TCSS 588
Bioinformatics

Ka Yee Yeung
kayee@uw.edu
Institute of Technology, UW-Tacoma
3/28/2018

1

Overview

- Jupyter notebooks
- Why do we care about statistics? What can we do with predictive models?
- Review on statistics
- Regression
- Building predictive models

2

Life cycle of an academic project

1. Individual exploratory work
2. Collaborative development
3. Parallel (scale up using cloud computing)
4. Publication and communication (reproducibly!)

https://www.slideshare.net/mbussonn/jupyter-a-platform-for-data-science-at-scale

Data science notebooks

- Mid 1980's: Start of computational notebooks: Matlab, Mathematica notebooks, Maple worksheets
  - GUI allowed for the interactive creation and editing of notebook documents that contain pretty-printed program code, formatted text
- 2010-2011: IPython web-notebooks and prototype
- 2014: Project Jupyter is a spinoff project from IPython

https://www.datacamp.com/community/blog/ipython-jupyter#gs.wngbPCE

jupyter

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text.

The name "Jupyter" is actually short for "Julia, Python and R"

".ipynb" files: JSON (JavaScript Object Notation) based files embedding input and output

http://jupyter.org/

Kernels

- A kernel is a program that runs and introspects the user's code: it provides computation and communication with the frontend interfaces, such as notebooks.
- The Jupyter Notebook Application has three main kernels: the IPython, IRkernel and IJulia kernels.
- Community maintained kernels: Ruby, Javascript, Scala, Perl, Octave etc.
- https://github.com/jupyter/jupyter/wiki/Jupyter-kernels

Go to your notebook now

7

Overview

- Jupyter notebooks
- Why do we care about statistics? What can we do with predictive models?
- Review on statistics
- Regression
- Building predictive models

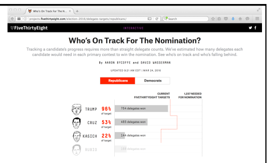
8

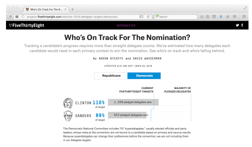
4 / 29

FiveThirtyEight's NFL Elo Playoff Odds
Dec. 23, 2014

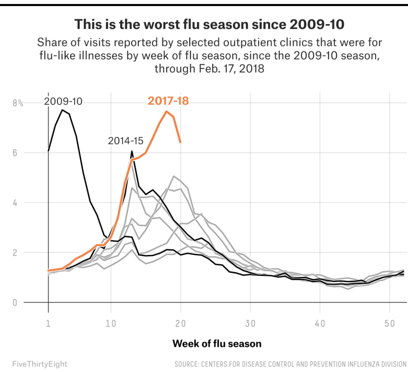
TEAM	ELO RATING	WINS	LOSSES	TIES	POINT DIFF.	WIN DIVISION	MAKE PLAYOFFS	WIN SUPER BOWL
AFC East								
New England	1730	12.8	3.2	0.0	+175	100%	100%	26%
Miami	1514	8.8	7.2	0.0	+38	0%	0%	0%
Buffalo	1501	8.2	7.8	0.0	+34	0%	0%	0%
N.Y. Jets	1357	3.2	12.8	0.0	-141	0%	0%	0%
AFC North								
Pittsburgh	1605	10.6	5.4	0.0	+62	61%	100%	3%
Cincinnati	1594	10.4	4.6	1.0	+27	39%	100%	3%
Baltimore	1574	9.8	6.2	0.0	+108	0%	46%	<1%
Cleveland	1384	7.2	8.8	0.0	-39	0%	0%	0%
AFC South								
Indianapolis	1568	10.8	5.2	0.0	+82	100%	100%	2%
Houston	1494	8.8	7.2	0.0	+71	0%	9%	<1%
Jacksonville	1283	3.2	12.8	0.0	-169	0%	0%	0%
Tennessee	1262	2.2	13.8	0.0	-177	0%	0%	0%
AFC West								
Denver	1667	11.9	4.1	0.0	+110	100%	100%	12%
San Diego	1548	9.4	6.6	0.0	+10	0%	43%	<1%
Kansas City	1530	8.6	7.4	0.0	+62	0%	2%	<1%
Oakland	1331	3.1	12.9	0.0	-181	0%	0%	0%
NFC East								
Dallas	1620	11.8	4.2	0.0	+99	100%	100%	5%
Philadelphia	1536	9.5	6.5	0.0	+65	0%	0%	0%
N.Y. Giants	1485	6.5	9.5	0.0	-11	0%	0%	0%
Washington	1311	4.2	11.8	0.0	-121	0%	0%	0%
NFC North								
Green Bay	1623	11.7	4.3	0.0	+134	67%	100%	7%
Detroit	1575	11.3	4.7	0.0	+43	33%	100%	3%
Minnesota	1451	6.7	9.3	0.0	-16	0%	0%	0%
Chicago	1381	5.3	10.7	0.0	-125	0%	0%	0%
NFC South								
Carolina	1489	6.4	8.6	1.0	-68	43%	43%	<1%
Atlanta	1476	6.6	9.4	0.0	-3	57%	57%	<1%
New Orleans	1462	6.6	9.4	0.0	-22	0%	0%	0%
Tampa Bay	1318	2.4	13.6	0.0	-134	0%	0%	0%
NFC West								
Seattle	1755	11.9	4.1	0.0	+140	94%	100%	32%
Arizona	1591	11.5	4.5	0.0	+14	6%	100%	3%
San Francisco	1522	7.5	8.5	0.0	-37	0%	0%	0%
St. Louis	1463	6.1	9.9	0.0	-30	0%	0%	0%

FIVETHIRTYEIGHT





Source: http://fivethirtyeight.com/ 10



Source: https://fivethirtyeight.com/features/america-should-have-stayed-home-this-flu-season/ 11

Motivation

Use data to identify relationships among variables and use these relationships to make predictions.

12

Looking at your data

- How are the data distributed?
  - Where is the center?
  - What is the range?
  - What is the shape of the distribution (e.g., Gaussian, uniform)?
- Are there “outliers”?
- Are there data points that don’t make sense?

13

Types of Variables: Overview

Categorical

Quantitative

binary

nominal

ordinal

discrete

continuous

2 categories

more categories +

order matters +

numerical +

uninterrupted

14

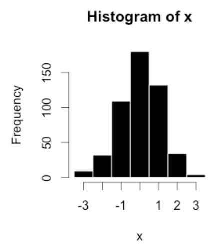
Continuous variables

- Histograms
- Box plots

15

Histogram

bin	count
-3.5	9
-2.5	32
-1.5	109
-0.5	180
0.5	132
1.5	34
2.5	4
3.5	9



Toy example from wikipedia: http://en.wikipedia.org/wiki/Histogram

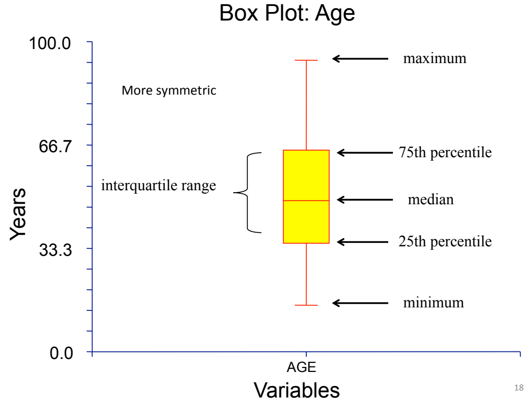
16

In-class exercise

- Predicting adverse drug reactions http://www.maayanlab.net/crowdsourcing/megatask1.php
- We will re-visit this task over this course.
- Basic data types in R: http://www.statmethods.net/input/datatypes.html

17

Box Plot: Age



18

Boxplots in R

- http://www.r-bloggers.com/box-plot-with-r-tutorial/
- https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/boxplot.html (advanced, more on this later)

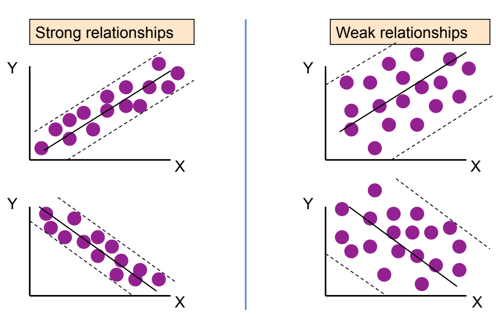
19

Correlation

- Measures the relative strength of the *linear* relationship between two variables
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

20

Linear Correlation

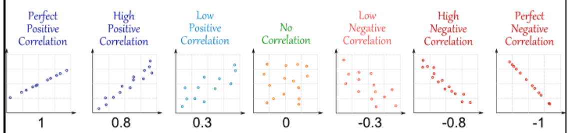


▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

21

Pearson's Correlation Coefficient

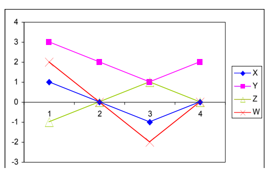
$$\rho = \frac{\sum_{j=1}^{p}(A[j]-\mu_A)(B[j]-\mu_B)}{\sqrt{\sum_{j=1}^{p}(A[j]-\mu_A)^2 \sum_{j=1}^{p}(B[j]-\mu_B)^2}} \qquad \mu_A = \frac{\sum_{j=1}^{p} A[j]}{p}$$



22

Examples

X	1	0	-1	0
Y	3	2	1	2
Z	-1	0	1	0
W	2	0	-2	0



Correlation (X,Y) = 1

Correlation (X,Z) = -1

Correlation (X,W) = 1

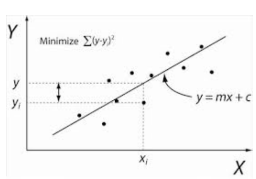
In-class exercise.

23

Linear regression

Regression analysis describes the relationship between two (or more) variables.

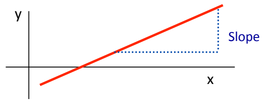
In correlation, the two variables are treated as equals. In regression, one variable is considered **independent (=predictor) variable (*X*)** and the other the **dependent (=outcome, response) variable *Y***.



24

Simple linear regression

- Relation between 2 continuous variables



$$y = \alpha + \beta_1 x_1$$

- Regression coefficient $\beta 1$
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
  - Least squares method
- Residual = actual data point – fitted value

25

Predicted value for an individual…

$$\hat{y}_i = \underbrace{\alpha + \beta^* x_i}_{\text{Fixed – exactly on the line}} + \boxed{\text{random error}_i}$$

Fixed – exactly on the line

Follows a normal distribution

26

Assumptions (or the fine print)

- Linear regression assumes that…
  - The relationship between X and Y is linear
  - Y is distributed normally at each value of X
  - The variance of Y at every value of X is the same (homogeneity of variances)
  - The observations are independent

27

Simple linear regression using a single predictor X.

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The *hat* symbol denotes an estimated value.

4 / 48

Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1-\hat{\beta}_0-\hat{\beta}_1 x_1)^2+(y_2-\hat{\beta}_0-\hat{\beta}_1 x_2)^2+\ldots+(y_n-\hat{\beta}_0-\hat{\beta}_1 x_n)^2.$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ are the sample means.

5 / 48

Credits

- Textbook Ch. 3
- Inspired by slides from Trevor Hastie, Robert Tibshirani, Kristin Sainani, Steven Buechler
- Data: http://www.maayanlab.net/crowdsourcing/megatask1.php

30