94-775/95-865 Lecture 3: Finding Possibly Related Entities, Visualizing High-Dimensional Vectors

George Chen
Last Time: Co-Occurrences

- Joint probability $P(A, B)$ can be a poor indicator of whether $A$ and $B$ co-occurring is “interesting”

- Find interesting relationships between pairs of items by looking at PMI

  - Intuition: “Interesting” co-occurring events should occur more frequently than if they were to co-occur independently

- Find interesting relationship between types of items (and *not* specific pairs of items) using chi-square (or equivalently phi-square)
Co-occurrence Analysis Applications

- If you're an online store/retailer: anticipate *when* certain products are likely to be purchased/rented/consumed more
  - Products & dates
- If you have a bunch of physical stores: anticipate *where* certain products are likely to be purchased/rented/consumed more
  - Products & locations
- If you're the police department: create "heat map" of where different criminal activity occurs
  - Crime reports & locations
Co-occurrence Analysis Applications

- If you're an online store/retailer:
  anticipate when certain products are likely to be purchased/rented/consumed more

- Products & locations
- Products & dates

- If you have a bunch of physical stores:
  anticipate where certain products are likely to be purchased/rented/consumed more

- If you're the police department:
  create "heat map" of where different criminal activity occurs
  Crime reports & locations

Examples of data to take advantage of:

- data collected by your organization
- social networks
- news websites
- blogs

Web scraping frameworks can be helpful:

- Scrapy
- Selenium (great with JavaScript-heavy pages)
Continuous Measurements

- So far, looked at relationships between discrete outcomes
- For pair of continuous outcomes, use a scatter plot

Image source: https://plot.ly/~MattSundquist/5405.png

Of course, not all trends look like a line
(so don’t just do linear regression!)
The Importance of Staring at Data

In general: not obvious what curve to fit (if any)

Not enough data => might think there's a pattern when it's just noise

In general: not obvious if some points are outliers and should be excluded
Correlation

Beware: Just because two variables appear correlated doesn't mean that one can predict the other
Correlation ≠ Causation

Moreover, just because we find correlation in data doesn't mean it has predictive value!

Important: At this point in the course, we are finding possible relationships between two entities

We are not yet making statements about prediction (we'll see prediction later in the course)

We are not making statements about causality (beyond the scope of this course)
Causality

Studies in 1960's: Coffee drinkers have higher rates of lung cancer. *Can we claim that coffee is a cause of lung cancer?*

Back then: coffee drinkers also tended to smoke more than non-coffee drinkers (smoking is a **confounding variable**).

To establish causality, groups getting different treatments need to appear similar so that the only difference is the treatment.

Image source: George Chen
Establishing Causality

If you control data collection

Users

- Treatment Group
- Control Group

Randomly assign

Compare outcomes of two groups

Randomized controlled trial (RCT) also called A/B testing

Example: figure out webpage layout to maximize revenue (Amazon)
Example: figure out how to present educational material to improve learning (Khan Academy)

If you do not control data collection

In general: not obvious establishing what caused what
Part I: Exploratory data analysis

*Identify structure present in “unstructured” data*

- Frequency and co-occurrence analysis
- Visualizing high-dimensional data/dimensionality reduction
- Clustering
- Topic modeling (a special kind of clustering)

Part II: Predictive data analysis

*Make predictions using structure found in Part I*

- Classical classification methods
- Neural nets and deep learning for analyzing images and text
Visualizing High-Dimensional Vectors

The next two examples are drawn from: http://setosa.io/ev/principal-component-analysis/
Imagine we had hundreds of these. How to visualize these for comparison?

Using our earlier analysis:

- Compare pairs of food items across locations (e.g., scatter plot of cheese vs cereals consumption)

But unclear how to compare the locations (England, Wales, Scotland, N. Ireland)!
The issue is that as humans we can only really visualize up to 3 dimensions easily.

Goal: Somehow reduce the dimensionality of the data preferably to 1, 2, or 3.
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Simplest thing to try: flatten to one of the red axes
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Simplest thing to try: flatten to one of the red axes
(We could of course flatten to the other red axis)
Principal Component Analysis (PCA)

How to project 2D data down to 1D?
Principal Component Analysis (PCA)

How to project 2D data down to 1D?
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

But notice that most of the variability in the data is not aligned with the red axes!

Most variability is along this green direction

Rotate!
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Most variability is along this green direction.
Principal Component Analysis (PCA)

How to project 2D data down to 1D?

Most variability is along this green direction.

The idea of PCA actually works for 2D → 2D as well (and just involves rotating, and not “flattening” the data).
Principal Component Analysis (PCA)

How to project 2D data down to 1D?
How to rotate 2D data so 1st axis has most variance

The idea of PCA actually works for 2D → 2D as well (and just involves rotating, and not “flattening” the data)

2nd green axis chosen to be 90° (“orthogonal”) from first green axis
Principal Component Analysis (PCA)

- Finds top $k$ orthogonal directions that explain the most variance in the data
  - 1st component: explains most variance along 1 dimension
  - 2nd component: explains most of remaining variance along next dimension that is orthogonal to 1st dimension
  - ...

- “Flatten” data to the top $k$ dimensions to get lower dimensional representation (if $k < \text{original dimension}$)
Principal Component Analysis (PCA)

3D example from:
http://setosa.io/ev/principal-component-analysis/
Principal Component Analysis (PCA)

Demo
PCA reorients data so axes explain variance in “decreasing order” → can “flatten” (project) data onto a few axes that captures most variance
PCA would just flatten this thing and lose the information that the data actually lives on a 1D line that has been curved!
PCA would squash down this Swiss roll (like stepping on it from the top) mixing the red & white parts.
2D Swiss Roll
2D Swiss Roll
2D Swiss Roll
2D Swiss Roll
2D Swiss Roll
2D Swiss Roll

This is the desired result
3D Swiss Roll

Projecting down to any 2D plane puts points that are far apart close together!
3D Swiss Roll

Projecting down to any 2D plane puts points that are far apart close together!

Goal: Low-dimensional representation where similar colored points are near each other (we don’t actually get to see the colors)
Manifold Learning

- Nonlinear dimensionality reduction (in contrast to PCA which is linear)
- Find low-dimensional “manifold” that the data live on

Basic idea of a manifold:

1. Zoom in on any point (say, $x$)
2. The points near $x$ look like they’re in a lower-dimensional Euclidean space (e.g., a 2D plane in Swiss roll)
Do Data Actually Live on Manifolds?

Do Data Actually Live on Manifolds?

Do Data Actually Live on Manifolds?

Do Data Actually Live on Manifolds?