NASA GeneLab Project Proposal

Author: Alex D. Seo

Basic steps to follow for Data Analysis

*Details about particular theory will not be dealt in this paper. Meaning that, for example, I won't explain what Simple Linear Regression is.
*This paper has been proof read by Dr. Sebastian Raschka (red text)

Step 0: Data Collection : Data will be provided through NASA GeneLab
- Otherwise...Data scraping, Data Crawling

Step 1: Define the Question/Problem
- What is the purpose of the question?(Prediction? Decision Making? Data Mining? etc…)
    - Is this a *clustering, association rule, anomaly detection* or *predictive modeling task?*
    - If it is predictive modeling, is it
        - Classification
        - Regression
        - Or Ordinal Regression?
- What kind of data do we have? What do we know?**(Data Understanding)**
    1. To understand the data we need some background knowledge (How is it collected?Who collected?Is there relevant study?)
    2. To understand the data we should look at what data do we have
       ex)checking units of quantitative data(is it ratio?is it interval?is it binary?)
       checking if there are qualitative data(Is it nominal? Is it ordinal?)
       checking if there are time series data / sequential data etc….
       categorical: nominal, ordinal
       numeric: interval, ratio

    This sub-step can be followed by **Data Exploration**.

- How should we interpret our data?(There could be different perspectives of looking at the dataset)
- What kind of information can we obtain from the dataset?
- What's special of our data? What traits does our dataset have?(ex) Dimensionality, Sparsity, Outliers)

This step will determine how you will going to analyze the data

Step 2 : Acquire training data and testing data
- Usually we need to divide the datasets into training dataset and testing dataset.
    - Shuffle the dataset before splitting (sometimes datasets are ordered by target variable
    - If this is a classification problem, make sure you use a stratified split
- FYI- Training dataset is dataset that is used to build your model
  Testing dataset is dataset to test how well your model performs on unseen data
  Normally, training : testing = 7:3 or 8:2
  Use validation set (subset of the training set) or cross-validation (k-fold, leave one out) during model tuning

- Sometimes, we divide the dataset to also have validation dataset but let's use….
  - K-fold Cross-Validation (Hyperparameter Tuning)
    - Split your training dataset into k equal partition(folds); the smaller the dataset, the larger the number of folds
    - Pick one partition(let's say ith partition), and use that ith partition for testing and use all other partition for training
    - Repeat this procedure k times
    - Then, get the average of performance(ex) error)

  *This will be used for model selection and model validation

Step 3 : Wrangle, prepare, cleanse the data **(Data preprocessing)**

*Before this step you should have fully understand the dataset that you have

- Data Cleaning

  -Data in the real world will not be clean like data from the course! They are nasty.

  -They are incomplete(Missing data), noisy(Outliers), inconsistent(F,Female,Girl)
  - Incomplete Data
    - For some missing data, there could be pattern. Why are they missing? Is it intended? Or completely a mistake? Think about it to choose the best solution
    - Sol1- Delete the data point. That was easy! (But, data is too valuable..)
    - Sol2- Assign statistics derived from non-missing values (ex) mean, median, mode, global constant..) But, this might affect your whole statistics.
    - Sol3- Assign value derived from similar non missing instance. What makes similar? Use Clustering
    - Sol4- Assign value derived from simple model(ex) SLR)
    - Sol5- Splitting and Pooling, Divide your datasets into multiple dataset that has different values for the missing data(use any solution) and integrate the final result, or pick the best dataset.
    - Sol6- Assign value based on the probability distribution of the non-missing data(ex) Maximum Likelihood Estimation)

    Well… What's the best solution? Nobody knows. It depends on your dataset :)

    But, always try the simple one first and revisit this part and investigate how important the strategy is for the modeling outcome!

    Maybe we can also find the best solution using statistical method.

  - Noisy Data
    - What is noise? How are you going to make the noisy data shut up? Noise : Random error or variance in a measured variable
    - Noisy Data can actually refer to data with any kind of problem(Incomplete, Inconsistent…) But let's deal with outliers at this part.
    - What are you going to do with outliers?But, First, how do you know if it is outlier? There are too many methods to apply with outliers….
    - Let's look at some most common and simple method to find them
    - Extreme values
      - Cook's Distance, Leverage points
      - IQR
      - Z-Score
    - Outliers are not always extreme, then, how do you detect them?

        - Perform clustering(ex) k-means..)
        - Distance based (ex) mahalanobis distance, k-nn...)

    Always double check the outlier! They might carry important information.

- Inconsistent Data
    - Find it & Fix(Unify) it.
    - There might be duplicate records but seems different due to different units. Always unify the units and scales.
      If they are duplicate records than delete one.

Also, check for class imbalance (if this is a classification problem)

- Again, use the simplest approach first (=not dealing with it), but later, explore different strategies (oversampling, SMOTE, etc., see: https://github.com/scikit-learn-contrib/imbalanced-learn)

- Data Integration
    - There is a case when you need to integrate datasets into one big dataset.
    - In this case, you need to deal with tone of inconsistent data.
    - There might be variables that seems different but actually the same! Watch out for them and try to delete that variable.
      Fastest way to check is by checking correlation.
    - More details can be related to DB study.

- Data Transformation
    - By transforming data you can make your model more explainable!
    - Distributional Transform
        - Commonly used distributional transformations are log(log or ln), square root(sqrt(x), inverse(1/x)
    - Normalization / Standardization
        - What is normalization? Why do we normalize the value?
        - The goal of normalization is to make an entire set of values have a particular property
        - To see if we need normalization we check the similarity between two data. Normally we use distance between two data(ex) Euclidean distance)
        - Min-Max Normalization
            - Let's say the minimum value of the feature(or variable) is a, maximum value of the feature is b
            - If we want to change the feature's range to [x,y], we perform min-max normalization
            - V (new value) = (v - a / b - a) * (y - x)
        - Z-score Standardization
            - To change the value to a common distribution with average of 0 and sd of 1.
            - V (new value) = v- mean / sd
    - Discretization
        - Sometimes we need to change continuous data into categorical(binary) data, because sometimes you don't need exact precision
        - Binning(ex) equal-width , equal-frequency) , Clustering
    - Feature Construction

        - You can also create new variable from existing variables!
          ex) Date of Birth => Age
    - Qualitative Data
        - Transforming Qualitative data to Quantitative data is fairly normal case
          ex) Male, Female => 0,1 , Dummy Variable
        - However, it depends. You can also analyze your data with qualitative data
            - Text Mining
- Data Reduction
    - If your dataset is too big, or you want to be efficient you can reduce some data!
    - Sampling
        - Sampling can be the best way to reduce your dataset and still have representatives of the data.
        - Sampling with replacement? Or Sampling without replacement?
        - Simple Random Sampling? Or Systematic Sampling?
        - Stratified Sampling? Or Cluster Sampling?
        - What should we do with imbalanced data?
            - Under Sampling? Or Over Sampling?
        - Choose the best sampling method for your dataset!
        - Always double check if the your sample is biased!
    - Feature selection
        - Filters
        - Wrappers
    - Dimensionality reduction
        - PCA (Eigenvalues and Eigenvectors)
        - Also consider non-linear techniques: kernel PCA, locally linear embedding, t-SNE, UMAP (https://github.com/lmcinnes/umap),

Good Job! Your Dataset is finally ready :)

Step 4: Analyze, identify patterns, and explore the data **(Data Exploration)**
*Simple is the Best way! Let's use **Plots**!

- By this step, you should already have bunch of plots ready from data understanding
- Then, Let's deeply explore data!
- Univariate Analysis (Fancy name, but it's the easiest part)
    - Mean, Median, Standard Deviation, Distribution check etc…
- Multivariate Analysis
    - Correlation Coefficient
    - Covariance
- Build a Hypothesis and test it
    - Parametric Tests
        - Fitness Test
        - Chi-squared Test
        - 1-sample, 2-sample t test
        - ANOVA
    - Non-parametric Tests
        - Permutation test (http://rasbt.github.io/mlxtend/user_guide/evaluate/permutation_test/)
        - Sign Test
        - Wilcoxon Test
        - Signed ranks Test
        - Spearman Test

        - Kruskal-Wallis Test
        - Friedman Test
        - Run Test
        - Mann-Whitney Test
- Measure of Similarity
    - Euclidean distance (if data is normalized, and this is reasonable)
    - Jaccard distance (for binary variables)
    - SMC(Simple Matching Coefficient)
    - Mahalanobis Distance

Step 5: Build a Model to predict / solve the problem(+Diagnosis / Validation)

- There are so many models. You need to select your model that is most suitable to your question purpose. Then, choose the best performing model.
- Methods can be divided into three ways according to how you train the dataset.
    - Training Methods
        - Supervised Learning
            - Labeled
            - When you have teacher
            - When you have input & output (ex) Regression, Classification)
        - Unsupervised Learning
            - Unlabeled
            - When you have no teacher
            - When you only have input data (ex) Clustering, Association)
        - Reinforcement Learning
            - When you have a teacher but treats you like a dog
            - Reward-based learning
            - Needs environment that changes according to agent's action
- Models
    - Regression
        - Simple Linear Regression
        - Multiple Linear Regression
        - Polynomial Model (Non Linear Regression)
        - Regularized Regression
            - Gradient Descent Method
            - Ridge
            - LASSO
            - Elastic Net
        - Decision tree regression, random forest regression
        - Support vector machine regression
        - Robust regression (RANSAC)
    - Classification
        - Generalized Linear Model
            - Logistic Regression
            - Multinomial Classification
        - Naive Bayesian Classifier
        - K-Nearest Neighbors
        - Decision Tree
        - Random Forest
        - Support Vector Machine
        - Ensemble
    - Neural Network
        - Logistic regression, multinomial logistic regression (single-layer neural network)

        - Artificial Neural Network (multilayer perceptron)
        - Convolutional Neural Network
        - Recurrent Neural Network
    - Clustering
        - K-Means
        - Hierarchical clustering
        - DBSCAN
    - Association Rules
        - Apriori Algorithm + frequent patterns (http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/, http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/)

- Model Selection
    - K-fold Cross-Validation
    - Model Accuracy
        - Confusion Matrix (Used for classification model)
          Accuracy = Number of correct predictions / Total Number of predictions
          = TP+TN / TP+TN+FP+FN
    - F1 score
    - ROC curve (Used for classification model); ROC area under the curve
    - Precision recall curves (similar to ROC)
- Model Validation
    - Is the model underfitting?or overfitting? How does your model perform? This is extension of model selection! Because, You need to pick a valid model.
    - Bias-Variance tradeoff
    - Holdout
    - K-fold Cross-Validation, Leave one out cross validation
    - Bootstrapping
- Diagnosis
    - Linearity? Homoscedasticity? Normality? It all depends on your models' hypothesis

Step 6 : Visualize to report / present the problem solving steps and final solutions

- Some terms you need to consider in **Data Visualization**, ACCENT (D.A. Burn)
    - Apprehension
    - Clarity
    - Consistency
    - Efficiency
    - Necessity
    - Truthfulness
- Principles for visualization (Mike Frank & Ed Vul)
    - Be true to your Research
    - Maximize information, Minimize ink (Again, Simple is the best way)
    - Organize Hierarchically (What to view first? What to see deeper?)
- Some Charts…
    - Histogram (Choose the range that can represent the data effectively!);
    - ECDF (Empirical cumulative distribution plot)
    - Pie Chart
    - Scatter plot
    - Box plot

- Heat map
- Violin plot
- Line Graph

Tip - Use Hybrid plots!

Information vs Readability… You choose!

Step 7: Get the results 😊