Lecture 4

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

Implementation of Data Science Techniques

- Understand the data using `pandas`
- Visualize data
- Comment Code
- Plot Histogram for the variables in the data
- Plot pair-wise scatter plots (Diagonals plot the histogram)
  - diagonals represent the histogram for the variables
  - off-diagonal terms represent the correlation between variables depending on the inclination angle.
- Plot the correlation matrix to see the correlation clearly
  - Can remove the variables having perfect correlation with an other variable.
- Plot side-by-side boxplot to identify variables which can help in our ML-related task such as classification. These variables can be termed as `discriminatory` variables

Basic Statistical Testing

Up until now, we were doing informal digging of data, known as exploratory data analysis. Now, we will test statistically for distributions to analyze data.

- IID assumption (done in Lecture 2)
- MLE of parametrized distribution (done previously in Lecture 3)
  - Examples for finding ML estimate

Now, statistical testing

Recipe for statistical testing

1. **Explore** - Explore reasonable assumptions about the data, e.g. distribution type (including "cannot be assumed"), mean, variance, etc. and ask what do we want to verify
2. **Null Hypothesis** - Form null hypothesis $\mathbf{H_0}$ that we want to reject, e.g. "The two means are NOT different"

3. **Alternative Hypothesis** - Form alternative hypothesis $\mathbf{H_1}$ that we hope is true, e.g. “The two means are different”
4. Decide on a significance level (1 confidence) to reject the null hypothesis BEFORE performing a test, e.g. p < 0.05 or p < 0.01
5. Perform the test by performing the calculations
6. Check if the result was significant enough to reject the null hypothesis and accept the alternative hypothesis, i.e., the alternative hypothesis was not just a chance outcome, but we are 95% or 99% confident that it is more likely than the null hypothesis

Confidence Interval

Given a sample $x_1, x_2, \ldots, x_N$ and sample mean $\bar{x}$

Find the interval $\bar{x} \pm \epsilon$ within which the true mean lies within confidence $1 - \alpha$ -

$$Pr(|\bar{x} - \mu| > \epsilon) < \alpha$$

When sample std. dev. is not known, replace with true std. deviation.

Comparing two independent set of samples

Welch's t-test:

$$t = \frac{\mu_X - \mu_Y}{\sqrt{\frac{\sigma_x^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

is matched to a table for the appropriate degrees of freedom (DoF):

$$\frac{\left(\frac{\sigma_x^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)^2}{\frac{\sigma_x^4}{n_X^2(n_X-1)} + \frac{\sigma_Y^4}{n_Y^2(n_Y-1)}}$$

Comparing means of paired samples

Take the differences of the two sample values and store in a column Δ

With an assumed distribution

Without assuming a distribution

- Wilcoxon signed rank test :
  - Calculate Δ for the variables.
  - Add all ranks of positive and negative Δ separately (ranked in order of $|\Delta|$)

- Pick the smaller sum of ranks as test stat w_{test}.

$$w_{test} = \min\left(\sum_{i:d_i \geq 0} r_i, \sum_{j:d_j<0} r_j\right)$$

- Test stat w_{test} should be smaller than $w_{critical}$ (obtained from the table) for the given N
- Compare the Δ distribution.

Linearly related variables

Two paired continuous variables

```
Correlation does not imply Causation
```

Pearson's Correlation Coefficient:

- Ranges from -1 (perfectly negative correlated) to +1 (perfectly positive correlated)
- Represented by $\rho_{X,Y}$

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[X,Y] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2}\sqrt{\mathbb{E}[Y^2] - \mathbb{E}[Y]^2}}$$

- For a sample

$$r_{x,y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Here, we will be using **Spearman's Correlation Coefficient** which is same as Pearson's correlation coefficient but for **ranks**. Used for non-parametric testing.

$$r_S = \rho_{R(X),R(Y)} = \frac{\mathrm{Cov}(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} = 1 - \frac{6\sum(R(x_i) - R(y_i))^2}{n(n^2-1)}$$

Choosing a stat test

- Frame your problem
  - Predictor and Outcome variable
- Check for widely acceptable tests among data-scientists.
- Check if the assumptions made by the test hold for your scenario.
- Else, make your own test by using an existing test as a base.

Goal: Check for statistical significance

Graphs

Basic Graphs

- Bar Graph
- Line Chart
- Pie Chart
- Scatter Plot

Chart Information

- Chart Title
- Axes Titles
- Axes Units
- Grid Lines (Horizontal and/or Vertical)
- Legend
- Error Bars
- Confidence Intervals

Graph Beautification

- Font Legibility
- Font Consistency
- Information-to-Ink Ratio (Add whitespace)
- Color Usage
- Sorting (Value/Alphabetical)
- Highlighting
- Boxes (to highlight specific regions in graphs)
- Call-Outs and Pointers
- Breakout Pie-Chart

Advanced Graphs

- Histogram
- Stacked Bar (Value or Percentage)
- Column Chart
- Area Chart
- CDF
- Dual Y-Axis Bar/Line Charts

- Donut Charts
- Mixed Charts
- Bubble Chart (3-D curves)
- Kaplan-Meier (or Survival) Curves - to show when failure occurs
- Box and Whisker Plots
- Violin Plot - shows several distributions together
- Gantt Chart
- Radar or Spider Chart, etc.

What to use?

- Dimensionality Reduction for high-dimensional data
- Data Transformations (log-scale, ratio, polar form)

HOMEWORK: RDBMS and SQL