Notes from the Berkeley textbook

Textbook is Inferential and Computational Thinking

Content

Here's the summary, with annotations

- Data Science
  - Introduction
    - Computational Tools
    - Statistical Techniques
  - Why Data Science?
  - Plotting the Classics
    - Literary Characters
    - Another Kind of Character
- Causality and Experiments
  - John Snow and the Broad Street Pump
  - Snow's "Grand Experiment"
  - Establishing Causality Is it just an association? The idea of a control group. Comparing groups that only differ in the attribute of interest.
  - Randomization Randomization as a way to make a good control group. Randomized controlled trials. Blind trials.
  - Endnote
- Programming in Python
  - Expressions
  - Numbers
  - Names
    - Example: Growth Rates
  - Call Expressions Calling functions. round, max. Different numbers of arguments.
- Data Types
  - Strings Adding strings. Converting numbers to strings with `str`
    - String Methods The idea of a function attached to an object. `upper`, `replace`.
  - Comparisons Including boolean variables. Comparison of strings.
  - Sequences Straight to arrays, with `make_array`. `sum`.
  - Arrays String and numeric arrays. Elementwise * and + with scalars. `size`, `sum`, `mean` methods. `import numpy as np`.
  - Ranges In fact, `arange`.
  - More on Arrays Elementwise combination of arrays.
- Tables Creating tables from data. Loading Minard data table. Adding columns to a table. Number of columns. Number of rows. Column names. Renaming columns. Getting data from columns. Formatting print output of columns. Selecting columns. Dropping columns.

  - Sorting Rows `show` to show some rows. `sort` by a column. `descending` as a option, and keyword arguments.
  - Selecting Rows By index, integer(s). By features - `t.where('SALARY'), are.above(10))`.
  - Example: Population Trends Subtracting columns.
  - Example: Trends in Gender Increasing F:M ratio as a function of age.
- Visualization IMDB etc databases. Scatter plots, line graphs.
  - Categorical Distributions Ice cream example. Bar charts, particularly horizontal bar charts. Sorting categories.
  - Numerical Distributions Histograms. Bins. Unequal bins. Counts vs proportions. Differences between bar charts and histograms. Grouping. Numerical variables as categories, and confusion therefrom.
  - Overlaid Graphs Categories on scatter plots, histograms, bar charts, line graphs.
- Functions and Tables Defining functions. Local scope. Docstrings. Multiple arguments.
  - Applying Functions to Columns Functions as values. Passing a function. Applying a function to a row of data.
  - Classifying by One Variable Group, then count. Group and sum, arbitrary functions.
  - Cross-Classifying Classifying by more than one variable. `group` again, and `pivot`.
  - Joining Tables by Columns The `join` method.
  - Bike Sharing in the Bay Area Using various table methods. `datascience` classes `Marker.map_table`, `Circle.map_table`.
- Randomness `np.random.choice`; comparisons, and booleans. Comparing strings. Comparisons in arrays. `count_nonzero`.
  - Conditional Statements `if`, `elif`, `else`, demonstrated inside functions.
  - Iteration `for`, using `np.arange`. Unrolling loops. `np.append`.
  - Simulation The idea of simulation. How to simulate. Heads and tails. Histograms, for loops. Rolling two dice in monopoly.
  - The Monty Hall Problem Simulations with functions and arrays.
  - Finding Probabilities Probabilities of combined events - multiplying. Conditional probabilities. Probabilities when there are two different ways of something happening.
- Sampling and Empirical Distributions Random and not-random samples. Sampling with uneven but known probability. Sampling every N, after starting at a random point (*systematic* sample). Sampling with and without replacement.
  - 10.1 Empirical Distributions Theoretical distributions. Observed distributions. Law of averages - as the sample number increases, the empirical distribution becomes more like the theoretical.
  - 10.2 Sampling from a Population Distribution of a random samples from population becomes more like distribution of population, as sample size increases.

  - 10.3 Empirical Distibution of a Statistic Numerical quantities from a population = *parameters*. Numerical quantities from samples: *statistics*. Statistic != Parameter. How different could it have been?
- 11. Testing Hypotheses Answers for yes/no questions from data.
  - 11.1 Assessing Models Model == "a set of assumptions about data". Statistic. Predicting statistic under the model. Compare prediction and data. Swain's jury. Mendel's flowers.
  - 11.2 Multiple Categories Race from California jury selections. Total variation distance as measure of distance between distributions. Simulating jurors, and variation distances. Explanations for failure of model.
  - 11.3 Decisions and Uncertainty Null-hypothesis, alternative hypothesis, test statistic, distribution under the null, conclusion. How to decide whether statistic is too unusual. Ronald Fisher and 0.05.
- 12. Comparing Two Samples Instead of comparing sample to known population.
  - 12.1 A/B Testing Mean difference of birth weights to smokers / non-smokers. Permutation.
  - 12.2 Deflategate Permutation test for Deflategate pressure measurements.
  - 12.3 Causality Proportion of patients with pain relief from Botulinum toxin. Random assignment of patients. Permutation test of proportions. Difficulty of drawing firm conclusions.
- 13. Estimation Sample statistic to estimate parameter. "How different could this estimate have been, if the sample had come out differently?"
  - 13.1 Percentiles Calculating percentiles. Quartiles.
  - 13.2 The Bootstrap Bootstrap for estimating variability of statistic. Salaries in SF. Estimate of median from sample using bootstrap. Do estimates *capture* the parameter.
  - 13.3 Confidence Intervals Bootstrap for 95% (etc) confidence intervals, mean baby birth weight. Bootstrap for proportions. Caveats for bootstrap.
  - 13.4 Using Confidence Intervals Testing whether a parameter is plausible. Hodgkin's Lymphoma, comparing drop in lung function over time, after treatment. Comparing drop to 0, using confidence intervals.
- 14. Why the Mean Matters Empirical distribution of mean is generally normal.
  - 14.1 Properties of the Mean Mean as "smoother". Depends only on distribution of values (not numbers). Balance point of histogram. Relationship of median and mean.
  - 14.2 Variability Deviations from average. Squared deviations. Mean squared deviation == variance -> SD. Most observations are a few standard SDs from the mean. Chebychev.
  - 14.3 The SD and the Normal Curve Point of inflection. Standard units, and the standard normal curve. The CDF. Proportion of area between +/-1 and 2 SDs.

  - 14.4 The Central Limit Theorem Simulation of sum of red / black choice in red -> normal. Average flight delay (where flight delay is positive-skewed). Proportion of purple flowers in Mendel's data. Width of sampling distribution as a function of sample size. No relationship to population size. Large samples make result valid for any distribution.
  - 14.5 The Variability of the Sample Mean Width of sampling distribution, sample size. Root N as scaling for SD.
  - 14.6 Choosing a Sample Size Using SDs to predict confidence intervals, and therefore, required sample size for given levels of confidence.
- 15. Prediction Predicting child height (y) from parents' height (x), by averaging over x intervals.
  - 15.1 Correlation MPG / acceleration data. Changing to standard units. Correlation and scatter, in standard units. "r measures the extent to which the scatter plot clusters around a straight line". r and the dot product of x, y in standard units. Not causation. Only measures linear association.
  - 15.2 The Regression Line The regression line and the 45 degree line. Regression slope from correlation coefficient. Prediction from the regression line. Meaning of the regression slope.
  - 15.3 The Method of Least Squares "Best" line. Line, predictions, and prediction error. Root mean squared error. Trying different lines. Regression line as best in terms of RMSE. Finding the line by numerical optimization.
  - 15.4 Least Squares Regression Shot-put weight lift / shot put distance correlation. Least squares line still least squares line, even if plot is not rugby-ball shaped. Find quadratic line using `minimize`.
  - 15.5 Visual Diagnostics Plotting residuals against parents' heights. Good regression -> "no pattern" the residuals. Residual plot of whale length and age reveals non-linearity in data. Heteroscedasticity ("uneven spread"). Acceleration (x) vs MPG (y). Residuals more variable for lower acceleration. Regression estimates less accurate for lower acceleration values.
  - 15.6 Numerical Diagnostics x vs residual plot has slope (very near) 0. Average of residuals always 0. SD of residuals predictable from SD of y and r.
- 16. Inference for Regression From regression line in sample to inference on the population.
  - 16.1 A Regression Model Regression model as true line plus random noise. Generating some points from the true line. The regression line as estimate.
  - 16.2 Inference for the True Slope Bootstrapping the regression line. 95% confidence interval. Non-zero regression slope when true line has slope 0. Inference on the slope with null of 0.
  - 16.3 Prediction Intervals Bootstrap regression to get prediction inter-

vals for all points on line.

- 17. Classification Examples of classifiers (fraudulent orders, compatible matches for dating websites, diagnosis of cancer, predicting votes. Observations. Attributes. Class (fraudulent, not-fraudulent). Training and testing sets. Can be imperfect.
  - 17.1 Nearest Neighbors Kidney disease. Hemoglobin, glucose. Classify new point. Nearest neighbor. Decision boundary. More difficult classifier - WBC and glucose. k nearest neighbors.
  - 17.2 Training and Testing Testing against unknown points - the testing set. Problem of testing nearest neighbor classifier on training set. Split data in half.
  - 17.3 Rows of Tables Rows as attributes of observations. Making rows into arrays. Pythagoras and Euclidean distance. Distance function. Applying functions to rows in table with `apply`. Apply "distance to new point" function. Select top five rows for this distance. Take majority vote on classification.
  - 17.4 Implementing the Classifier Classifying banknotes. Complex patterns of class on scatterplot. Three attributes and Pythagoras. More than three. Distance function for N-D. Wine dataset, classifying grape species. General k-nearest neighbors classifier.
  - 17.5 The Accuracy of the Classifier Hold-out method (splitting into test and training). 95% prediction success for test set. Breast cancer, school competition. kNN classifier. 96% accuracy.
  - 17.6 Multiple Regression House prices in Iowa. Prediction with slopes for each predicting variable. RMSE as criterion. Use `minimize` to find best slopes for training set. Residual plot - underestimation of high priced houses. kNN prediction (average of price of the kNNs).
- 18. Updating Predictions
  - 18.1 A "More Likely Than Not" Binary Classifier Bayes rule, students in second and third years, majors declared or not. The tree diagram.
  - 18.2 Making Decisions False positives, false negatives. Test for rare disease with low false positive rate. Table expressing known probabilities to demonstrate Bayes rule. Effect of priors (subjective probability).

Homework

For example, see https://github.com/data-8/data8assets/blob/gh-pages/materials/su17.

Datasets

- Larry Winner's dataset list.
- https://medium.com/datadriveninvestor/the-50-best-public-datasets-for-machine-learning-d80e9f030279.