Stats 170A: Project in Data Science

Data Visualization and Exploratory Data Analysis

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine
Overview

• Lectures/Homeworks up to this point
  – Data management (relational DBs, query languages, PostgreSQL)
  – Data manipulation in Python (Pandas)
  – Data formats (JSON, XML)
  – Practical experience with Twitter data, IMDB data

• Next 2 weeks
  – Review of data visualization and exploration
  – Basic principles of machine learning (and some statistics)
  – Machine learning with text data
How this Course will work

• Q1: Weeks 1 to 6: Lectures and Assignments
  – Review general principles of data science
  – Weeks 1 to 3: databases, data extraction, data cleaning
  – Weeks 4 to 6: text analysis, data exploration, machine learning
  – Combination of lectures, assignments, and background reading

• Q1: Weeks 7 to 10: Project Proposals
  – Project proposals from student teams
  – Feedback from instructors, refine proposal, oral presentation at end of quarter

• Q2: Work on Projects
  – Build and use a prototype system/pipeline
  – Develop ideas, implement algorithms, make use of libraries and packages
  – Conduct experiments with real data sets
  – Test and evaluate your system in a systematic manner
  – Communicate your results (presentations and reports)
Assignment 5

Refer to the Wiki page

Due noon on Monday February 12th to EEE dropbox

Note change: due before class (by 2pm)
seaborn: statistical data visualization

Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

For a brief introduction to the ideas behind the package, you can read the introductory notes. More practical information is on the installation page. You may also want to browse the example gallery to get a sense for what you can do with seaborn and then check out the tutorial and API reference to find out how.

To see the code or report a bug, please visit the github repository. General support issues are most at home on stackoverflow, where there is a seaborn tag.

Documentation

- An introduction to seaborn
- What's new in the package
- Installing and getting started
- Example gallery
- Seaborn tutorial
- API reference

Features

- Style functions: API | Tutorial
- Color palettes: API | Tutorial
- Distribution plots: API | Tutorial
- Categorical plots: API | Tutorial
- Regression plots: API | Tutorial
- Axis grid objects: API | Tutorial

© Copyright 2012-2017, Michael Waskom.
Created using Sphinx 1.3.3.
Types of Data
Types of Data for a Single Variable

• Real-valued, continuous
  – e.g., a person’s weight or income
  – values may be discretized and bounded, but we will think of as on the real line

• Integer
  – e.g. Year of birth, number of years in college
  – Could a be a real-valued variable that is quantized (age in years)

• Ordinal
  – e.g., education level = \{kindergarten, high school, college, grad school,\ldots\}

• Categorical
  – e.g., \{red, blue, yellow\} or \{CA, MA, NY, AZ, \ldots\} or text strings

(Note that many visualization and machine learning techniques implicitly assume real-valued data, and other data types are converted to reals or rep)
Multiple Variables

- More than 1 variable, often referred to as multivariate or multidimensional

- Often interested in relationships between variables and geometric structure of the data (for real-valued data), e.g., is it clustered?

- Small numbers of variables can plot the data and look at relationships

- For large numbers we use exploratory techniques
  - E.g., clustering and dimension reduction

- Note that many visualization and machine learning techniques implicitly assume real-valued data
  ....categorical data types are often converted to reals (e.g., binary) or represented via grouping, colors, or icons
Data with Context

• Time-series data
  – A variable whose values are indexed by time
  – We can also have multidimensional time-series

• Sequence data
  – A variable indexed by position
  – E.g., words (categorical) in text, or DNA sequences

• Spatial data
  – Data whose values are indexed spatially, e.g., by lat/lon or by city
  – Can also have multidimensional time-series

• Spatio-temporal
  – Indexed by both space and time, e.g., storm tracks, vehicle trajectories, etc
Stock Market Indices last week

- **S. & P. 500 Index**
- **Dow Jones Industrial Average**

Source: Reuters

At market close 02/02/2018
Night Lights from North and South Korea

From https://www.vox.com
Where People Run

From: https://flowingdata.com/2014/02/05/where-people-run/#jp-carousel-33695
Relational Data

- N entities, \( i = 1, \ldots, N \)
- N x N relations:
  - can be represented as an array \( y(i, j) = 1 \) if \( i \) is connected to \( j \), 0 otherwise
  - Example: a social network

- Can combine with other data, e.g.,
  - Each relation could have metadata, e.g., text
  - Each relation could be time-dependent, \( y(i, j, t) \) is a time series over time \( t \)
Visualization of an email network using 2-dimensional graph drawing or “embedding”

Data from 500 researchers at Hewlett-Packard over approximately 1 year.

Various structural elements of the network are apparent
Philosophy behind this Class

• Provide an experience of how data science works in the real-world
  – Defining a problem
  – Identifying, understanding, exploring relevant data
  – Extracting, cleaning, management of data
  – Exploration and analysis of data
  – Building models from data (e.g., via machine learning)
  – Evaluating models: how well do they predict
  – Communicating your results to others

• Tie together ideas from different courses you have taken and give you experience in applying these ideas to real-world data
  – Databases, software, algorithms, machine learning, statistics
Data Science: from Data to Actions

- **Raw Data**
- **Data Wrangling**
- **Data Management**
- **Databases, Algorithms, Software Engineering**
- **Exploratory Data Analysis**
- **Predictive Modeling**
- **Machine Learning, Statistics**
- **Consumers**
- **External Business Customers**
- **Internal Business Customers**
- **Government**
- **Scientists**

**Domain knowledge**

**Business knowledge**
Why Visualization and Exploration?

- People are good at pattern recognition
  - At spotting clusters, trends, outliers, structure...that computers many miss

- Usually two types of users
  1. The data scientist who wants to explore/analyze/understand
     - For the data scientist, visualization and exploration are part of an iterative process
  2. The person who needs a quick summary to make a decision
     - For the consumer we want to communicate information quickly and clearly
     - e.g., for a medical doctor, for a policy-maker, for a consumer

- For data scientists...its always a good idea to look at your data
  - Helps to understand where the semantics of the data...what the measurements actually mean
What is Exploratory Data Analysis?

- Broader than just visualization

- EDA = \{visualization, clustering, dimension reduction, \ldots\}

- For small numbers of variables, EDA = visualization

- For large numbers of variables, we need to be cleverer
  - Clustering, dimension reduction, embedding algorithms
  - These are techniques that essentially reduce high-dimensional data to something we can look at

- Pioneered by John Tukey (statistician at Bell Labs, Princeton) in the 1960’s
  - “let the data speak”
Exploratory Data Analysis: Single Variables
Summary Statistics

Mean: “center of data”
Mode: location of highest data density
Variance: “spread of data”
Skew: indication of non-symmetry

Range: max - min
Median: 50% of values below, 50% above
Quantiles: e.g., values such that 25%, 50%, 75% are smaller

Note that some of these statistics can be misleading
E.g., mean for data with 2 clusters may be in a region with zero data
Histogram of Unimodal Data

1000 data points simulated from a Normal distribution, mean 10, variance 1, 30 bins
**Histograms: Unimodal Data**

100 data points from a Normal, mean 10, variance 1, with 5, 10, 30 bins
Histogram of Multimodal Data

15000 data points simulated from a mixture of 3 Normal distributions, 300 bins
Histogram of Multimodal Data

15000 data points simulated from a mixture of 3 Normal distributions, 300 bins
Skewed Data

5000 data points simulated from an exponential distribution, 100 bins
Another Skewed Data Set

10000 data points simulated from a mixture of 2 exponentials, 100 bins
Same Skewed Data after taking Logs (base 10)

10000 data points simulated from a mixture of 2 exponentials, 100 bins
What will the mean or median tell us about this data?
Histogram with Outliers

Pima Indians Diabetes Data,
From UC Irvine Machine Learning Repository

Number of Individuals

X values
Histogram with Outliers

Pima Indians Diabetes Data, From UC Irvine Machine Learning Repository

Number of Individuals

Diastolic Blood Pressure

blood pressure = 0?
Box Plots: Diabetes Data

Two side-by-side box-plots of individuals from the Pima Indians Diabetes Data Set

Body Mass Index

Note: significant overplotting here that could easily be missed
Box Plots: Diabetes Data

Two side-by-side box-plots of individuals from the Diabetes Data Set

- **Body Mass Index**

  - **Healthy Individuals**
    - **Q2** (median)
    - **Q1**
    - **Box = middle 50% of data**
    - **Upper Whisker**
    - **Lower Whisker**

  - **Diabetic Individuals**
    - **Q3**
    - **Q3 - Q1**
    - **Plots all data points outside “whiskers”**
    - **1.5 x**

- **CLASS**

  - **Healthy Individuals**
  - **Diabetic Individuals**
Multiple Box Plots: Diabetes Data

- **Diastolic Blood Pressure**
- **24-hour Serum Insulin**
- **Plasma Glucose Concentration**
- **Body Mass Index**

Healthy vs. Diabetic comparisons.
Horizontal BoxPlot for Planet Data

From: https://seaborn.pydata.org/examples/horizontal_boxplot.html
Exploring Pairs of Variables
Relationships between Pairs of Variables

• Say we have a variable Y we want to predict and many variables X that we could use to predict Y

• In exploratory data analysis we may be interested in quickly finding out if a particular X variable is potentially useful at predicting Y

• Options?
  – Linear correlation

  – Scatter plot: plot Y values versus X values
Linear Dependence between Pairs of Variables

• Covariance and correlation measure linear dependence

• Assume we have two variables or attributes $X$ and $Y$ and $n$ objects taking values $x(1), \ldots, x(n)$ and $y(1), \ldots, y(n)$. The sample covariance of $X$ and $Y$ is:

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x(i) - \bar{x})(y(i) - \bar{y})$$

• The covariance is a measure of how $X$ and $Y$ vary together.
  – large and positive if large values of $X$ are associated with large values of $Y$ and small $X \Rightarrow$ small $Y$

• (Pearson Linear) Correlation = scaled covariance, varies between -1 and 1

$$\rho(X, Y) = \frac{\sum_{i=1}^{n} (x(i) - \bar{x})(y(i) - \bar{y})}{\left( \sum_{i=1}^{n} (x(i) - \bar{x})^2 \sum_{i=1}^{n} (y(i) - \bar{y})^2 \right)^{1/2}}$$
# Data Set on Housing Prices in Boston

(widely used data set in research in regression (prediction) research)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CRIM</td>
<td>per capita crime rate by town</td>
</tr>
<tr>
<td>2</td>
<td>ZN</td>
<td>proportion of residential land zoned for lots over 25,000 ft²</td>
</tr>
<tr>
<td>3</td>
<td>INDUS</td>
<td>proportion of non-retail business acres per town</td>
</tr>
<tr>
<td>4</td>
<td>NOX</td>
<td>Nitrogen oxide concentration (parts per 10 million)</td>
</tr>
<tr>
<td>5</td>
<td>RM</td>
<td>average number of rooms per dwelling</td>
</tr>
<tr>
<td>6</td>
<td>AGE</td>
<td>proportion of owner-occupied units built prior to 1940</td>
</tr>
<tr>
<td>7</td>
<td>DIS</td>
<td>weighted distances to five Boston employment centres</td>
</tr>
<tr>
<td>8</td>
<td>RAD</td>
<td>index of accessibility to radial highways</td>
</tr>
<tr>
<td>9</td>
<td>TAX</td>
<td>full-value property-tax rate per $10,000</td>
</tr>
<tr>
<td>10</td>
<td>PTRATIO</td>
<td>pupil-teacher ratio by town</td>
</tr>
<tr>
<td>11</td>
<td>MEDV</td>
<td>Median value of owner-occupied homes in $1000's</td>
</tr>
</tbody>
</table>
Matrix of Pairwise Linear Correlations

Data on characteristics of Boston housing

- Crime Rate
- Industry
- Nitrous oxide
- Average # rooms
- Proportion of old houses
- Highway accessibility
- Property tax rate
- Student-teacher ratio

- Percentage of large residential lots
- Distance to employment centers
- Median house value
Examples of X-Y plots and linear correlation values
Examples of X-Y plots and linear correlation values
Lack of linear correlation does not imply lack of dependence
Summary Statistics for Anscombe’s 4 Data Sets


4 data sets, each with 2 variables X and Y, with the same summary statistics (imagine that Python reports these summaries and we have not yet looked at the data)

**Summary Statistics of Data Set 1**

\[ N = 11 \]
Mean of X = 9.0
Mean of Y = 7.5

**Summary Statistics of Data Set 2**

\[ N = 11 \]
Mean of X = 9.0
Mean of Y = 7.5

**Summary Statistics of Data Set 3**

\[ N = 11 \]
Mean of X = 9.0
Mean of Y = 7.5

**Summary Statistics of Data Set 4**

\[ N = 11 \]
Mean of X = 9.0
Mean of Y = 7.5
Anscombe’s 4 Data Sets

Guess the Linear Correlation Values for each Data Set

### Summary Statistics for each Data Set

#### Summary Statistics of Data Set 1
- $N = 11$
- Mean of $X = 9.0$
- Mean of $Y = 7.5$
- Intercept = 3
- Slope = 0.5
- Correlation = 0.82

#### Summary Statistics of Data Set 2
- $N = 11$
- Mean of $X = 9.0$
- Mean of $Y = 7.5$
- Intercept = 3
- Slope = 0.5
- Correlation = 0.82

#### Summary Statistics of Data Set 3
- $N = 11$
- Mean of $X = 9.0$
- Mean of $Y = 7.5$
- Intercept = 3
- Slope = 0.5
- Correlation = 0.82

#### Summary Statistics of Data Set 4
- $N = 11$
- Mean of $X = 9.0$
- Mean of $Y = 7.5$
- Intercept = 3
- Slope = 0.5
- Correlation = 0.82

**Lesson:** Summary statistics can be misleading.

---

Dangers of searching for correlations in high-dimensional data

Simulated 50 random Gaussian/normal data vectors, each with 100 variables
Results in a 50 x 100 data matrix

Below is a histogram of the 100 choose 2 pairs of correlation coefficients

Even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.
Correlations in a Large Random Data Set

From: https://seaborn.pydata.org/examples/many_pairwise_correlations.html
Conclusions so far?

• Summary statistics are useful.....up to a point

• Linear correlation measures can be misleading

• There really is no substitute for plotting/visualizing the data
Scatter Plots

- Plot the value of one variable against the other

- Simple...but can be very informative, can reveal more than summary statistics

- For example, we can...
  - See if variables are dependent on each other (beyond linear dependence)
  - Detect if outliers are present
  - Can color-code to overlay group information (e.g., color points by class label for classification problems)
MEDIAN HOUSEHOLD INCOME

MEDIAN PER CAPITA INCOME

(units = dollars)

(from US Zip code data: each point = 1 Zip code)
Constant Variance versus Changing Variance

variation in $Y$ does not depend on $X$

variation in $Y$ changes with the value of $X$

e.g., $Y = \text{annual tax paid}$, $X = \text{income}$
Scatter-Plot Matrices: Example for Diabetes Data
Using Color to Show Group Information in Scatter Plots

Iris classification data set, 3 classes

Figure from www.originlab.com
Another Example with Grouping by Color

Figure from hci.stanford.edu
Outlier Detection

- Definition of an outlier?
  - No precise definition
  - Generally....”A data point that is significantly different to the rest of the data”
  - But how do we define “significantly different”? (many answers to this.....)
  - Typically assumed to mean that the point was measured in error, or is not a true measurement in some sense

Outliers in 1 dimension

Outlier in 2 dimensions

Diastolic Blood Pressure (mm Hg)

X VALUES

Y VALUES
Assignment 5

Refer to the Wiki page

Due noon on Monday February 12\textsuperscript{th} to EEE dropbox

Note change: due before class (by 2pm)