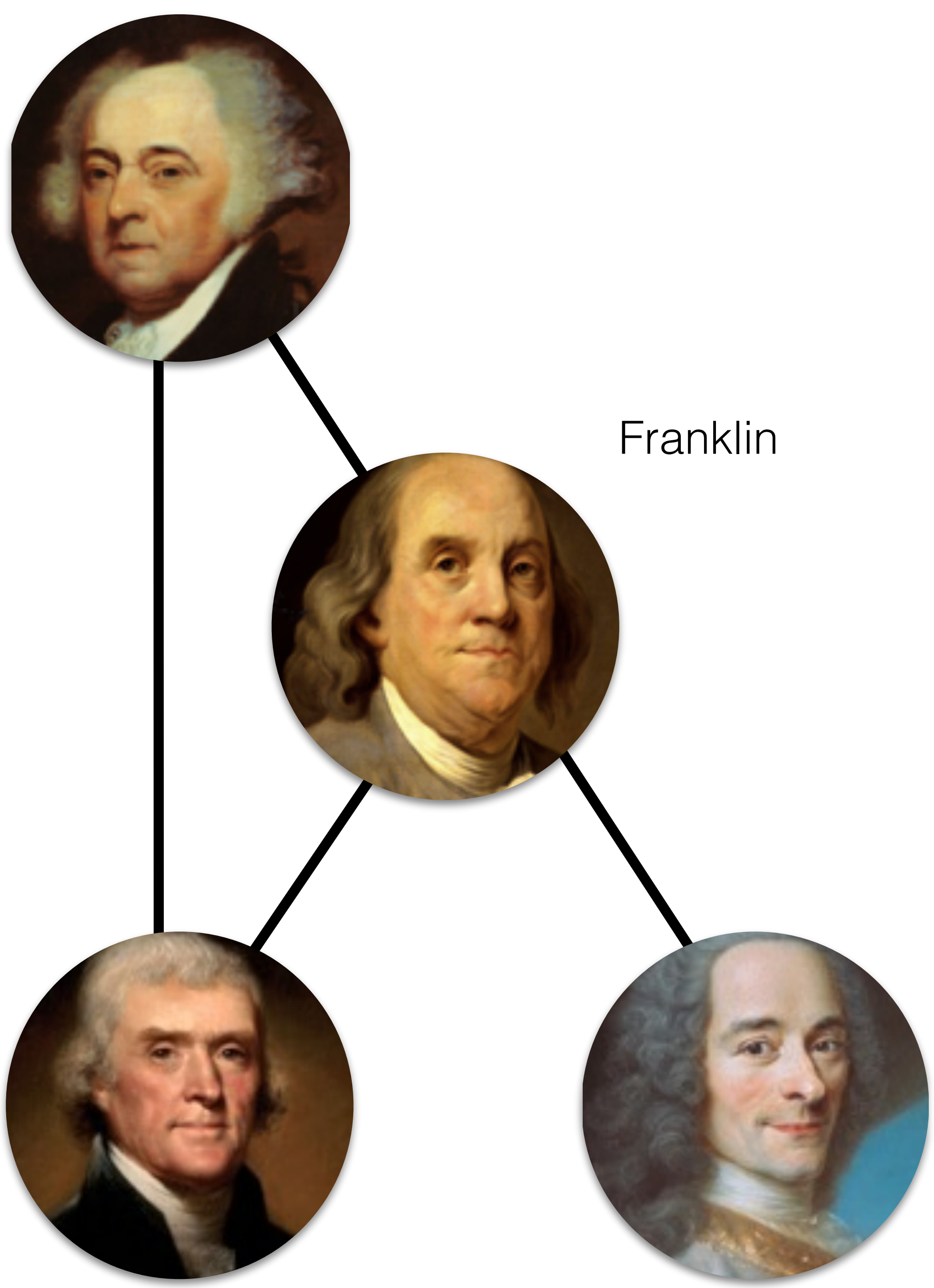
- Are the labels y_j and y_k for two different data points x_j and x_k independent? During learning and prediction, would your guess for y_j help you predict y_k?



Label dependence

- Object recognition in images
- Neighboring pixels tend to have similar values (building, sky)

Label dependence



- Homophily in social networks
- Friends to have similar attribute values

Big differences

- Are the labels y_j and y_k for two different data points x_j and x_k independent? During learning and prediction, would your guess for y_j help you predict y_k?

- [Part of speech tagging, network homophily, object recognition in images]

- Sequence models (HMMs, CRFS, LSTMs) and general graphical models (MRFs) but come at a high computational cost

Big differences

- How do the features in x interact with each other?
    - Independent? [Naive Bayes]
    - Potentially correlated but non-interacting? [Logistic regression, linear regression, perceptron, linear SVM]
    - Complex interactions? [Non-linear SVM, neural networks, decision trees, random forests]

Feature interactions

training data

I like the movie	1
I hate the movie	-1
I do not like the movie	-1
I do not hate the movie	1

how predictive is:

- like
- hate
- not
- not like
- not hate

What do you need?

1. Data (emails, texts)
2. Labels for each data point (spam/not spam, which author it was written by)
3. A way of “featurizing" the data that’s conducive to discriminating the classes
4. To know that it works.

What do you need?

Two steps to building and using a supervised classification model.

1. Train a model with data where you know the answers.

2. Use that model to predict data where you don't.

Recognizing a Classification Problem

- Can you formulate your question as a *choice* among some universe of possible classes?

- Can you create (or find) labeled data that marks that choice for a bunch of examples? Can *you* make that choice?

- Can you create features that might help in distinguishing those classes?

Uses of classification

Two major uses of supervised classification/regression

Prediction

Train a model on a sample of data <x, y> to predict values for some new data x′

Interpretation

Train a model on a sample of data <x, y> to understand the relationship between x and y

Clustering

- Clustering (and unsupervised learning more generally) finds *structure* in data, using just X

X = a set of skyscrapers

What is structure?

- Unsupervised learning finds *structure* in data.
  - clustering data into groups
  - discovering “factors”

Methods differ in the kind of structure learned

Structure

- Partitioning *X* into N disjoint sets [K-means clustering, PGMs]
- Assigning *X* to hierarchical structure [Hiearchical clustering]
- Assigning *X* to partial membership in N different sets [EM clustering, PGMs, PCA]
- Learning a representation of *x* in *X* that puts similar data points close to each other [Deep learning]

Uses of clustering

Exploratory data analysis

- Discovering interesting or unexpected structure can useful for hypothesis generation

→ Input to supervised models

- Unsupervised learning generates alternate representations of each x as it relates to the larger X.

→ Input to supervised models

Brown clusters trained from Twitter data: every word is mapped to a single (hierarchical) cluster

^001010110 (29)	never neva nvr gladly nevr #never neverr nver neverrr nevaa nevah nva neverrrr letchu letcha ne'er -never neveer glady #inever bever nevaaa neever nerver enver neeever nevet neeeever nevva
^001010111 (23)	ever eva evar evr everrr everr everrrr evah everrrrr everrrrrr evaa evaaa everrrrrrr nevar eveer evaaaa eveeer everrrrrrrr everrrrrrrrr evea eveeeer evaaaaa evur
^00101100 (16)	only onli onlyy ony onlii 0nly -only olny onlyyy onlt onlly onyl onlu onlee onle inly

http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

Recognizing a Classification/Regression/Clustering Problem

- I want to predict a star value {1, 2, 3, 4, 5} for a product review
- I want to find all of the texts that have allusions to *Paradise Lost*.
- Optical character recognition
- I want to associate photographs of cats with animals in a taxonomic hierarchy
- I want to reconstruct an evolutionary tree for languages

boyd and Crawford

- danah boyd and Kate Crawford (2012), “Critical Questions for Big Data,” Information, Communication and Society

- Specifically about “big data” but we can read it as a commentary on much quantitative practice using social data

1. “big data” changes the definition of knowledge

- How do computational methods/quantitative analysis pragmatically affect epistemology?

- Restricted to what data is available (twitter, data that’s digitized, google books, etc.). How do we counter this in experimental designs?

- Establishes alternative norms for what “research” looks like

2. claims to objectivity and accuracy are misleading

- What is still subjective in data/empirical methods? What are the interpretive choices still to be made?
- Interpretation introduces dependence on individuals. Is this ever avoidable?
- What does an experiment (or results) “mean”?

2. claims to objectivity and accuracy are misleading

- Data collection, selection process is subjective, reflecting belief in what matters.

- Model design is likewise subjective

    - model choice (classification vs. clustering etc.)
    - representation of data
    - feature selection

- Claims need to match the sampling bias of the data.

3. bigger data is not always better data

- Uncertainty about its source or selection mechanism [Twitter, Google books]
- Appropriateness for question under examination
- How did the data you have get there? Are there other ways to solicit the data you need?
- Remember the value of small data: individual examples and case studies

4. taken out of context, big data loses it meaning

- A representation (through features) is a necessary approximation; what are the consequences of that approximation?

- Example: quantitative measures of "tie strength" and its interpretation (e.g., articulated, behavior, personal networks).

5. just because it is accessible does not make it ethical

- Twitter, Facebook, OkCupid
- Anonymization practices for sensitive data (even if born public)
- Accountability both to research practice and to subjects of analysis

6. limited access to "big data" creates new digital divides

- Inequalities in access to data and the production of knowledge

- Privileging of skills required to produce knowledge

Tuesday 1/24: Classification

- Bring examples of hard problems that would fall under the domain of classification, and how you could approach training data collection