Midterm

- In-Class Midterm Friday 2/7
  - What is Data Science , phases of data science
  - What is EDA and why it is important
  - Why are databases important
  - Can you write a simple SQL query or select a correct query from a multiple choice selection
  - Do you know why we need to apply various statistical analysis like mean, standard deviation, correlation
  - Why is data cleaning important and what are methods for dealing with missing values
  - Will NOT be expected to write code in Python

- Lab Midterm Tuesday 2/4 and Thursday 2/6
  - Will be given a python notebook with a datasets and questions
  - You are free to use labs, lectures, WWW to complete the midterm coding portion
  - You HAVE to turn in the midterm coding portion by the end of lab to receive a grade.

Data Cleaning

What is Data Cleaning?

- Data Cleaning is post or concurrent phase of EDA (Exploratory Data Analysis).

- But what is Data Cleaning?
  - Given one ore more datasets, you will be asked to answer specific questions or investigate a hypothesis.
  - The data, as presented, may not directly answer your questions, hence, you may need to:
    - remove columns
    - modify columns
    - remove duplicate values
    - deal with missing values
    - deal with outlier data
    - normalize or scale data to make the data fit within a range.

Fact about Data Cleaning

- Interesting Fact:
  - Over **70%** of the work you will do as a Data Scientist on any Data Science or Statistics project is cleaning your data and manipulating it to make it ready for modelling and analysis.

Why is Data Cleaning Important?

- Data cleaning is important because it impacts efficiency and results of the other phases of the project.
- If your data is messy and inconsistent, then you are not likely to draw any interesting conclusions because the algorithms may not work.
  - Basically “If Your Data Is Bad, Your Machine Learning Tools Are Useless” – Harvard Business Review
- Clean data can improve or contribute to :
  - Processing time
  - More accurate predications / or ability to learn a model from data

Examples of Data Cleaning Tasks

- *Need to identify missing data*
  - Sometimes, instead of NULL/NAN/empty-cell, data sets use a different response category to record missing data.
  - For example, if the number 9 is used to represent a missing value, you must designate in your program that this value represents missing data.

- *Need to recode responses to "no" based on skip patterns*
  - There are a number of skip outs in some data sets.
  - For example, if we ask someone whether or not they have ever smoked, and they say "no", it would not make sense to ask them more detailed questions about smoking.
  - Hence, when we analyze the more detailed question, be sure you are not recording those that answered "No" to the first question as missing data.

Examples of Data Cleaning Tasks

- *Need to collapse response categories*
  - If a variable has many response categories, it can be difficult to interpret the statistical analyses in which it is used.
  - Alternatively, there may be too few subjects or observations identified by one or more response categories to allow for a successful analysis.
  - In these cases, you would need to collapse across categories.

- *Need to aggregate variables/fields/columns*
  - Consider if you want to combine multiple variables into one.

Examples of Data Cleaning Tasks

- *Need to create continuous variables*
  - If you are working with a number of items that represent a single construct, it may be useful to create a composite variable/score.

- *Labeling variable responses/values*
  - Given that nominal and ordinal variables have, or are given numeric response values (i.e. dummy codes), it can be useful to label those values so that the labels are displayed in your output.

Strategies for Dealing with Missing Data

- There are several strategies for dealing with missing data.

- **Reducing dataset**
  - The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all samples with missing values.

- **Replace Missing Value with Mean**
  - This method replaces each missing value with mean of the attribute (Kantardzic, M. 2003). The mean is calculated based on all known values of the attribute. This method is usable only for numeric attributes and is usually combined with replacing missing values with most common attribute value for symbolic attributes.

- **Replace Missing Value with Median for the Given Class**
  - Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given attribute is replaced by the median of all known values of that attribute in the class where the instance with the missing value belongs (Acuña, E. & Rodriguez, C. 2004). This method is usable only for numeric attributes and requires existence of classes or possibility to create classes as previous method

Missing Values (Cont.)

- **Replace Missing Value with Most Common Attribute Value**
  - This method simply uses most common attribute value for missing value imputation (Grzymala-Busse J. W., Hu M. 2001). The most common value of all values of the attribute is used. This method is usable only for symbolic attributes and is usually combined with replacing missing values with missing values imputation using mean for numeric attributes.

- **Closest Fit**
  - The closest fit algorithm (Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse, W. J. & Zheng X. 2005) for missing attribute values is based on replacing a missing attribute value with an existing value of the same attribute from another case that resembles as much as possible the case with missing attribute values.
  - This approach can be generalized to using K nearest neighbor, which will look for more than one (“k”) similar cases with known values of the attribute (Hand, D., Mannila, H. & Smyth, P. 2001).

Noisy Data

- Noise is a random error or variance in a measured variable.
- Noise can occur in data due to
  - Faulty data collection instruments
  - Data entry problems
  - Errors in data integration
- Lets take a look at the following data smoothing techniques

Data Cleaning (Cont.)

- Binning method
  - First sort data and partition into (equi-depth) bins
  - Smooth by bin means, smooth by bin median, or smooth by bin boundaries.
  - For example, in smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
  - When smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.
- Clustering
  - Detect and remove outliers
- Regression
  - Smooth data values using Linear regression

Data Cleaning

- Actually, data cleaning is not just one or two steps, it's a process of data transformation and exploration.

- Typically, data transformation includes:
  - **Smoothing –** works to remove noise from the data using techniques such as binning, regression or clustering.
  - **Attribute construction (or feature construction)** – creates new attributes from the given set of attributes to help the data mining process.
  - **Normalization** – attributes are scaled so as to fall within a smaller range, such as -1 to 1, or 0 to 1. Normalization can include (min-max normalization or z-score normalization).
  - **Discretization** – raw values of a numeric attribute (e.g. age) are replaced by interval labels (e.g. 0 – 10, or 11 – 20, etc.) or conceptual labels (e.g. youth, adult, senior).
  - **Concept hierarchy generation for nominal data** - attributes, such as street name, can be generalized to higher-level concepts, such as city or country.

Data glitches are also an Issue

- Systemic changes to data which are external to the recorded process.
  - Changes in data layout / data types
    - Integer becomes string, fields swap positions, etc.
  - Changes in scale / format
    - Is it Dollars or euros ? Or both are recorded !!
  - Temporary reversion to defaults
    - Failure of a processing step
  - Gaps in time series
    - Especially when records represent incremental changes.

Conventional Definition of Data Quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
    - Special problems in federated data: time consistency.
- Consistency
  - The data agrees with itself.