Udemy- Data Science

Machine Learning

- Data Acquisition
- Data Cleaning
- Train, Val (For adjusting Hyperparameters) & Test Data
- Model Training and Building
- Model Testing
- Adjust Model parameters
- Deploy models

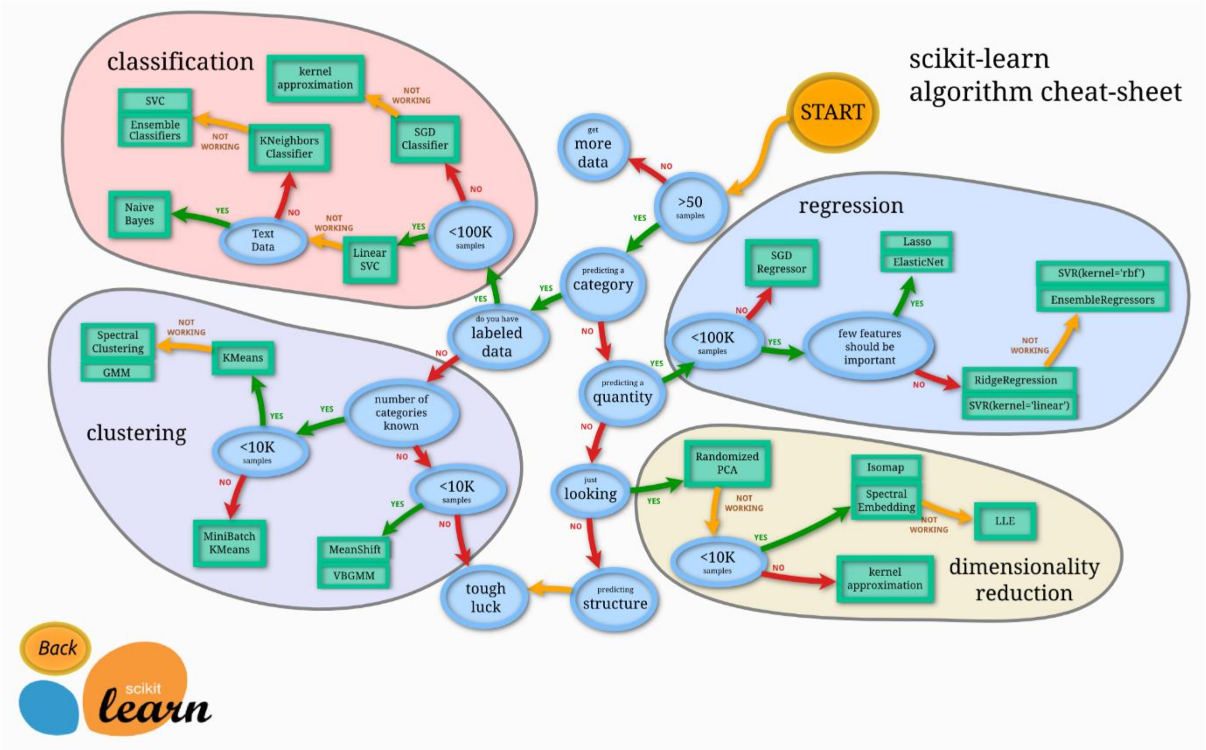
Model Evaluation

1. **Key Classification Metric**
   a. **Accuracy**
      i. correct / total
      ii. Useful when target classes are well balanced
   b. **Recall**
      i. TP / (TP + FN)
      ii. Ability of a model to predict all relevant cases
   c. **Precision**
      i. TP / (TP + FP)
      ii. Ability of a model to predict only relevant cases.
   d. **F1- Score**
      i. 2 * (P * R) / (P + R)
      ii. Combination of Precision and Recall, Harmonic Mean.
      iii. Harmonic Mean, HM punishes extreme values
   e. **Confusion Matrix**
      i. **TP FN**
      ii. **FP TN**

2. **Key Regression Metric**
    a. **Mean Absolute Error**
    b. **Mean Squared Error**
        i. Punishes extreme values
    c. **Root Mean Squared (RMS)**
        i. Square root of MSE
        ii. Most popular, has same unit of 'y'

Scikit Learn

1. From sklearn.family import Model (Estimator)
2. Split Train & Test data
3. Model.fit(X_train, y_train)
4. Model.predict(X_train)
5. Model.score()



Bias Variance Trade off

- Point where we are just adding noise by adding model complexity
- Training error goes down, but test error is starting to go up
- After this point, model begins to overfit

Logistic Regression

- Sigmoid / Logistic Function can take values from 0 & 1
- Confusion Matrix can be used for model evaluation.

KNN

- Classification algorithm
- High prediction cost
- Not for high dimension
- Not for categorical features

Decision Tree

- **Nodes,** split for the value of a certain attribute
- **Edges,** outcome of a split to next node
- Entropy and Information Gain are the methods to choose best split

Random Forest (Ensemble)

- Many trees with a random sample of features chosen as split.
- **Advantage**
    - If a very strong feature in data set, then most of the bagged trees will use that feature as top split, resulting in highly corelated ensemble.

    - By randomly leaving out candidate features from each split, Random Forest decorrelate the trees.
    - Averaging then will reduce the variance.

Support Vector Machines (SVM)

- Choose a hyperplane which maximise margin between classes
- Expand this idea to non-linear space using kernel trick

K Means

- Unlabelled data in Unsupervised Learning
- Group similar clusters together
    - Compute cluster centroid by taking mean vector of points.
    - Assign each data point to the cluster for which the centroid is closest.

Principal Component Analysis (PCA)

- General factor analysis
- Unsupervised
- Find which feature explains most variance in data

Recommendation Systems

1. **Content Based**
    a. Focus on attributes of items
    b. Based on similarity between items
2. **Collaborative Filtering**
    a. Wisdom of crowd
    b. Based on the knowledge of user's attitude to items
    c. More commonly used, better result

d. Memory Based or Model Based

Natural Language Processing (NLP)

- **Bag of Words,** A document represented as a vector of words.
- Cosine similarity on the vectors to determine similarity.
    - sim (A, B) = cos(theta) = A.B / (|A| * |B|)
- Improve in Bag of words by adjusting word counts based on their frequency in corpus.
- We can use **TF / IDF**
    - **TF (Term Frequency)**
        - Importance of term within document
        - Number of occurrences of term in the document
    - **IDF (Inverse Document Frequency)**
        - Importance of term in the corpus
        - IDF(t) = log(D/t)
        - log (Total Documents / no. of documents with term)
    - **TF-IDF = TF * IDF**

Big Data

Distributed System, which distribute data over multiple machines.

- Easier to scale

Hadoop is a way to distribute very large files across multiple machines.

- **HDFS, Hadoop Distributed File System**
    - Allows to work with large dataset
    - Duplicate blocks for fault tolerance
    - **MapReduce** allows computation on data
    - Has a name node & various data nodes attached.

    - Uses blocks of data of 128MB default each replicated 3 times.
    - Blocks are distributed in a way to support fault tolerance.
    - **Map Reduce** is a way to split a computation task to a distributed set of files such as HDFS.
    - It consists of Job Trackers and Task Trackers.
    - Job trackers send code to run on task tracker.
    - Task trackers allocate memory and CPU and monitor the task on worker node.

Spark

- Quickly and easily handle big data.
- Open Source on Apache
- Flexible alternate to MapReduce
- Spark can use data stored in a variety of formats
    - Cassandra, S3, HDFS … etc

Spark vs MapReduce

- Requires files to be stored in HDFS, Spark does not
- Spark can perform 100x faster
    - Spark keeps all the data in memory after each transformation, MapReduce only to disk.
    - Spark can use disk if memory is full

At the core RDD, Resilient Distributed Dataset

- RDD has 4 main features
    1. Distributed collection of data
    2. Fault Tolerant
    3. Parallel operation – partitioned
    4. Ability to use many data sources
- RDDs are immutable, lazily evaluated, and cacheable

- 2 types of RDD operations
    - Transformations
        - Filter, apply filter and return elements that are true
        - Map, transform each element
        - FlatMap, transform each element to 0-N elements.
    - Actions
        - Collect, return all elements of RDD as an array
        - Count, number of elements
        - First, first element in RDD
        - Take, return array with first n element

- Offer RDD is holding values in tuples (key, value)
    - Reduce, aggregate RDD elements using function that return single value
    - ReduceByKey,
        - Similar to Group by