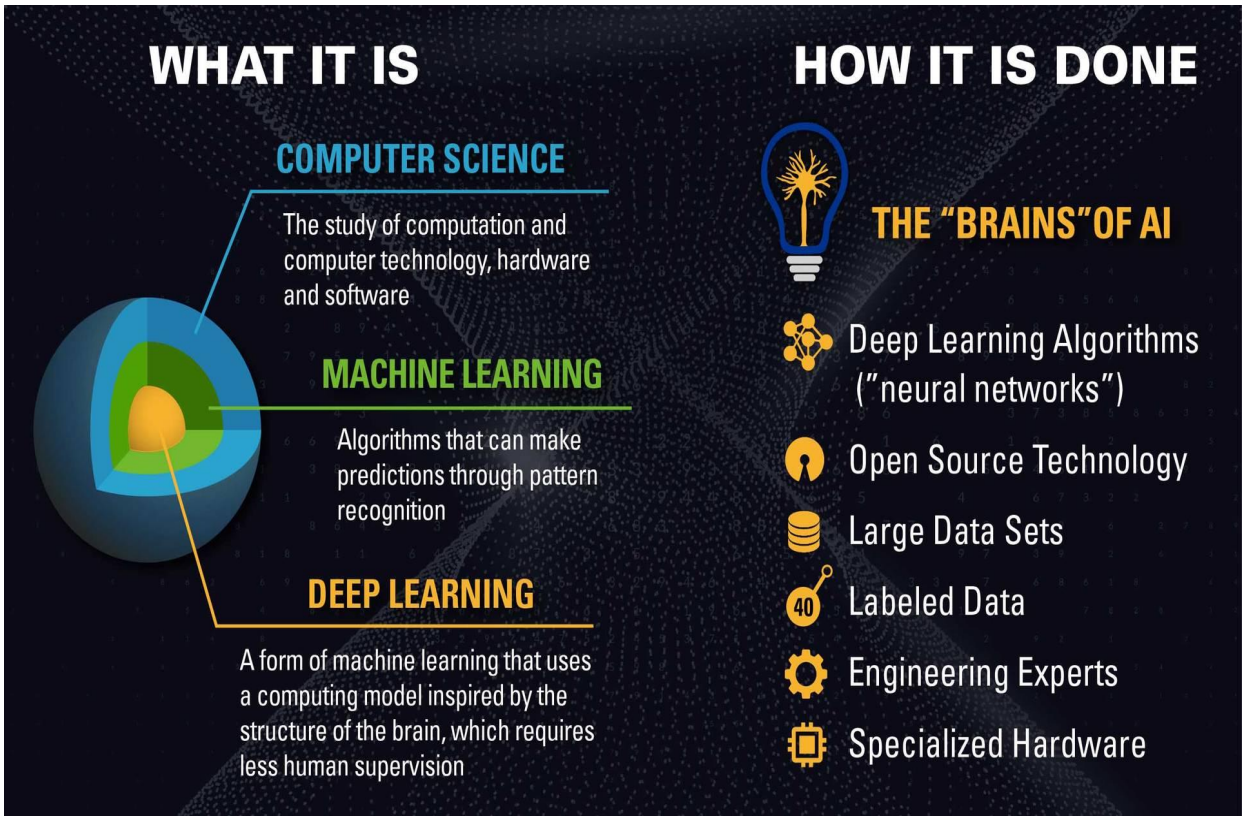

WHAT IT IS
COMPUTER SCIENCE
The study of computation and computer technology, hardware and software
MACHINE LEARNING
Algorithms that can make predictions through pattern recognition
DEEP LEARNING
A form of machine learning that uses a computing model inspired by the structure of the brain, which requires less human supervision
HOW IT IS DONE
THE "BRAINS" OF AI
Deep Learning Algorithms ("neural networks")
Open Source Technology
Large Data Sets
Labeled Data
Engineering Experts
Specialized Hardware

Index

Reference

1. https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/
2. https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/#:~:text=Bagging%20is%20a%20way%20to,based%20on%20the%20last%20classification.
3. https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca/
4. https://www.analyticsvidhya.com/blog/2017/09/30-questions-test-tree-based-models/
5. https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide
6. https://www.analyticsvidhya.com/blog/2017/10/svm-skilltest/
7. https://www.analyticsvidhya.com/blog/2017/03/questions-dimensionality-reduction-data-scientist/
8. https://emkademy.com/research/toolbox/2020-03-02-accuracy-precision-recall
9. https://en.wikipedia.org/wiki/Type_I_and_type_II_errors
10. https://en.wikipedia.org/wiki/Sensitivity_and_specificity
11. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc
12. https://machinelearningmastery.com/probability-calibration-for-imbalanced-classification/
13. https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/
14. https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392
15. https://towardsdatascience.com/supervised-machine-learning-model-validation-a-step-by-step-approach-771109ae0253 **(for model check),** https://slideplayer.com/slide/9174541/
16. https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f
17. https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79

1. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA? 1**. PCA is an unsupervised method 2. It searches for the directions that data have the largest variance 3. Maximum number of principal components <= number of features 4. All principal components are orthogonal to each other** **all the above are true**

2. In what type of **learning labeled training data** is used **Supervised learning**

3. **PCA can be used for projecting and visualizing data in lower dimensions.** **True**

4. **Dimensionality Reduction Algorithms are one of the possible ways to reduce the computation time required to build a model** **True**

5. **..Random Forest.**..... is a widely used and effective machine learning algorithm based on the idea of bagging.

6. **It is not necessary to have a target variable for applying dimensionality reduction algorithms** **True**

7. **OHE** is dimensionality reduction and numerical encoding.

8. **Linear Regression** = y =mx +c, y = f(x)

9. **SD** is the square root of variance

10. Mean 0 and SD is 1 always.

11. Movie recommendation is **Reinforcement learning**.

12. **Normalization** we can do anywhere, not only in the sklearn.

13. **K-means** is not lazy learning

14. Data should have always 92% accuracy

15. Feature always columns

16. After splitting only we normalize the data.

17. An under fitted model has **high bias and low variance**.

18. How we show plot for Outliers = **Boxplot, Z- Score, Scatter plot**

19. **Confusion Matrix** is used mostly in Supervised learning.

20. **Over sampling** is used when the amount of data collected is insufficient.

21. **Undersampling** is appropriate when there is plenty of data for an accurate analysis.

22. **RBF** is default in Kernel.

23. **Random Forest** and **Extra Tree** don't have learning rate as hyperparameter.

24. Distance measure clusters are **Manhattan distance, Euclidean distance, Minkowski distance, Jaccard similarity.**

25. If the **mean, median and the mode** of a set of numbers are **equal**, it means the **distribution is symmetric.**

26. we have to rescale the data before split or after? **Post split**

27. A perceptron adds up all the weighted inputs it receives, and if it exceeds a certain value, it outputs a 1, otherwise it just outputs a 0. **True**

28. What are support vectors?
all the examples that have a non-zero weight ??k in a svm, the only examples necessary to compute f(x) in an svm.

29. What is the p**urpose of the Kernel Trick**?

to transform the data from nonlinearly separable to linearly separable

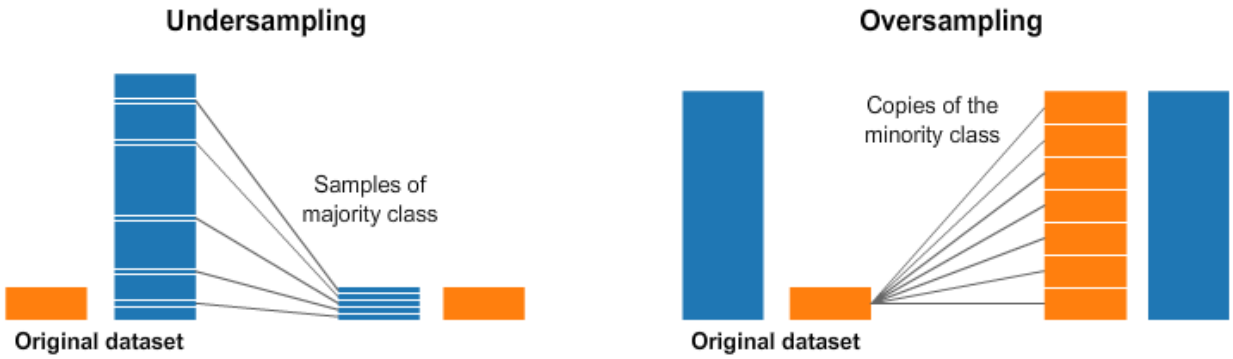
30. The firing rate of a neuron
is more analogous to the output of a unit in neural net than the output voltage of the neuron

Oversampling and Undersampling

There are two main approaches to random resampling for imbalanced classification; they are oversampling and undersampling.

- **Random Oversampling:** Randomly duplicate examples in the minority class.
- **Random Undersampling**: Randomly delete examples in the majority class.

Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset. **Random undersampling** involves randomly selecting examples from the majority class and deleting them from the training dataset.



- What is **the best technique for dealing with heavily imbalanced datasets**?

Resampling Technique

A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling).

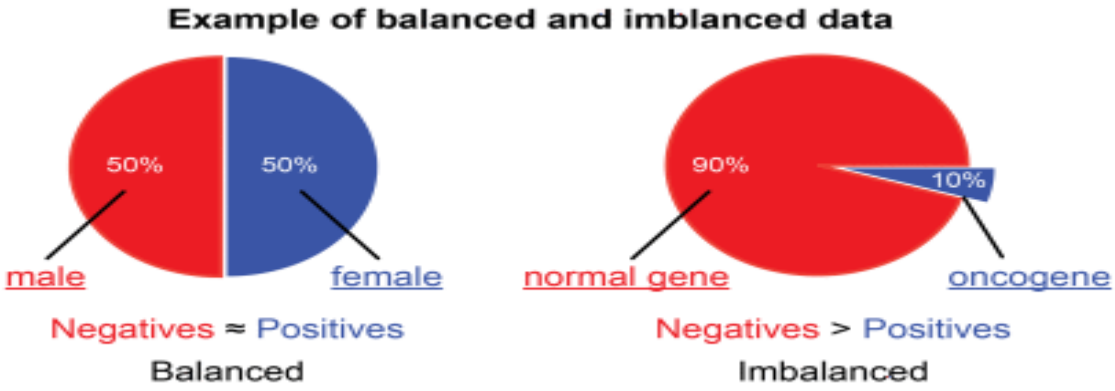
- What is an **imbalanced dataset with an example**?

A typical example of imbalanced data is **encountered in e-mail classification problems** where emails are classified into ham or spam. The number of spam emails is usually lower than the number of relevant (ham) emails. So, using the original distribution of two classes leads to an imbalanced dataset.

- What do you mean by **imbalanced data**?

A classification data set with skewed class proportions is called imbalanced. Classes that make up a **large proportion of the data set** are called **majority classes**. Those that make up a **smaller proportion are minority classes**.

In Simple words, Imbalanced data refers **to those types of datasets where the target class has an uneven distribution of observations**, i.e one class label has a very high number of observations.



Random forests

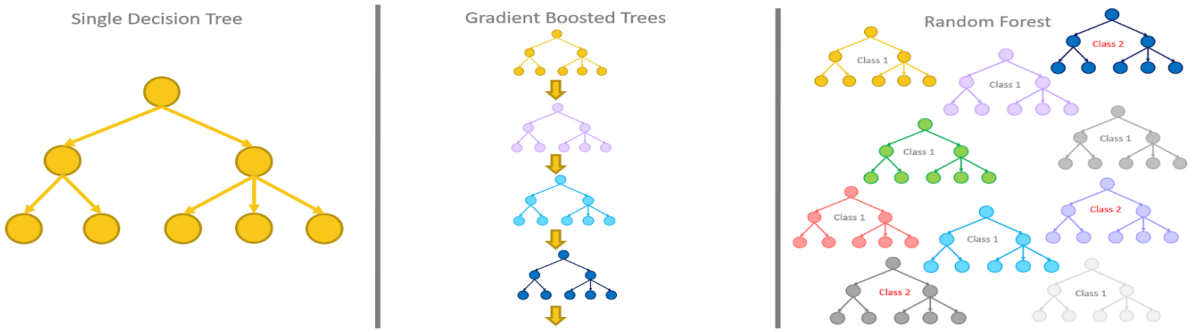
It is based on the concept of **bootstrap aggregation** (aka bagging). This is a theoretical foundation that shows that sampling **with replacement** and then building an ensemble **reduces the variance of the forest without increasing the bias.**

The same theoretical property is not true if you sample without replacement, **because sampling without a replacement would lead to pretty high variance**.

Let's say we're building a random forest with 1,000 trees, and our training set is 2,000 examples. If we sample without replacement we would train on 2 examples per tree. This is obviously impractical.

- Is random forest with replacement?

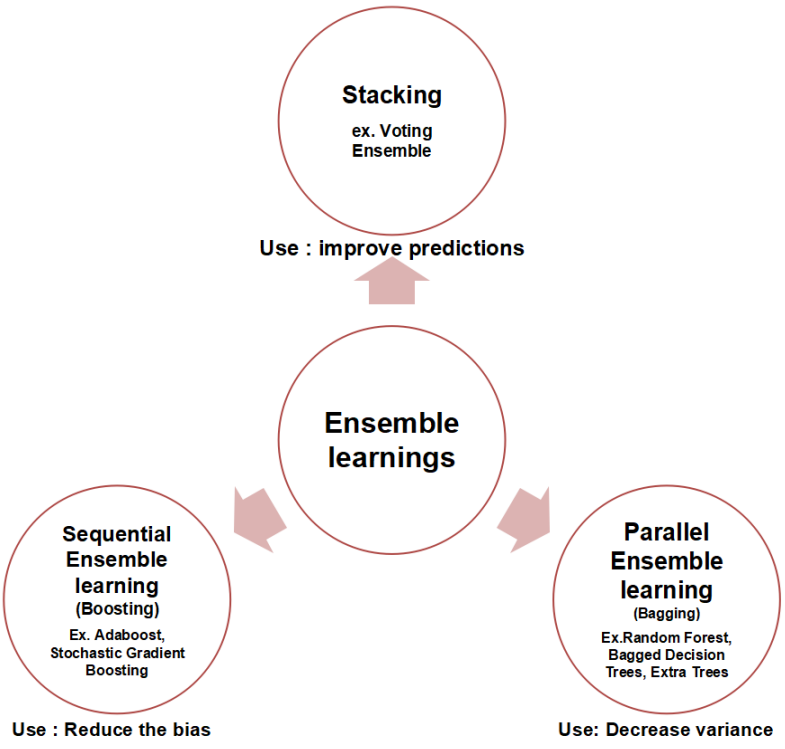
When training, each tree in a random forest learns from a random sample of the data points. **The samples are drawn with replacement, known as bootstrapping**, which means that some samples will be used multiple times in a single tree.

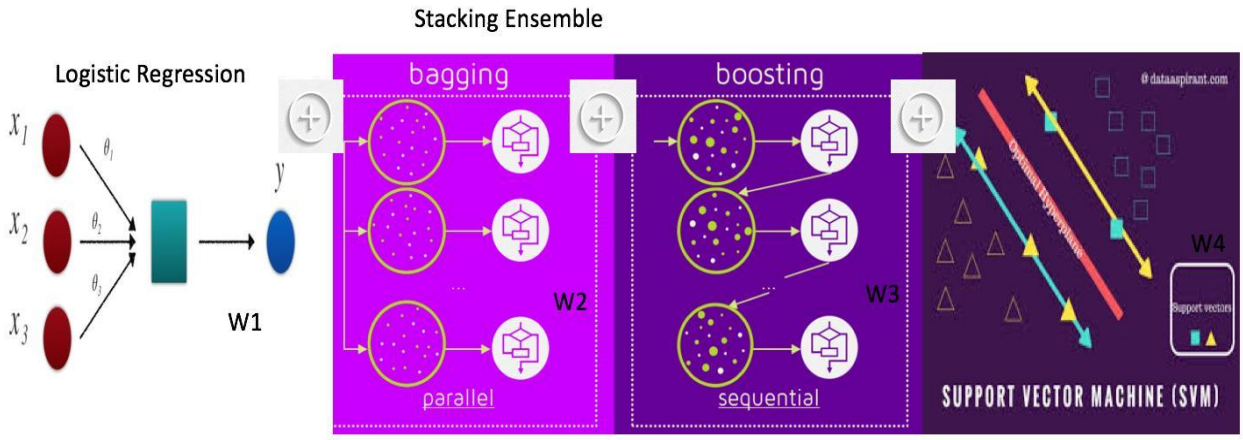


Ensemble method

Ensemble is a Machine Learning concept in which the idea is to train multiple models using the same learning algorithm. The ensembles take part in a bigger group of methods, called **multi classifiers, where a set of hundreds or thousands of learners with a common objective are fused together to solve the problem.**

The main causes of error in learning are due to **noise, bias and variance**. Ensemble helps to **minimize** these factors. These methods are designed **to improve the stability and the accuracy of Machine Learning algorithms.**





These are weight algorithmus we have in supervised learning.

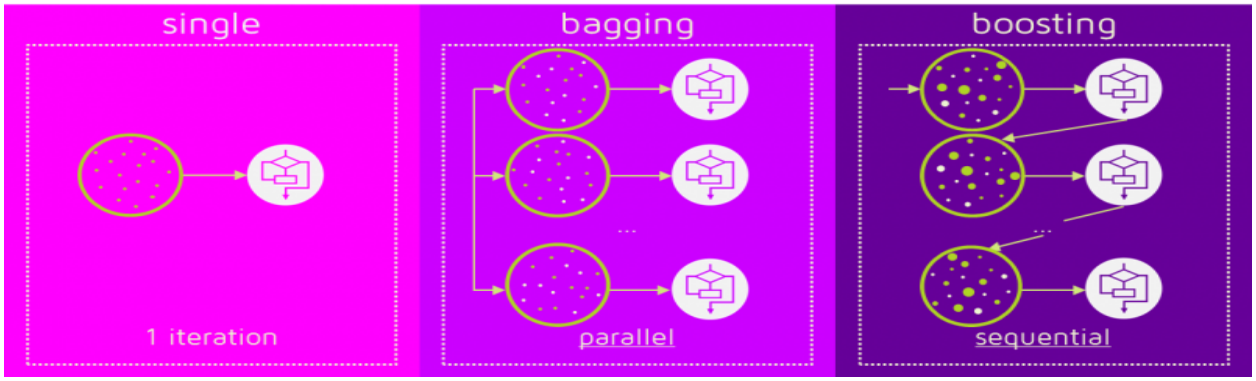
- Why do we need **bagging and boosting**?

Bagging attempts to tackle the over-fitting issue. If the classifier is **unstable** (**high variance, Overfitting**), then we need to apply **bagging**.

Boosting tries to **reduce bias**. If the classifier is **steady and straightforward** (**high bias, Underfitting**), then we need to apply **boosting**.

- What is a **bagging method**?

Bagging, also known as bootstrap aggregation, is the **ensemble learning method that is commonly used to reduce variance within a noisy dataset.** In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen from more than one.



Bagging	Boosting
Both the ensemble methods get N learners from 1 learner. But..	
..follows parallel ensemble techniques, i.e. base learners are formed independently.	..follows Sequential ensemble technique, i.e. base learners are dependent on the previous weak base learner.
Random sampling with replacement.	Random sampling with replacement over weighted data.
Both gives out the final prediction by taking average of N learners. But..	
..equal weights is given to all model. (equally weighted average)	..more weight is given to the model with better performance. (weighted average)
Both are good at providing high model scalability. But..	
..it reduces variance and solves the problem of overfitting.	..it reduces the bias but is more prone to overfitting. Overfitting can be avoided by tuning the parameters.

Bagging	Boosting
Individual trees/models are independent of each other.	Individual trees are not independent of each other.
There is no concept of learning from each other in bagging. Each individual model/tree will be fed with a sample of features/columns from the whole training set along with a sample of observations/rows for those features.	In boosting, each of the trees will learn from the mistakes of the previous tree and try to minimize the residual error as it keeps moving forward sequentially. For e.g.: I have a boosting algorithm with 3 sequential models M1, M2 and M3. M2 tries to reduce the residual error generated by M1 and M 3 will try to bring down the residual error generated by the M2 close to zero thereby giving the best accuracy.
Learning rate is not used as a hyper parameter in Bagging methods as the trees are independent of each other.	Learning rate is used as a hyper parameter in Boosting methods as each of the trees learns from the previous iterations.
Examples: Random Forest, Extra Tree algorithms.	Examples: Gradient Boosting, ADABoost, XGBoost.
Helps reduce variance.	Helps reduce both bias and variance.

DECISION TREE

- Easy to interpret (**non-parametric model**), powerful, versatile.
- Effective learning with small training data, however **very sensitive** to small variations in the training data.
- Doesn't require feature scaling.
- Very few assumptions about the training data(the data is non-linear).
- Very likely to **overfit the training data**(not generalizing well).
- Loves **orthogonal** decision boundaries.
- Decision Trees are **not sensitive to noisy data or outliers**.
- The decision tree algorithm is effective for balanced classification, although **it does not perform well on imbalanced datasets**.
- They are **unstable**, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.

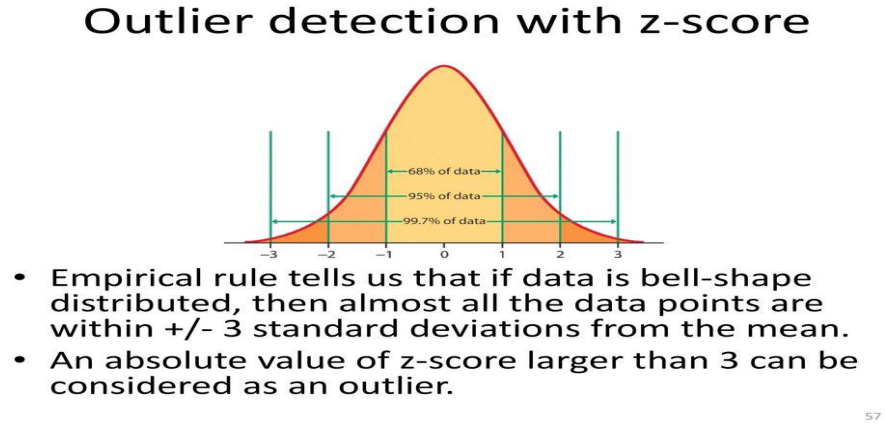
Which models are **sensitive to outliers**?

List of Machine Learning algorithms which are sensitive to outliers:

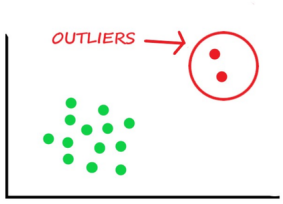
- **Linear Regression.**
- **Logistic Regression.**
- **Support Vector Machine.**
- **K- Nearest Neighbors.**
- **K-Means Clustering.**
- **Hierarchical Clustering.**
- **Principal Component Analysis.**

OUTLIERS

- An outlier is a **data point that is noticeably different from the rest**.
- The extreme values in the data are called **outliers**. The outliers are a part of the group but are far away from the other members of the group.
- The simplest way to detect an outlier is **by graphing the features or the data points**. **Visualization** is one of the best and easiest ways to have an inference about the overall data and the outliers. **Scatter plots and box plots** are the most preferred visualization tools to detect outliers.
- How do you find outliers in machine learning?



- Do outliers affect machine learning?



Machine learning algorithms are sensitive to the range and distribution of attribute values. **Data outliers can spoil and mislead the training process resulting in longer training times, less accurate models and ultimately poorer results**

SVM

- What is the **goal of SVM**?

The objective of applying SVMs is **to find the best line in two dimensions or the best hyperplane in more than two dimensions in order to help us separate our space into classes.** The hyperplane (line) is found through the **maximum margin**, i.e., the maximum distance between data points of both classes.

- What is **C and gamma** in SVM?

C is a hypermeter which is set **before the training model and used to control error**. **Gamma** is also a hypermeter which is se**t before the training model and used to give curvature weight of the decision boundary**.

- What is **Gamma**?

Gamma is used when we use the **Gaussian RBF kernel**. If you use a **linear or polynomial kernel then you do not need gamma only you need a C hypermeter**. Somewhere it is also used as **sigma**. Actually, sigma and gamma are related.

1. Gamma **high** means **more curvature**.
2. Gamma **low** means **less curvature**.

- What is **C**?

C adds **a penalty to each misclassified point**. If the C value is **small**, then essentially, the penalty for misclassified points is also small, thus resulting in a **high margin** based boundary. C value is **Large, soft margin**.

- What do you mean by **hard margin**?

A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing **overfitting**.

- What is a **soft margin**?

we still can try **to find a line to separate red and green dots, but we tolerate one or few misclassified dots** (e.g. the dots circled in red). This is called the **Soft Margin**. Or we can try **to find a non-linear decision boundary to separate red and green dots**. This is called the **Kernel Trick.**

1. Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting. Which of the following is the best option? Would you be more likely to consider iterating SVM next time? **you will try to calculate more variables**
2. How can **SVM be classified**? **It is a model trained using supervised learning. It can be used for classification and regression.**
3. Which of the following are **real world applications of the SVM**?
   text and hypertext categorization, image classification, clustering of news articles
4. Which of the following **evaluation metrics can not be applied in case of logistic regression** output to compare with target? **mean-squared-error**
5. Which of the following **evaluation metrics can be used to evaluate a model while modeling a continuous output variable**? **mean-squared-error**
6. The **cost parameter in the SVM** means: **the tradeoff between misclassification and simplicity of the model**
7. The **kernel trick**
   exploits the fact that in many learning algorithms, the weights can be written as a linear combination of input points
8. How does the bias-variance decomposition of a **ridge regression estimator compare with that of ordinary least squares regression**? **ridge has larger bias, smaller variance**
9. What is **kernel in SVM**? **1. Kernel functions map low dimensional data to high dimensional space 2. It's a similarity function**
10. You trained a **binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data**. Which is false.
    A. this is an instance of overfitting

    B. **this is an instance of underfitting** **False**

    C. the training was not well regularized

    D. the training and testing examples are sampled from different distributions

11. Suppose your model is demonstrating **high variance across the different training sets**. Which of the following is NOT a valid way to try and reduce the variance?

    A. increase the amount of training data in each training set

    B. **improve the optimization algorithm being used for error minimization. NOT VALID**

    C. decrease the model complexity

    D. reduce the noise in the training data

12. Suppose you are using the **RBF kernel in SVM with high Gamma value**. What does this signify? **1. the model would consider even far away points from hyperplane for modeling, 2. the model would consider only the points close to the hyperplane for modeling**
13. We usually use **feature normalization before using the Gaussian kernel in SVM**. What is true about feature normalization? **1. We do feature normalization so that new features will dominate others, 2. Some times, feature normalization is not feasible in case of categorical variables**
14. **Wrapper methods are hyper-parameter selection methods** that **are useful mainly when the learning machines are "black boxes"**
15. Which of the following methods **can not achieve zero training error on any linearly separable dataset**? **15-nearest neighbors**
16. Let S1 and S2 be the set of support vectors and w1 and w2 be the learnt weight vectors for a linearly separable problem using hard and soft margin linear SVMs respectively. Which of the following are correct? **s1 may not be a subset of s2**
17. Imagine, you are solving classification problems with highly imbalanced classes. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case? **1. Accuracy metric is not a good idea for imbalanced class problems. 2. Precision and recall metrics are good for imbalanced class problems.**

18. The **minimum time complexity for training an SVM is O(n2)**. According to this fact, what sizes of datasets are not best suited for SVM's? **large dataset**
19. **Perceptron Classifier** is **supervised learning algorithm**
20. The **SVMs are less effective** when **the data is noisy and contains overlapping points**
21. Which of the following are **components of generalization Error**? **bias, variance**
22. Which of the following methods is used for **multiclass classification**? **one vs rest**
23. Which one of the following is suitable? **1. When the hypothesis space is richer, overfitting is more likely. 2. when the feature space is larger , overfitting is more likely.**
24. Which of the following is categorical **data**? **branch of bank**
25. The **soft margin SVM is more preferred than the hard-margin SVM** when **the data is noisy and contains overlapping points**
26. In SVM, RBF kernels with appropriate parameters to perform binary classification where the data is **non-linearly separable**. In this scenario **the decision boundary in the transformed feature space in linear**
27.

MISSING VALUES

- What **percentage of missing values** is acceptable?

The overall percentage of data that is missing is important. Generally, **if less than 5% of values are missing then it is acceptable to ignore them (REF)**. However, the overall percentage missing alone is not enough; you also need to pay attention to which data is missing.

Type of handling missing values

1. **Deleting Rows**
2. **Replacing With Mean/Median/Mode**

This works better if the data is **linear**.

3. **Assigning An Unique Category**

A **categorical feature** will have a definite number of possibilities, such as gender.

4. **Predicting The Missing Values**

Using the **features** which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a **very high variance**.

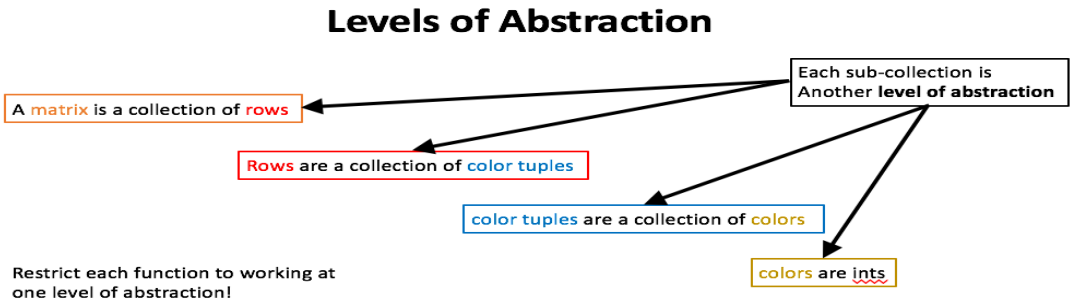
5. **Using Algorithms Which Support Missing Values**

- **KNN** is a machine learning algorithm which works on the **principle of distance measure**. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values.
- Another algorithm which can be used here is **RandomForest**. This model produces a robust result because **it works well on non-linear and categorical data**. It adapts to the data structure taking into consideration the **high variance or the bias, producing better results on large datasets**.
- How do you **handle missing data in data analysis**?

When dealing with missing data, data scientists can use two primary methods to solve the error: **imputation or the removal of data**. The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low.

IMAGE

- The **lowest level of abstraction** describes how a system actually stores data.
- What are the **4 layers of abstraction** in color image?



-

There are four levels of abstraction found in an **RGB numpy array**. These consist of a **matrix of rows (or columns), a row of RGB tuples, a RGB tuple of colors and the colors themselves**.

- **View Level** is the **highest level of abstraction**.
- 3 Levels of abstraction are : **Physical Level, Conceptual Level and View Level**

1. What function of **PIL.Image loads an image**? **Open**
2. How can you **load an image into an np.array**? **Casting the image to np.array**
3. When passing an image to an array is it important to specify the type? **Yes**
4. What is the **shape of a PNG**? **WxHx4**
5. when we work with images **we shape them so W == H**
6. What function can give us a hint as to **how to reshape the data** **sqrt()**
7. If we are using PIL what NEEDS to be the shape of the image in order to show it? **(28, 28)**
8. How can you make an **array into a PIL.Image** **using fromarray()**

Refer AUC-ROC Curve

- What is **Scale Invariance**?

A **system, function, or statistic** has scale invariance **if changing the scale by a certain amount does not change the system, function, or statistic's shape or properties**. **For example**, if you zoom in on a Koch snowflake, it looks the same.

Zooming in on a Koch snowflake.

- What does **well calibrated** mean?

Being well-calibrated means you are **usually right when you predict something will happen or say something is true**. **For example**, Superforecasters – people who are very good at predicting geopolitical events – are extremely well-calibrated.

Text Feature

Machine learning algorithms cannot work with raw text directly; **the text must be converted into numbers**. Specifically, vectors of numbers.This is called **feature extraction or feature encoding**.

- What is a **Bag-of-Words**?

A popular and simple method of feature extraction with text data is called the bag-of-words or BoW model of text. i.e. **considering each word count as a feature**. You can consider this method as a **N-Hot Encoding where we put a 1 if the word appears and 0 otherwise**.

	good	movie	not	a	did	like
good movie	1	1	0	0	0	0
not a good movie	1	1	1	1	0	0
did not like	0	0	1	0	1	1

In **sklearn** there is a similar method called **Count Vectorizer where we count the number of occurences of a word**.

Simple text cleaning techniques that can be used as a first step, such as:

- Ignoring case (Convert **all characters to lowercase** before tokenizer**)**
- Ignoring punctuation
- Remove too specific or rare words.
- Ignoring frequent words that don't contain much information, called **stop words**, like "a," "of," etc.
- Fixing misspelled words.

- Reducing words to their stem (e.g. “play” from “playing”) using **stemming and lemmatization algorithms**.

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Word	*Lemmatization*	*Stemming*
was	be	wa
studies	study	studi
studying	study	study

Stemming identifies the common root form of a word **by removing or replacing word suffixes** (e.g. “flooding” is stemmed as “flood”), while **lemmatization identifies the inflected forms of a word and returns its base form** (e.g. “better” is lemmatized as “good”).

N-Gram(used to find the relationship between the words)

Each word or token is called a **“gram”**. **An N-gram is an N-token sequence of words: a 2-gram (more commonly called a bigram) is a two-word sequence of words like “please turn”, “turn your”, or “your homework”, and a 3-gram (more commonly called a trigram) is a three-word sequence of words like “please turn your”, or “turn your homework”.**

Scoring methods

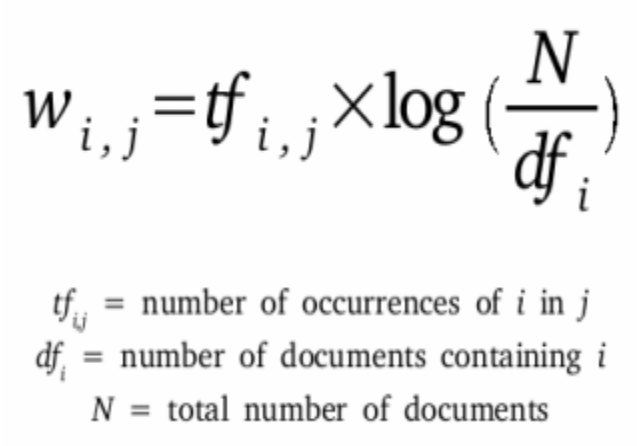
- **Counts**. Count the number of times each word appears in a document.
- **Frequencies**. Calculate the frequency that each word appears in a document out of all the words in the document.

TF-IDF**(we convert it into percentage with formula)**

A problem with scoring word frequency is that **highly frequent words start to dominate in the document** (e.g. larger score), but may not contain as much “informational content” to the model as rarer but perhaps domain specific words. One approach is to rescale the

frequency of words by how often they appear in all documents, so that the scores for frequent words like “the” that are also frequent across all documents are penalized.

This approach to scoring is called **Term Frequency – Inverse Document Frequency, or TF-IDF** for short, where:



- **Term Frequency**: is a scoring of the frequency of the word in the current document(**helps to normalize in frequency**).
- **Inverse Document Frequency**: is a scoring of how rare the word is across documents(**helps to remove most frequency** words).

The scores are a weighting where not all words are equally as important or interesting.

The scores have the effect of highlighting words that are distinct (contain useful information) in a given document.

Thus the idf of a rare term is high, whereas the idf of a frequent term is likely to be low.

- What are **N-Grams good for**? **Helps to use local context**
- How can we **remove too specific terms**? **By removing all the terms have a lower document frequency than X**
- How can we **remove common words** **By removing all the terms have a higher document frequency than X**
- Stem may not be an **actual word** whereas lemma is an **actual language word**.

Time Series

Definition of Time Series:

An **ordered sequence of values of a variable at equally spaced time intervals**.

- What is **time series forecasting**?

Time series forecast uses **historical data and patterns to predict new trends and future data behavior**. This method is used on cyclical data patterns.

- What is the **time series used for**?

Time series forecasting is used to **predict future behavior, trends, and patterns by analyzing a large amount of old data**.

- Does the **time interval need to be constant** to keep integrity? **Yes**
- What is a **time series** **a concurrent set f data points**
- Do we need to **treat all columns the same**? **No**
- In a time series … **Data is taken in groups o**
- What do **we use to make the predictions**? **The data we extrapolated**

Cross-Validation

- What does cross-validation mean in machine learning?

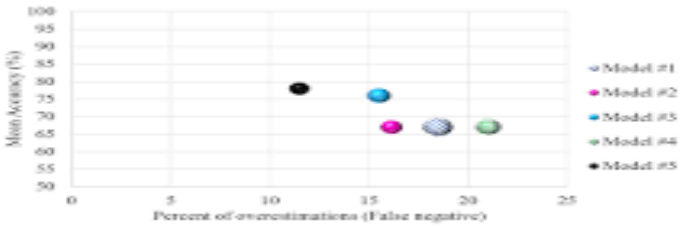
Cross-validation is **a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data**. Use cross-validation to detect overfitting, ie, failing to generalize a pattern.

Definition. Cross-Validation is **a statistical method of evaluating and comparing learning algorithms by dividing data into two segments**: one used to learn or train a model and the other used to validate the model.

- What is cross-validation and why is it necessary?

When you use cross validation in machine learning, you **verify how accurate your model is on multiple and different subsets of data**. Therefore, you ensure that it generalizes well to the data that you collect in the future. It improves the accuracy of the model.

- What is the **role of cross-validation**?



The goal of cross-validation is **to test the model's ability to predict new data that was not used in estimating it**, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

- What are the different types of cross-validation?

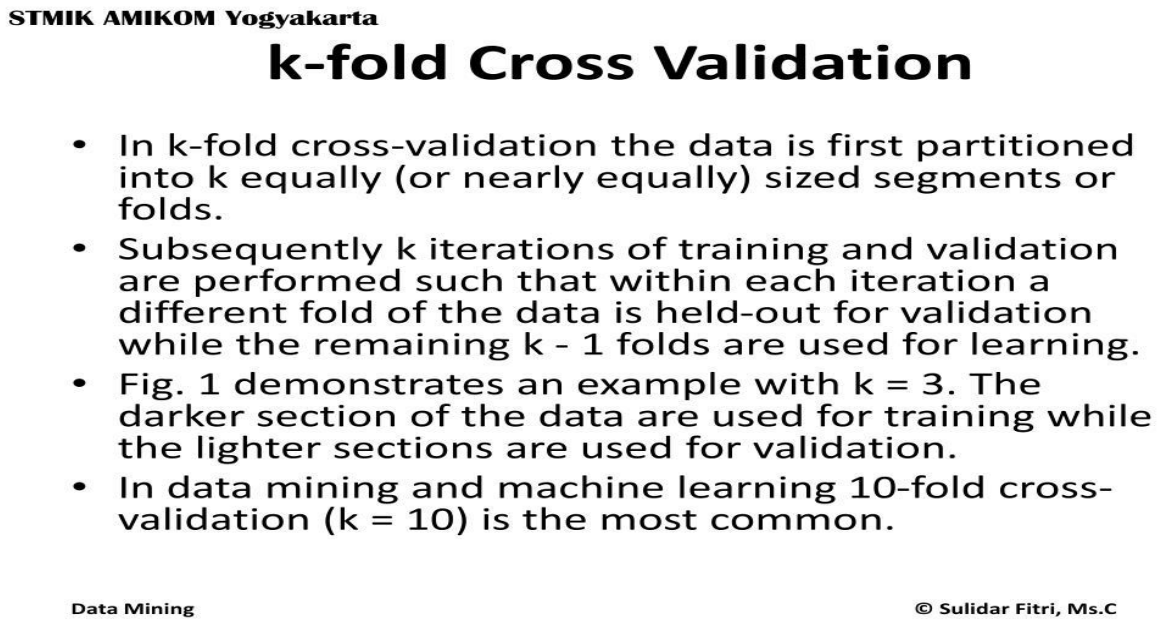
There are various types of cross-validation. However, mentioned above are the 7 most common types - **Holdout, K-fold, Stratified k-fold, Rolling, Monte Carlo, Leave-p-out, and Leave-one-out method**. Although each one of these types has some drawbacks, they aim to test the accuracy of a model as much as possible.

- Should I shuffle KFold?

The general procedure for cross - validation requires the dataset to be shuffled randomly. **If data is unordered in nature (i.e. non - Time series) then shuffle = True is the right choice**.

- Why is a CV stratified?

Implementing the concept of stratified sampling in cross-validation **ensures the training and test sets have the same proportion of the feature of interest as in the original dataset**. Doing this with the target variable ensures that the cross-validation result is a close approximation of generalization error.



The holdout method

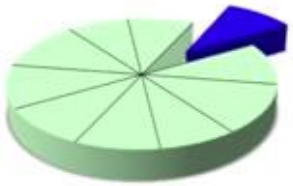
- The **holdout method** is the simplest kind of cross validation.
- The data set is separated into two sets, called the training set and the testing set.
- The function approximator fits a function using the training set only.
- Then the function approximator is asked to predict the output values for the data in the testing set (it has never seen these output values before).

Cross-Validation

- Cross-Validation (CV) is the standard Data Mining method for evaluating performance of classification algorithms. Mainly, to evaluate the Error Rate of a learning technique.
- In CV a dataset is partitioned in n folds, where each is used for testing and the remainder used for training. The procedure of testing and training is repeated n times so that each partition or fold is used once for testing.
- The standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use a stratified 10-fold cross-validation.
- Stratification implies making sure that when sampling is done each class is properly represented in both training and test datasets. This is achieved by randomly sampling the dataset when doing the n fold partitions.

Cross-validation

10-fold cross-validation



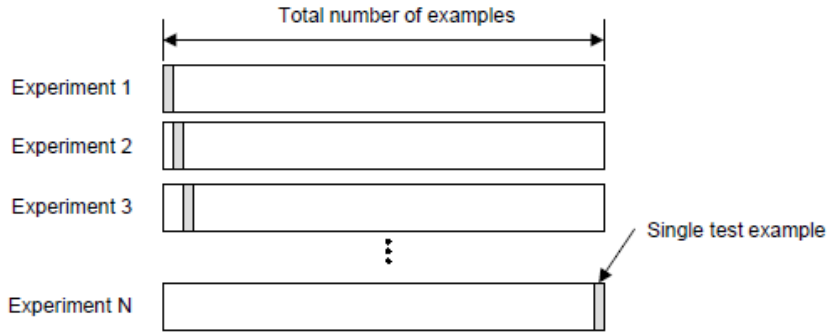
- Divide dataset into 10 parts (folds)
- Hold out each part in turn
- Average the results
- Each data point used once for testing, 9 times for training

Stratified cross-validation

- Ensure that each fold has the right proportion of each class value

Leave-one-out Cross Validation

- **Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples**
  - For a dataset with N examples, perform N experiments
  - For each experiment use N-1 examples for training and the remaining example for testing

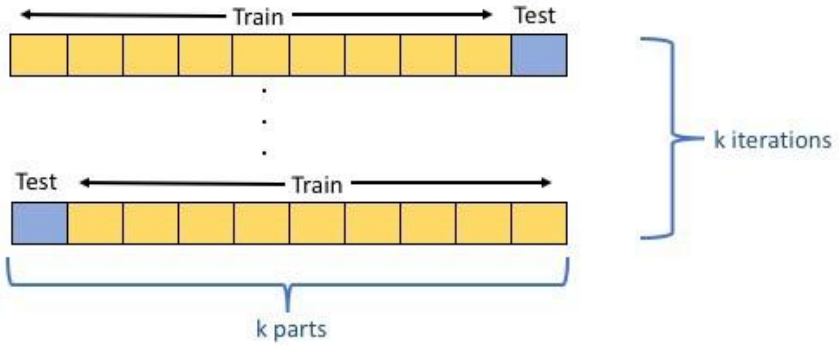


- **As usual, the true error is estimated as the average error rate on test examples**

$$E = \frac{1}{N}\sum_{i=1}^{N} E_i$$

K Folds Cross Validation Method

1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations



- Why do **we need to split the data**? To be able to test the model
- When building a Machine Learning model, we use both the training and test set to fit the model False
- What is the **target**? What we want to predict
- What are **the labels for a ML model**? The target column
- What are **the features in the process of building a ML model**? The data without the labels
- How do we usually indicate **the features**? with X
- How do we usually indicate **the target**? with y
- The difference between the **actual Y value and the predicted Y value** found using a **regression equation** is called the slope
- Which of the following methods/methods do we use to find the best fit line for data in **Linear Regression**? least square error

- Which of the following methods do we use to best fit the data in **Logistic Regression**? **maximum likelihood**
- In the regression equation Y = 75.65 + 0.50X, the **intercept** is 75.65 and **slope** 0.50
- Suppose, you get a situation where you find that your **linear regression model is under-fitting the data**. In such a situation which of the following options would you consider? **1. I will add more variables, 2. I will start introducing polynomial degree variables**

- We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in a training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, **What do you expect will happen with the mean training error**? **can't say**
- We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in a training set and test set randomly. **What do you expect will happen with bias and variance as you increase the size of training data**? **bias increases and variance decreases**
- In terms of bias and variance. Which of the following is true when you **fit a degree 2 polynomial**? **bias will be high, variance will be low**
- f X and Y in a regression model are **totally unrelated**, **the coefficient of determination would be 0**
- Problem in **multi regression** is ? **both multicollinearity & overfitting**

- Choose the correct statement with respect to **'confidence' metric in association rules**; it is **the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.**

- Which of the following sentences are correct in reference to **Information gain**?
    1. **It is biased towards multi-valued attributes, 2. ID3 makes use of information gain, 3. The approach used by ID3 is greedy**

- **Multivariate split** is where the partitioning of tuples is based on a combination of attributes rather than on a single attribute. **true**
- Gain ratio tends to prefer **unbalanced splits in which one partition is much smaller than the other** **true**

- The **gini index is not biased towards multivalued attributes**. **false**
- Gini index **does not favor equal sized partitions.** **false**
- When the **number of classes is large Gini index is not a good choice**. **True**
- This clustering approach initially assumes that each data instance represents a single cluster. **agglomerative clustering**
- Which statement is true about the **K-Means algorithm**? **all attributes must be numeric**

- The _____ step eliminates the extensions of (k-1)-itemsets which are not found to be frequent,from being considered for counting support. **pruning**

- What is the approach of a **basic algorithm for decision tree induction**? **Greedy**
- Which of the following classifications would best suit the **student performance classification systems**? **if...then... analysis**
- This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration **k-means clustering**

- A **good clustering method will produce high quality clusters** with **high intra class similarity**

- Which statement is true about **neural networks and linear regression models**? **both models require input attributes to be numeric**
- Which **Association Rule** would you prefer **low support and high confidence**

- The **apriori property** means **if a set cannot pass a test, its supersets will also fail the same test**

- **Clustering** is _____ and is example of _____learning descriptive and unsupervised
- **Association rules** from frequent item sets both minimum support and confidence are needed
- Which **Association Rule** would you prefer low support and high confidence
- **Classification rules** are extracted from decision tree
- How will you counter **overfitting in the decision tree**? by pruning the longer rules
- What are **two steps of tree pruning work**? post pruning and pre pruning

- Which of the following sentences are true?
    1. in pre-pruning a tree is \pruned\ by halting its construction early
    2. a pruning set of class labeled tuples is used to estimate cost complexity
    3. the best pruned tree is the one that minimizes the number of encoding bits

- Which of the following properties are **characteristic of decision trees**?
    1. High variance
    2. Lack of smoothness of prediction surfaces
    3. Unbounded parameter set

- Which Statement is not a true statement.
    1. k-means clustering is a linear clustering algorithm.
    2. k-means clustering aims to partition n observations into k clusters
    3. **k-nearest neighbor is same as k-means** True
    4. k-means is sensitive to outlier

- How to select the **best hyperparameters in tree based models**? measure performance over validation data

- What is true about **K-Mean Clustering**?
    1. K-means is extremely sensitive to cluster center initializations
    2. Bad initialization can lead to Poor convergence speed
    3. Bad initialization can lead to bad overall clustering

- What is **gini index**?
    4. gini index??operates on the categorical target variables
    5. B. it is a measure of purity
    6. C. gini index performs only binary split

- **Tree/Rule based classification algorithms** generate ... rule to perform the classification. If-then.

- **Decision Tree** is
    1. flow-chart
    2. structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label
- Which of the following is true about **Manhattan distance**? it can be used for continuous variables

- Which of the following can act as **possible termination conditions in K-Means**?

1. For a fixed number of iterations.
2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
3. Centroids do not change between successive iterations.
4. Terminate when RSS falls below a threshold.

- Which of the following statements is true about the **k-NN algorithm**?

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

- The K-means algorithm:

  minimizes the within class variance for a given number of clusters

- Which of the following metrics do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. Single-link
2. Complete-link
3. Average-link

- **High entropy** means that the partitions in classification are **Not pure**

Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.
It is a measure of disorder or purity or unpredictability or uncertainty.
Low entropy means less uncertain and high entropy means more uncertain.

- A machine learning problem involves four attributes plus a class. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?

Maximum possible different examples are the products of the possible values of each attribute and the number of classes;

3 * 2 * 2 * 2 * 3 = 72

- Suppose we would like to perform **clustering on spatial data such as the geometrical locations of houses**. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate? **Density-based clustering**

  The density-based clustering methods recognize clusters based on density function distribution of the data object. For clusters with arbitrary shapes, these algorithms connect regions with sufficiently high densities into clusters.

-

Naïve Bayes

The Naïve Bayes classifier assumes **conditional independence** between attributes and assigns the MAP class to new instances.

Naive Bayes is a classification algorithm for **binary (two-class) and multi-class classification problems**. The technique is easiest to understand when described using binary or categorical input values.

It is called naive Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed **to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on**.

Attributes are statistically independent of one another given the class value.