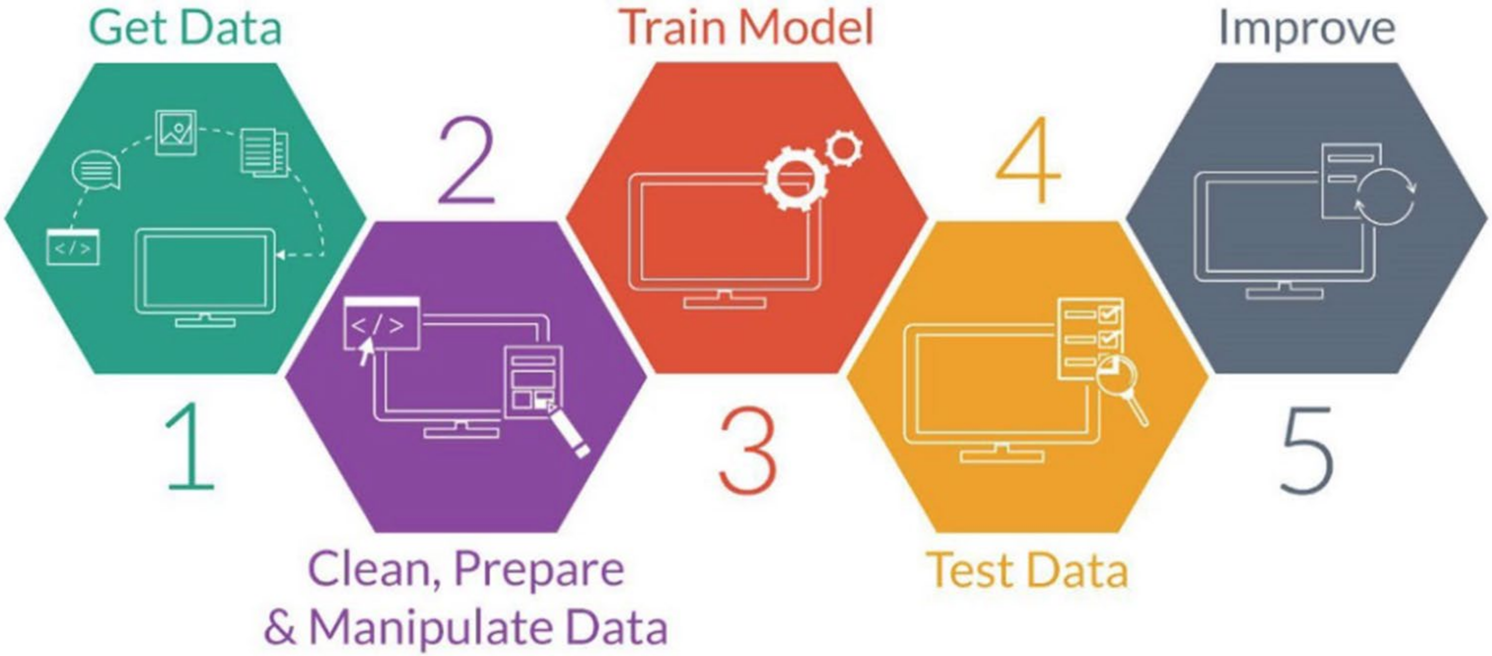
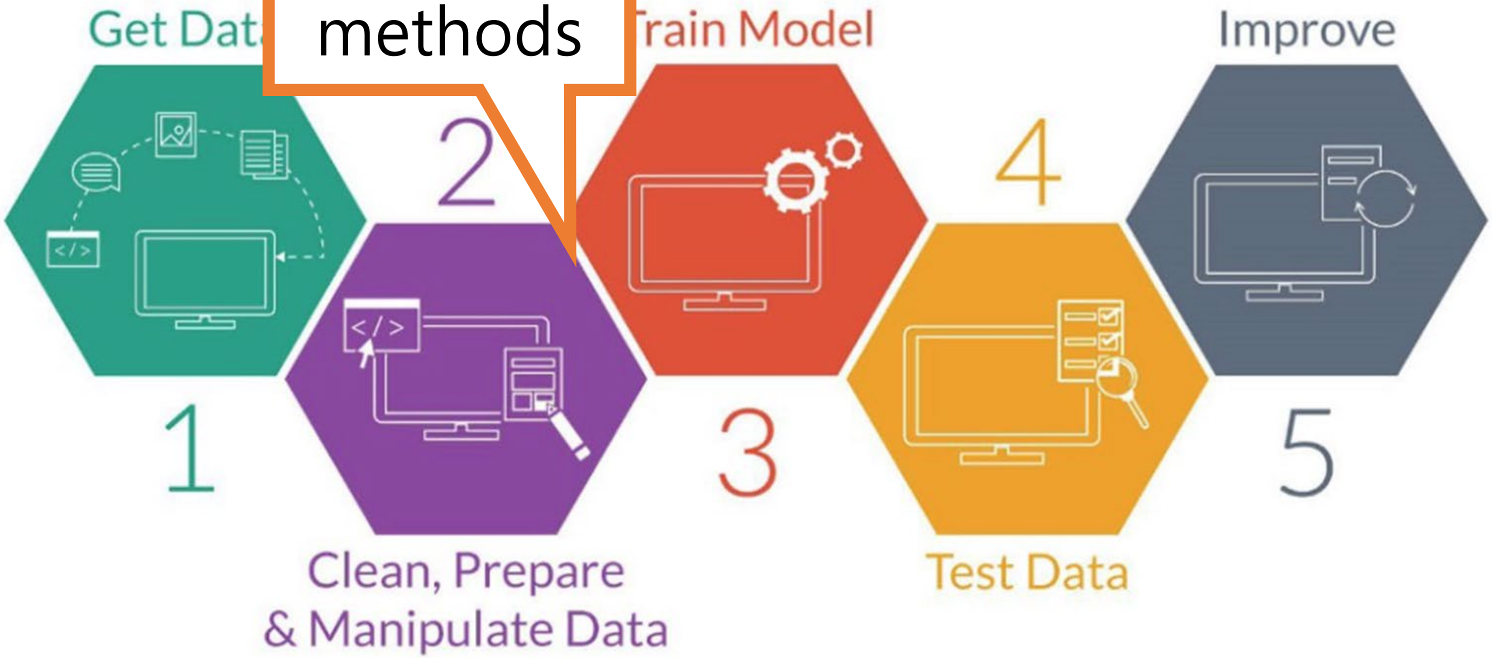
Details of ML part 6

Basic Processes for Building ML Models



https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p5.pdf

Some Methods MAY Improve Model



2.5 some methods

https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p5.pdf

Feature Scaling

- Feature Scaling is where we force the values from different columns to exist on the same scale, in order to enhance the learning capabilities of the model
  1. Standardization
  2. Normalization
  3. Min-max scaling

Standardization

- Standardization rescales the data to have a mean of 0 and a standard deviation of 1
- It assumes that your data has a Gaussian (bell curve) distribution

Normalization

- Normalization rescales data so that it exists in a range between 0 and 1
- It is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (bell curve)

Min-max scaling

- Min-max scaling performs a linear transformation on the original data
- This technique gets all the scaled data in the range (0, 1)
- It preserves the relationships among the original data values
- The cost of having this bounded range is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

Covariance and Correlation

- Both are mathematical concepts that are also used in statistics and probability theory
- Most useful in understanding variables
- If one variable goes to increasing direction the same as another variable goes that direction, it means a positive correlation
- If both variable are in the opposite direction then called negative correlation

Covariance

$$\operatorname{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$$

- Numerator show, the quantity of variance in x multiplied by the quantity of variance in y.
- Unit of covariance shows, Unit of x multiplied by a unit of y
- Hence if we change the unit of variables, covariance also has new value but sign will remain the same.
- However if it is positive then both variables vary in the same direction else if it is negative then they vary in the opposite direction.

Correlation

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

- Correlation between two variables which is a normalized version of the covariance
- The range of correlation coefficients is always between -1 to 1. The correlation coefficient is also known as Pearson’s correlation coefficient

Correlation

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

- -1 and +1 indicate that both variables have a perfect linear relationship.
- Negative means they are inversely proportional to each other with the factor of correlation coefficient value.
- Positive means they are directly proportional to each other mean vary in the same direction with the factor of correlation coefficient value.
- if the correlation coefficient is 0 then it means there is no linear relationship between variables.

Differences

- Correlation is simply a normalized form of covariance. It is obviously important to be precise with language when discussing the two, but conceptually they are almost identical
- The value of the correlation coefficient ranges from [-1 – 1]. -1 is indicate for a negative relationship. 1 means a positive relationship. 0 means no relationship

Correlation Matrix

- Correlation matrix is simply a table which displays the correlation coefficients for different variables
- The matrix depicts the correlation between all the possible pairs of values in a table
- It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data

➔ Feature Selection

Feature Selection

- Feature Selection is the process used to select the input variables that are most important to the Machine Learning task
- In some cases, you may have access to a whole lot of set of potential predictors or variables. In this case, it can often be harmful to use all of these input variables or predictors in to your model. This is where feature selection comes in

Why use Feature Selection

- Improve Model Accuracy
- Lower Computational Cost
- Easier to Understand & Explain

Ways to conduct Feature Selection

1. Correlation Matrix
2. Univariate Testing
   - Regression Task
   - Classification Task
3. Recursive Feature Elimination with Cross-Validation (RFECV)

Univariate Testing

- Univariate Feature Selection or Testing applies statistical tests to find relationships between the output variable and each input variable in isolation
- Tests are conducted one input variable at a time
- The tests depends whether you are running a regression task or a classification task

Regression Task

- In a regression task, you may be provided with an f-score and a p-value for each variable and gives you a view of the statistical significance of their relationships between the input and the output variables
- This will help you assess how confident you should be with the variables you have used in your model

Classification Task

- Depending on what test you use, you might be provided a chi-square score and a p-value for each variable
- Again, this gives you a view of the statistical significance of their relationships between the input variables and the output variables

Recursive Feature Elimination with Cross-Validation (RFECV)

- Recursive Feature Elimination fits a model that starts with all the input variables, then iteratively removes those with the weakest relationship with the output until the desired number of features is reached
- It actually fits a model instead of just running statistical tests unlike the Univariate Testing

Dimension Reduction

- Previous Methods are belonged to dimension reduction
- Dimension reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension

Dimension Reduction

- Linear Discriminant Analysis (LDA) for supervised learning
- Principle Component Analysis (PCA) for unsupervised learning