Statistics 101C Discussion Week 1

Brice Randolph

University of California, Los Angeles

brandolph@ucla.edu

March 29, 2016

Agenda

What will we be doing in discussion section?

- Lecture Key Points
- Probability Review (Confidence intervals today)
- Interview Question(s)
- Homework questions
- Something new or an extension upon a model seen in class

About Discussion

Can you ditch discussion? (8 am seriously?)

- I'm going to make this discussion as useful as possible (applications & interview questions)
- Those who come to the 8 am as opposed to the 9 am get free coffee (first come first served)
- Email policy (brandolph@ucla.edu)
- Office hours (Poll?)

The Dude Recommends

Some things I have learned since graduating:

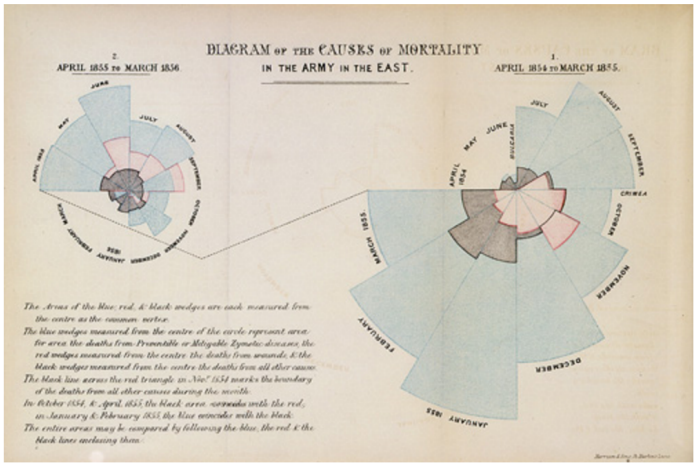
Github - take a few minutes to sign up.

Subscribe to a couple **data science blogs**

- http://www.r-bloggers.com/
- http://fivethirtyeight.com/
- http://fastml.com/
- https://www.quora.com/What-are-the-best-machine-learning-blogs-or-resources-available

The Dude Recommends

Think of your favorite example of a data visualization
(mine's the coxcomb -1856 Florence Nightingale)



Nightingale's coxcombs

Think of your favorite example of a data journalism
(Las Vegas Sun's "Do No Harm" about poor healthcare in south Nevada)

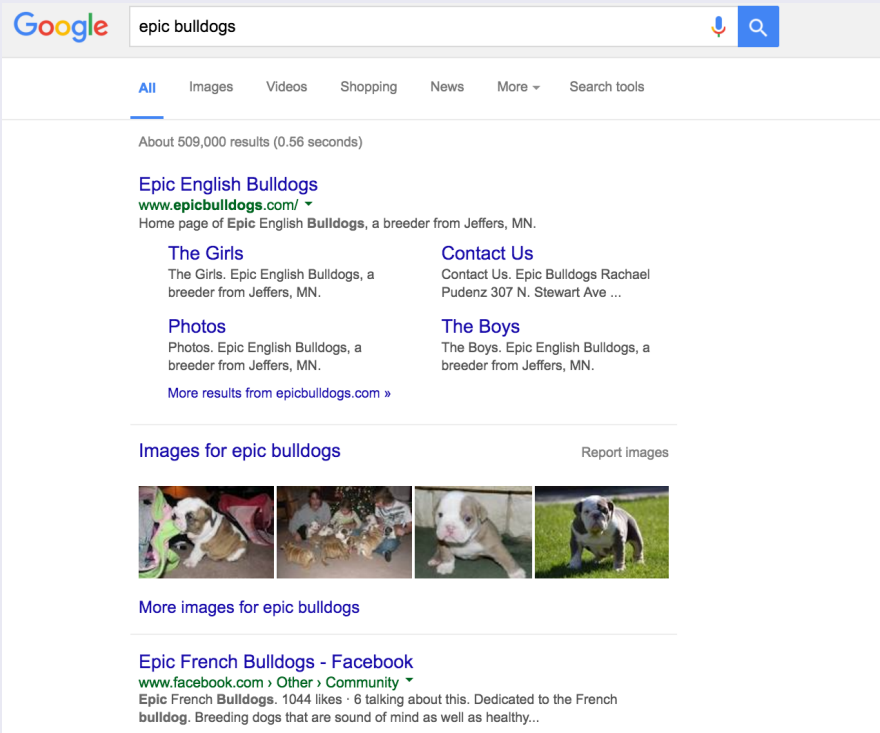
Can't think of one? Look one up and share it with your neighbor
(seriously, Go!)

Overview

What are statistical models used for?

Forecasting / Prediction:

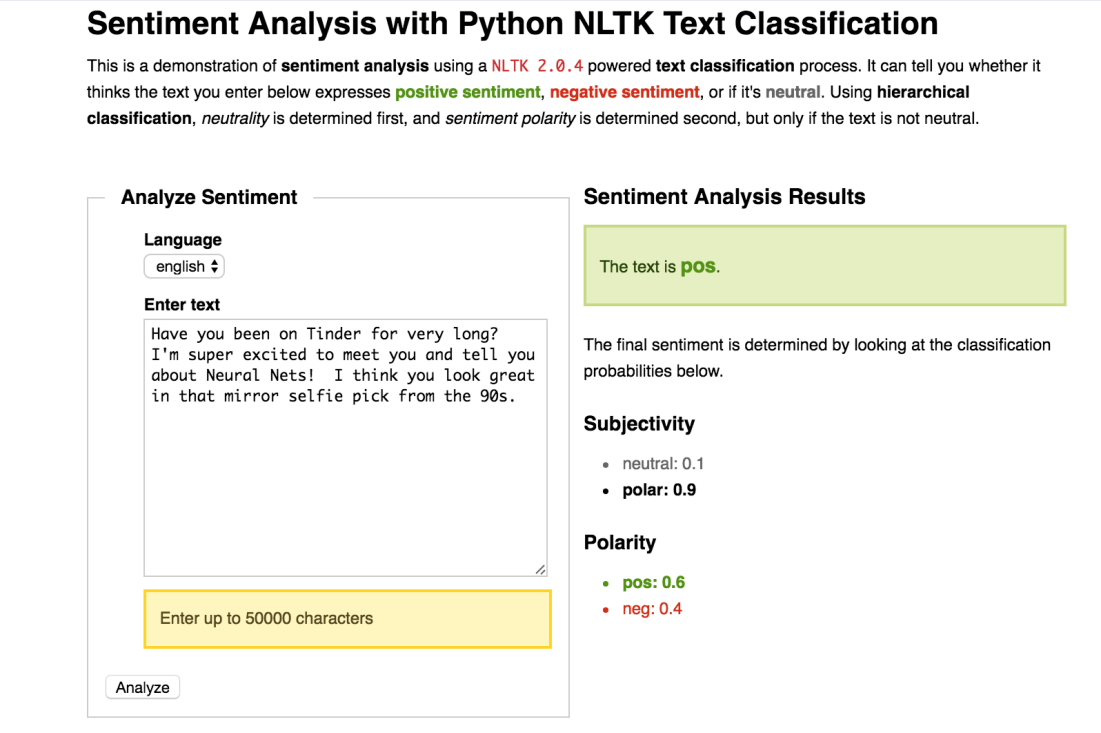
- Stock price
- Likelihood of contracting certain diseases
- Webpage ranking



Other Applications

Computer Vision/ Artificial Intelligence

- Object tracking
- Sentiment Analysis



Other Applications

Image classification/Scene Understanding

clarifai ABOUT PRICING DEVELOPER BLOG CONTACT US SIGN UP NC

Predicted Tags

no person ocean sea water sand beach travel vehicle seashore people

Similar Images

Image classification/Scene Understanding

clarifai

ABOUT PRICING DEVELOPER BLOG CONTACT US

*By using the demo you agree to our terms of service

Predicted Tags

portrait people woman one

adult girl facial expression

music fashion young

Your turn

I'm going to play a song for 30 seconds and I want you to tell me who sings it.

You can use any machine learning algorithms you have on you...

Shazam

Shazam uses a smartphone or computer's built-in microphone to gather a brief sample of audio being played. It creates an acoustic fingerprint based on the sample and compares it against a central database for a match. If it finds a match, it sends information such as the artist, song title, and album back to the user.

Final Exam (Competition)

Goal: **Predict** some outcome **Given** some predictors

Data: Euthanizing Animals (Predicting if an animal will be killed?)

Kaggle: You will upload your predictions here (go check out the website)

Team with best score wins

Machine learning in one sentence

Build a model to fit some process(static or dynamic) for prediction or inference.

You have learned some **pretty good** models already. Can you name some?

Spaghetti Model from class?

Why do I say they are good when they are so simple?

Why might these be used in industry? Think audience...

General Process Summary by Galit Shmueli

6

In Statistics, like in Data Mining, you start with data and a goal. In statistics there is a lot of focus on inference, that is, answering population-level questions using a sample. In data mining the focus is usually prediction: you create a model from your sample (training data) in order to predict test data.

The process in statistics is then:

1. Explore the data using summaries and graphs - depending on how data-driven the statistician, some will be more open-minded, looking at the data from all angles, while others (especially social scientists) will look at the data through the lens of the question of interest (e.g., plot especially the variables of interest and not others)
    1. Choose an appropriate statistical model family (e.g., linear regression for a continuous Y, logistic regression for a binary Y, or Poisson for count data), and perform model selection
    2. Estimate the final model
    3. Test model assumptions to make sure they are reasonably met (different from testing for predictive accuracy in data mining)
    4. Use the model for inference -- this is the main step that differs from data mining. The word "p-value" arrives here...

Take a look at any basic stats textbook and you'll find a chapter on Exploratory Data Analysis followed by some distributions (that will help choose reasonable approximating models), then inference (confidence intervals and hypothesis tests) and regression models.

I described to you the classic statistical process. However, I have many issues with it. The focus on inference has completely dominated the fields, while prediction (which is extremely important and useful) has been nearly neglected. Moreover, if you look at how social scientists use statistics for inference, you'll find that they use it quite differently! You can check out more about this here

share improve this answer

answered Nov 10 '10 at 14:38

community wiki
Galit Shmueli

ggplot2

Please download **ggplot2** if you haven't already

install.packages("ggplot2")

I will also be using *ggmap dplyr ggvis* and *shiny* at times

Graphics

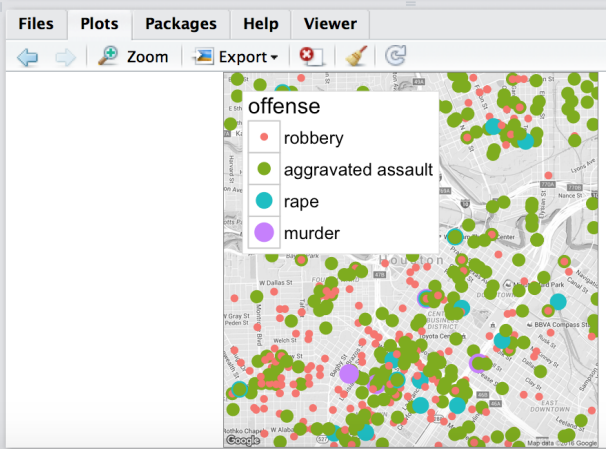
Only recently have graphical tools been easy to create.

They had only been used at the end of an analysis to summarize the report.

Take advantage of it!

```
107 library(ggplot2)
108 library(ggmap)
109 library(dplyr)
110 # find a reasonable spatial extent
111 qmap('houston', zoom = 13)
112 gglocator(2)
113
114 str(crime)
115 # only violent crimes
116 violent_crimes <- subset(crime,
117                          offense != "auto theft" & offense != "theft" & offens
118 # order violent crimes
119 violent_crimes$offense <- factor(violent_crimes$offense,
120                                  levels = c("robbery", "aggravated assaul
121 # restrict to downtown
122 violent_crimes <- subset(violent_crimes, -95.39681 <= lon & lon <= -95.34188
123                          29.73631 <= lat & lat <= 29.78400)
124
125 theme_set(theme_bw(16))
126 HoustonMap <- qmap("houston", zoom = 14, color = "bw") theme(legend.position=
127 HoustonMap +
128   geom_point(aes(x = lon, y = lat, colour = offense, size = offense),
129              data = violent_crimes)
130 HoustonMap +
131   stat_bin2d(
132     aes(x = lon, y = lat, colour = offense, fill = offense),
133     size = .5, bins = 30, alpha = 1/2,
134     data = violent_crimes
135   )
135:4  ggplot2 / ggmap example                                   R Script
Environment  History
To Console  To Source
data = violent_crimes)
```

```
1: Removed 11 rows containing non-finite values (stat_bin2d)
2: Removed 8 rows containing missing values (geom_tile).
>
>
> theme_set(theme_bw(16))
> HoustonMap <- qmap("houston", zoom = 14, color = "bw") the
(0.80, 0.3),legend.title=element_text(size=12))
Error: unexpected symbol in "HoustonMap <- qmap("houston", z
w") theme"
> HoustonMap +
+ geom_point(aes(x = lon, y = lat, colour = offense, size =
+ data = violent_crimes)
Warning message:
Removed 11 rows containing missing values (geom_point).
> #oh suhh dude
```



Visualizations

Check this out: one of the more advanced machine learning tools

http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Confidence intervals

Can someone tell me what a confidence interval might be used for?

Confidence intervals in the wild

Given the observed error (accuracy) of a model over a limited sample of data, how well does this error characterize its accuracy over additional instances?

Suppose we have

- a learned model h
- a test set S containing n instances drawn independently of one another and independent of h
- $n \geq 30$
- h makes r errors over the n instances

our best estimate of the error of h is

$$error_S(h) = \frac{r}{n}$$

Confidence intervals on error

With approximately N% probability, the true error lies in the interval

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

where z_N is a constant that depends on N (e.g. for 95% confidence, z_N =1.96)

Look familiar?

These last two examples came from here:
http://pages.cs.wisc.edu/ dpage/cs760/evaluating.pdf

Practice Interview Questions

Use R: Write some code that will simulate flipping a fair coin 100 times.

What are the basic assumptions used in linear regression? How do you empirically assess if errors are normally distributed? What are the most common estimation techniques for linear regression?

Any final questions/concerns/music recommendations?