Previously on 95-865…
What is PMI Measuring?

Probability of A and B co-occurring if they were independent

$$\frac{P(A, B)}{P(A) P(B)}$$

PMI measures (the log of) a ratio that says how far A and B are from being independent

There are lots of connections of information theory to prediction

Rough intuition:
Something surprising ↔ less predictable ↔ more bits to store
Looking at All Pairs of Outcomes

- PMI measures how $P(A, B)$ differs from $P(A)P(B)$ using a log ratio
- Log ratio isn’t the only way to compare!
- Another way to compare:

$$\Phi^2 = \sum_{A, B} \frac{[P(A, B) - P(A)P(B)]^2}{P(A)P(B)}$$

Phi-square is between 0 and 1

0 $\rightarrow$ pairs are all indep.

Chi-square = $N \times \Phi^2$

$N$ = sum of all co-occurrence counts (in upper right of triangle earlier)

Measures how close all pairs of outcomes are close to being indep.
### Example: Phi-Square Calculation

<table>
<thead>
<tr>
<th></th>
<th>Green</th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>1000</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>White</td>
<td>2000</td>
<td>2000</td>
<td>350</td>
</tr>
<tr>
<td>Black</td>
<td>200</td>
<td>2000</td>
<td>2000</td>
</tr>
</tbody>
</table>

\( N = 5750 \)

Sum comprises of 6 terms

\[
\text{Phi-square} = \sum_{A, B} \left[ \frac{P(A, B) - P(A) P(B)}{P(A) P(B)} \right]^2
\]

**Interpretation:** neighboring pixels not close to being indep.

\[
P(Green, White) = \frac{200}{5750}
\]

\[
P(Green, Black) = \frac{200}{5750}
\]

\[
P(White, Black) = \frac{350}{5750}
\]

\[
P(Green) = \frac{1400}{5750}
\]

\[
P(White) = \frac{2550}{5750}
\]

\[
P(Black) = \frac{2550}{5750}
\]

**Add these up to get:**

\[
\text{Phi-square} = 0.6470...
\]
Often we know what kind of named entities are found
Example: Elon Musk and Tim Cook are people,
Tesla and Apple are companies
→ can ask what people are related to what companies
Back to Earlier Example

<table>
<thead>
<tr>
<th></th>
<th>Tesla</th>
<th>Apple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elon Musk</td>
<td>300</td>
<td>1</td>
</tr>
<tr>
<td>Tim Cook</td>
<td>1</td>
<td>195</td>
</tr>
</tbody>
</table>

PMI, phi-square, chi-square calculations done same way as before

Main things to calculate first:

- $P(\text{Elon Musk}, \text{Tesla})$
- $P(\text{Elon Musk}, \text{Apple})$
- $P(\text{Tim Cook}, \text{Tesla})$
- $P(\text{Tim Cook}, \text{Apple})$
- $P(\text{Elon Musk})$
- $P(\text{Tim Cook})$
- $P(\text{Tesla})$
- $P(\text{Apple})$

The math here is actually a bit easier to think about because the rows and columns aren't indexing the same items
Back to Earlier Example

These are the joint probabilities!

$$\begin{array}{c|c|c}
 & \text{Tesla} & \text{Apple} \\
\hline
\text{Elon Musk} & 300 & 1 \\
\hline
\text{Tim Cook} & 1 & 195 \\
\hline
\end{array}$$

Total: 497

Divide by total

$$\begin{array}{c|c|c}
 & \text{Tesla} & \text{Apple} \\
\hline
\text{Elon Musk} & \frac{300}{497} & \frac{1}{497} \\
\hline
\text{Tim Cook} & \frac{1}{497} & \frac{195}{497} \\
\hline
\end{array}$$

Compute “marginals”

$$\begin{array}{c|c|c}
 & \text{Tesla} & \text{Apple} \\
\hline
\text{Elon Musk} & \frac{300}{497}+\frac{1}{497} & \frac{1}{497} \\
\hline
\text{Tim Cook} & \frac{1}{497}+\frac{195}{497} & \frac{195}{497} \\
\hline
\end{array}$$

$$P(\text{Elon Musk}) = \frac{300}{497}+\frac{1}{497}$$

$$P(\text{Tim Cook}) = \frac{1}{497}+\frac{195}{497}$$

$$P(\text{Tesla}) = \frac{300}{497}+\frac{1}{497}$$

$$P(\text{Apple}) = \frac{1}{497}+\frac{195}{497}$$

Not just for 2 by 2 tables (e.g., we could have many people, many companies)
Summary: 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

- In practice: some times it is helpful to generalize PMI and look instead at

\[
\text{PMI}_\rho(A, B) = \log_2 \frac{P(A, B)^\rho}{P(A) P(B)}
\]

Tune parameter $\rho > 0$

(we'll talk about parameter tuning later in the course)
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

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)

- Crime reports & locations
Example Application of PMI: Image Segmentation

Example Application of PMI:
Word Embeddings

Image source: https://deeplearning4j.org/img/countries_capitals.png

Omer Levy and Yoav Goldberg. Neural word embeddings as implicit matrix factorization. NIPS 2014.
Continuous Measurements

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

Of course, not all trends look like a line

![Graph showing computing improvements: Transistors Per Circuit](https://plot.ly/~MattSundquist/5405.png)
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

Negatively correlated

Not really correlated

Positively correlated

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
95-865 Outline

Part I: Exploratory data analysis

Identify structure present in “unstructured” data

• Frequency and co-occurrence analysis

• Clustering

• Topic modeling (a special kind of clustering)

Basic probability & statistics

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

George Chen

The next two examples are drawn from:
http://setosa.io/ev/principal-component-analysis/
Visualizing High-Dimensional Vectors

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 $k <$ original dimension)
Principal Component Analysis (PCA)

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

Demo