STATS 415: Overview

Prof. Liza Levina

Department of Statistics, University of Michigan

What’s in the name?

- Our course is called “Data mining”
- Our book is called “Statistical Learning”
- Other names you hear: machine learning, artificial intelligence, data science, data analytics, predictive analytics, ... and of course statistics!
- The differences are mostly historical and cultural. Understanding them is not worth your time.

Learning from Data

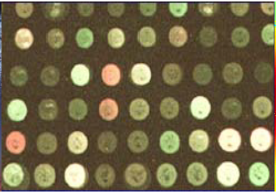
- Fact: The amount of data collected and stored is exponentially increasing, due to advances in data collection, computerization of many aspects of life and breakthroughs in storage technology.
- Consequence: Data analysis problems have dramatically increased in size and complexity.
- The data analyst's job: make sense of all these data! Identify patterns and trends, uncover "interesting" relationships among the variables and/or the observations, predict future behavior.

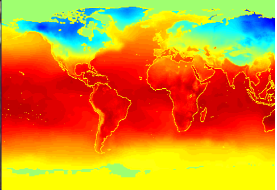
- Technology helps
  - Faster computers, more storage ⇒ more flexible and thus more powerful techniques ⇒ fewer modeling assumptions
  - New visualisation capabilities (a picture is worth a thousand words...)
- But not always!
  - Some problems are inherently computationally intractable
  - "Easy" black-box data analysis can lead to a lot of misuse and misunderstanding
  - Flexible models can overfit (too much of a good thing)
  - A famous example: Google Flu
  - Understanding underlying assumptions and interpreting conclusions correctly remains as important as ever

Learning from (big) data: Motivation

- There is often "hidden" information in the data that is not readily evident.
- Human analysts without large-scale algorithms may take months or years to discover useful information.
- Much of the data available are never analyzed at all.

Learning from data: Scientific viewpoint





- Data are collected and stored at enormous speeds
  - remote sensors on a satellite (NASA)
  - telescopes scanning the skies (SDSS)
  - microarrays generating gene expression data (MEDLINE)
  - scientific simulations generating terabytes of data (GIS)
- Statistical learning helps scientists with
  - classifying and clustering data
  - formulating hypotheses
  - validating hypotheses
  - predicting future behavior

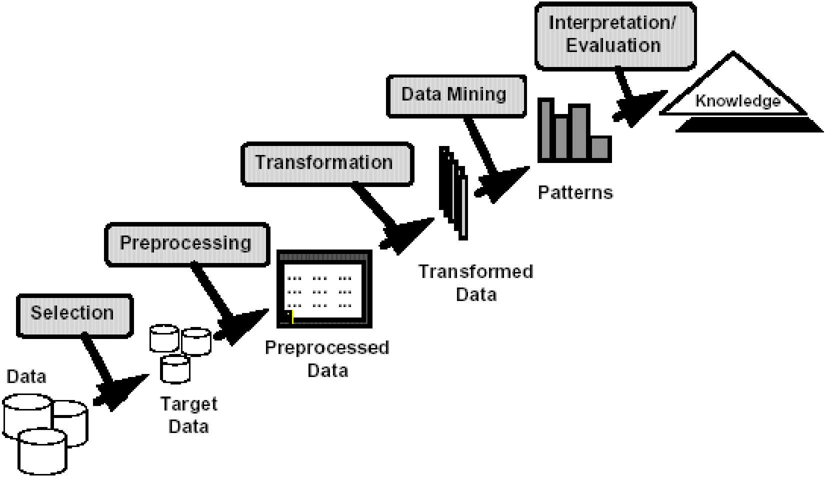
Learning from data: Commercial viewpoint

- Almost any commercial transaction generates data
  - Web searches, social connections (Google, Facebook, Twitter, etc)
  - Purchases both online and at stores (Amazon, eBay, Walmart)
  - Bank/credit card transactions (Bank of America, Visa, Mastercard)
  - Customized services and successful advertizing give competitive edge
  - Have to balance useful services vs annoyance and privacy concerns - a fine line!

Learning from data: Personal viewpoint

- We are bombarded with massive amounts of data
- We generate massive amounts of data (everything you do on your phone is data!)
- Humans are great at spotting and recognizing patterns, but also easy to trick
- If used right, data can help you (fitness trackers, good movie recommendations)
- If used wrong, data can harm you (link between social media and depression, fake news)
- Learning to critically think about data is one of the most imporant skills in the modern world

The process of learning from data



Some basic notation

- n: the number of observations (cases, data points)
- p: the number of variables (features, predictors)
- The variables can be quantitative, categorical, or a mix
- Data matrix: $n \times p$ matrix $X = \{X_1, \dots, X_p\}$, the variables $X_1, \dots, X_p$ can be quantitative, ordinal, categorical, or a mix of all of the above.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

- (Optional) Response: Y, which can be one variable or a vector of many – a special variable of interest

Supervised vs unsupervised learning

Unsupervised learning: only X is observed

- Goal: understand/summarize/visualize the relationships between the variables in X
- Examples: principal components analysis, clustering

Supervised learning: X and Y are observed

- Goal: understand/summarize/visualize the relationships between X and Y, and/or learn to predict future/ unknown values of Y from X

Examples

- Visualization: applicable to both supervised and unsupervised tasks, often with different plots
- Classification (supervised; Y is a categorical variable)
- Regression (supervised; Y is continuous variable)
- Clustering (unsupervised)
- ANOVA (supervised; categorical X, continuous Y)

Classification: Definition

- Given a collection of data points (training set) $(x_1, y_1), \ldots, (x_n, y_n)$, where y is a class label (categorical)
- Find a model or an algorithm that outputs the class label y as a function of the values of variables x
- Goal: previously unseen data points x should be assigned a class label y as accurately as possible
- A test set (previously unseen) is used to determine the accuracy of the model

Classification Example: Customer Scoring

- A bank has a database of 1m past customers, 10% of whom took out mortgages with the bank
- Use data mining to predict whether a customer will take out a mortgage or not, based on the customer's data
- Customer data
  - History of transactions with the bank
  - Other credit data
  - Demographic data

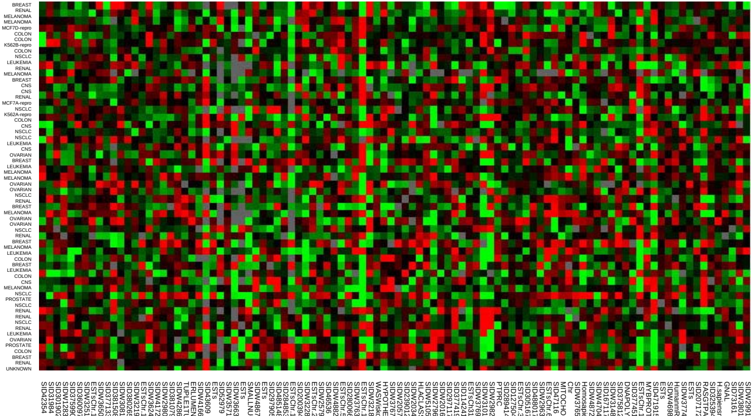
Classification Example: Spam Detection

- Customize an email spam detection system for an individual user.
- Relative frequencies in a message of most commonly occurring words and punctuation marks.

	george	you	your	hp	free	re	remove
spam	0.00	2.26	1.38	0.02	0.52	0.13	0.28
email	1.27	1.27	0.44	0.90	0.07	0.42	0.01

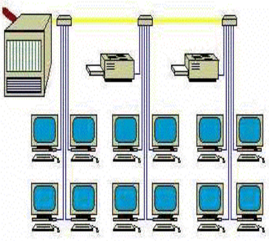
Classification Example: Microarray gene expression data

Predict developing cancer based on genetic profile



Classification Example: Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
    - Credit card fraud
    - Network intrusion



Credit Card Fraud Detection

- Credit card losses in the US are over 1 billion $ per year
- Roughly 1 in 50k transactions are fraudulent
- Fair-Issac's fraud detection software based on neural networks, led to reported fraud decreases of 30-50%
- Challenge: false alarm rate vs missed detection

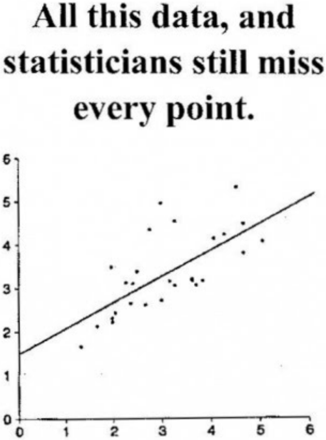
Regression: Definition

- Predict a value of a given continuous-valued variable based on the values of other variables
- Linear regression: predict Y from a linear combination of X
- Many other models are possible (nonlinear functions, trees, neural networks, etc)

Regression: Examples

- Predicting sales amounts of new product based on advertising expenditure
- Predicting wind velocity as a function of temperature, humidity, air pressure, etc
- Predict a student's freshman year GPA based on high school grades and SATs

Occam's razor: bias vs variance trade-off



All this data, and statisticians still miss every point.

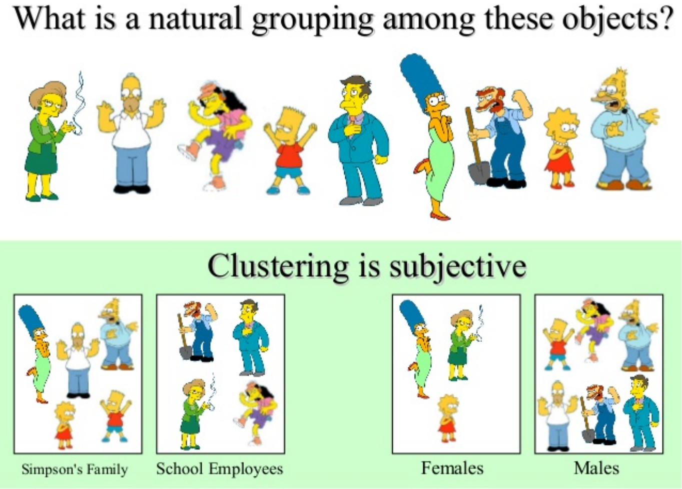
Prediction vs inference

- Prediction: the goal is to predict Y from X. The predictor could be a black box as long as it's accurate.
- Inference: the goal is to understand how

Clustering: Definition

- Given a set of data points, each having a set of variables, find clusters such that
    - data points in the same cluster are " more similar" to one another, and
    - data points in different clusters are "less similar" to one another.
- Similarity measures
    - Euclidean distance if variables are continuous
    - Other problem-specific measures

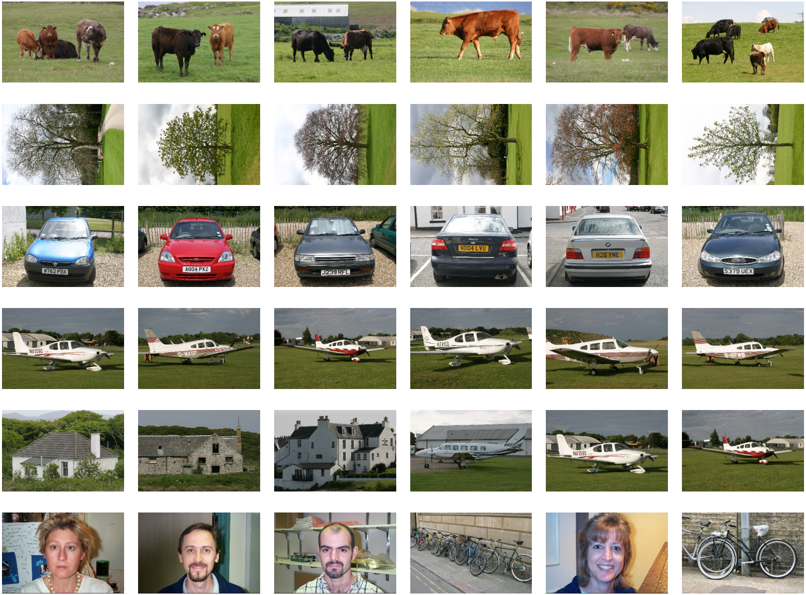
Importance of similarity measures



Clustering Example: Market Segmentation

- Goal: subdivide a market into distinct subsets of customers for targeted advertisement
- Collect different variables on customers (demographics like age, gender, marital status, education; geographical locations; lifestyle, hobbies, etc)
- Find clusters of similar customers

Clustering Example: Images



More Clustering Examples

- Cluster documents that are similar to each other based on the important terms appearing in them, for example to organize news articles by topics
- Cluster patients by symptoms and medical history, to develop personalized treatments
- Cluster stocks based on their movements every day, to find patterns in the market

The challenges in learning from data

- Hype: people often expect more than is realistic
- Data snooping and fishing: finding spurious structure that is not reproducible
- Trade-offs:
  - Prediction vs inference: may get great performance from a black box, and a more interpretable simpler model may not predict as well
  - False alarm vs missed detection (Type 1 vs Type 2 error)
  - Bias vs variance: flexibility vs overfitting

Ode to Statistics (machine learning, data mining, etc)

Google's chief economist Hal Varian, 2009:

I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would have guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades, not only at the professional level but even at educational levels... Because now we really do have essentially free and ubiquitous data. So the complementary scarce factor is the ability to understand that data and extract value from it.