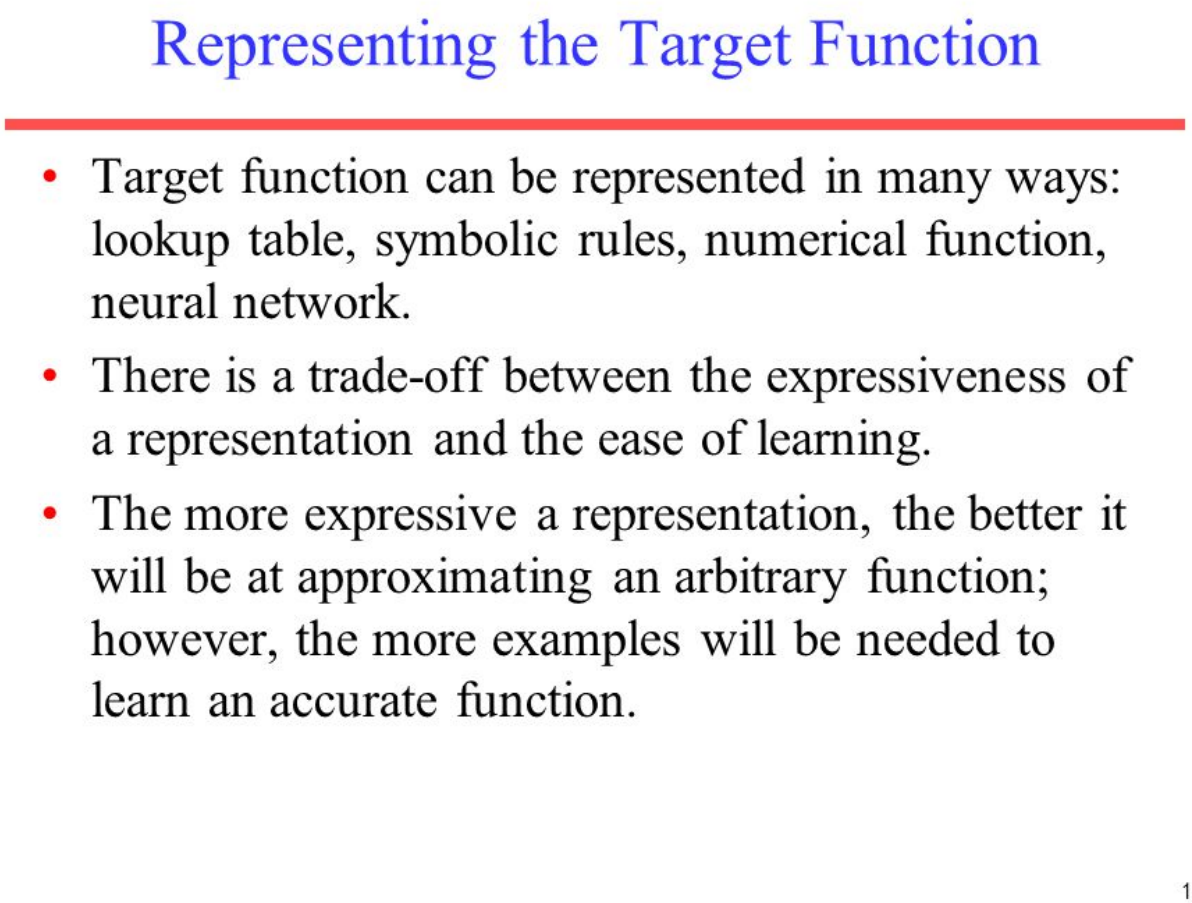
Assignment 07 Solutions

1. What is the definition of a target function ? In the sense of a real-life example, express the target function. How is a target function's fitness assessed ?

Ans: A target function, in machine learning, is a method for solving a problem that an AI algorithm parses its training data to find.The target function is essentially the formula that an algorithm feeds data to in order to calculate predictions.

Analyzing the massive amounts of data related to its given problem, an AI derives understanding of previously unspecified rules by detecting consistencies in the data. The observations of inherent rules about how the studied subject operates inform the AI on how to process future data that does not include an output by applying this previously unknown function.

"



"

2. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models ?

Ans: In short, predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data.

It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

The three main types of descriptive studies are **Case studies, Naturalistic observation, and Surveys**.

Some examples of descriptive research are: A specialty food group launching a new range of barbecue rubs would like to understand what flavors of rubs are favored by different people.

Case Studies are a type of observational research that involve a thorough descriptive analysis of a single individual, group, or event. There is no single way to conduct a case study so researchers use a range of methods from unstructured interviewing to direct observation.

"

Research Method	Advantages	Limitations
Naturalistic Observation	•More accurate than reports after the fact •Behavior is more natural	•Observer can alter behavior •Observer Bias •Not generalizable
Case Studies	•Depth •Takes advantage of circumstances that can not be replicated	•Not generalizable •Time consuming and expensive •Observational Bias
Surveys	• Immense amount of data •Quick and inexpensive •Generalizable •Replicable	• Poor sampling can skew results •Wording Effect •Social Desirability Bias

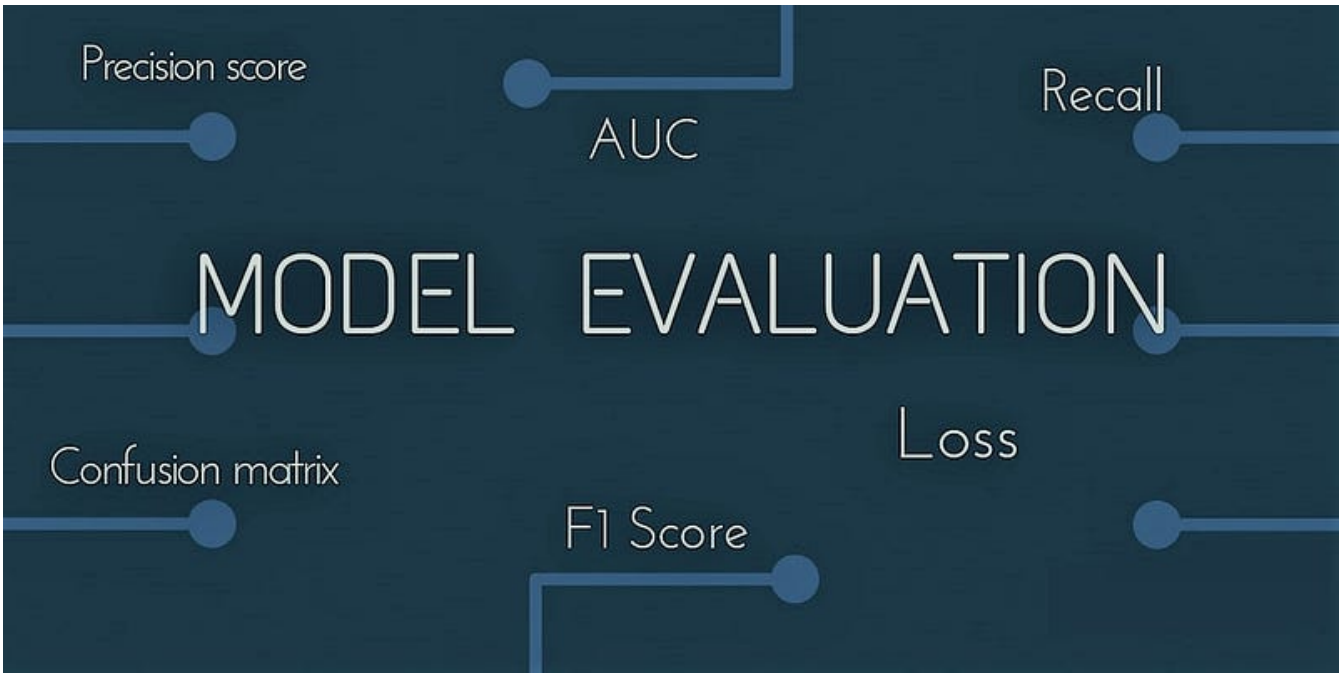
"

3. Describe the method of assessing a classification model's efficiency in detail. Describe the various measurement parameters ?

Ans: Logarithmic loss (or log loss) measures the performance of a classification model where the prediction is a probability value between 0 and 1.

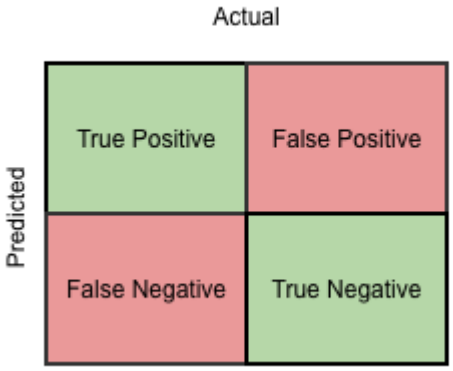
Log loss increases as the predicted probability diverge from the actual label. Log loss is a widely used metric for Kaggle competitions. Input on the most important basics for the measurement of the physical parameters: Temperature, flow velocity, humidity, pressure, CO2 and infrared. Tips on correct measurement and for avoiding measurement errors.

Other Model Evaluation Metrics are mentioned below: "



"

"



Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precission = TP/(TP+FP)

Recall = TP/(TP+FN)

True Positive Rate = TP/(TP+FN)

False Positive Rate = FP/(FP+TN)

"

4. Describe :

1. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting ?
2. What does it mean to overfit? When is it going to happen?
3. In the sense of model fitting, explain the bias-variance trade-off.

Ans: The following is the short notes on:

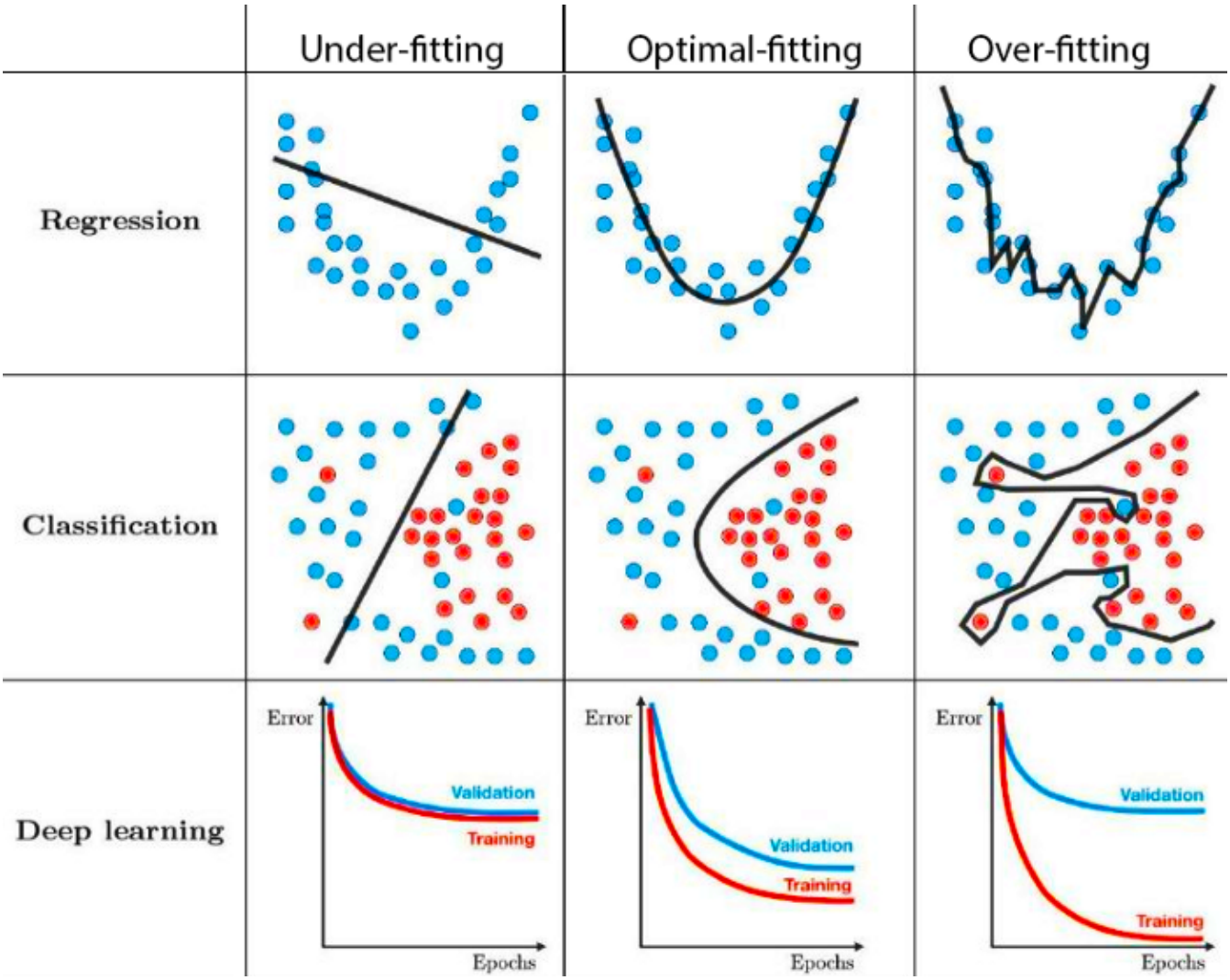
In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting:

Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data.

What does it mean to overfit? When is it going to happen

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.

"

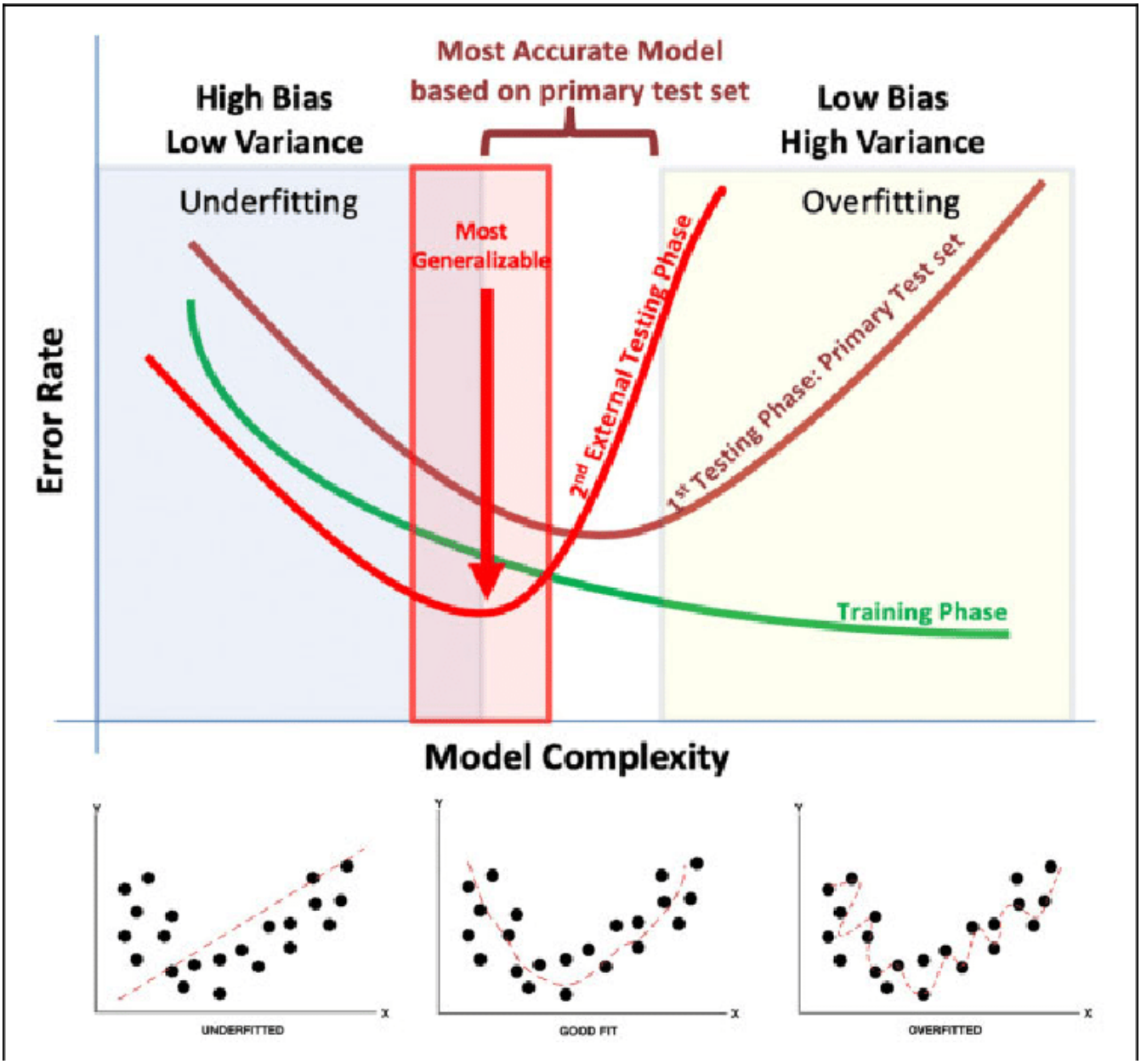


"

In the sense of model fitting, explain the bias-variance trade-off

The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set.

"



5. Is it possible to boost the efficiency of a learning model? If so, please clarify how ?

Ans: Building a machine learning model is not enough to get the right predictions, as you have to check the accuracy and need to validate the same to ensure get the precise results. And validating the model will improve the performance of the ML model. Some ways of boosting the efficiency of a learning model are mentioned below:

1. Add more Data Samples
2. Look at the problem differently: Looking at the problem from a new perspective can add valuable information to your model and help you uncover hidden relationships between the story variables. Asking different questions may lead to better results and, eventually, better accuracy.
3. Adding Context to Data: More context can always lead to a better understanding of the problem and, eventually, better performance of the model. Imagine we are selling a car, a BMW. That alone doesn't give us much information about the car. But, if we add the color, model and distance traveled, then you'll start to have a better picture of the car and its possible value.
4. Finetuning our hyperparameter: to get the answer, we will need to do some trial and error until you reach your answer.
5. Train our model using cross-validation

6. Expoerimenting with different Algorithms.

6. How would you rate an unsupervised learning model's success? What are the most common success indicators for an unsupervised learning model ?

Ans: In case of supervised learning, it is mostly done by measuring the performance metrics such as accuracy, precision, recall, AUC, etc. on the training set and the holdout sets whereas for Unsupervised Learning it is different. Since there is no pre-evidence or records for patterns, we cannot directly compute the accuracy by comparing actual and predicted outputs but there exist many evaluation metrics to measure the performance of unsupervised learning algorithms after the training process.

Some of them are-

Clustering - Jaccard similarity index, Rand Index, Purity, Silhouette measure, Sum of squared errors, etc.

Association rule mining – Lift, Confidence

Time series analysis – Root mean square error, mean absolute error, mean absolute percentage error, etc.

Autoencoders - Reconstruction errors

Natural Language processing (like sentiment analysis and text clustering) – Comparing the correlation between natural words after converting them to numerical vectors.

Principal component analysis – Reconstruction error, Scree plot

Generative adversarial networks – Discriminator functions

Recurrent neural networks and LSTM (In numerical series) – Root mean square error, mean absolute error, mean absolute percentage error, etc.

Recurrent neural networks and LSTM (In semantic series) - Word to vector correlation

Anomaly detection (like DBSCAN, OPTICS) – Cohesion, Separation, Sum of squared errors, etc.

Expectation/ Maximization problems – Log-likelihood

Survival analysis (Cox model 1) – Simple hazard ration, R Squared

Survival analysis (Cox model 2) – Two group hazard ratio and brier score, Log-rank test, Somers' rank correlation, Time-dependent ROC – AUC, Power validation, etc.

Few other examples of such measures are:

- Silhouette coefficient.
- Calisnki-Harabasz coefficient.
- Dunn index.
- Xie-Beni score.
- Hartigan index.

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer ?

Ans: Categorical Data is the data that generally takes a limited number of possible values. Also, the data in the category need not be numerical, it can be textual in nature. All machine learning models are some kind of mathematical model that need numbers to work with. This is one of the primary reasons we need to pre-process the categorical data before we can feed it to machine learning models.

If a categorical target variable needs to be encoded for a classification predictive modeling problem, then the LabelEncoder class can be used.

8. Describe the predictive modeling method for numerical values. What distinguishes it from categorical predictive modeling ?

Ans: predictive modeling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

Classification is the process of identifying the category or class label of the new observation to which it belongs.Predication is the process of identifying the missing or unavailable numerical data for a new observation. That is the key difference between classification and prediction.

9. Make quick notes on:

1. The process of holding out
2. Cross-validation by tenfold
3. Adjusting the parameters

Ans: The Quick notes on the following topics is below:

- **The process of holding out:**
  The hold-out method for training machine learning model is the process of splitting the data in different splits and using one split for training the model and other splits for validating and testing the models. The hold-out method is used for both model evaluation and model selection.
- **Cross-validation by tenfold:** 10-fold cross validation would perform the fitting procedure a total of ten times, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold out set for validation.
- **Adjusting the parameters:**
  A fancy name for training: the selection of parameter values, which are optimal in some desired sense (eg. minimize an objective function you choose over a dataset you choose). The parameters are the weights and biases of the network

10. Define the following terms:

1. Purity vs. Silhouette width

2. Boosting vs. Bagging
3. The eager learner vs. the lazy learner

Ans: The Following is the short notes on:

- **Purity vs Silhouette width:**
  - Purity is a measure of the extent to which clusters contain a single class. Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster.
  - The silhouette width is also an estimate of the average distance between clusters. Its value is comprised between 1 and -1 with a value of 1 indicating a very good cluster.

- **Boosting vs. Bagging:**
  - Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data.
  - Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

- **The eager learner vs. the lazy learner:**
  - A lazy learner delays abstracting from the data until it is asked to make a prediction.
  - while an eager learner abstracts away from the data during training and uses this abstraction to make predictions rather than directly compare queries with instances in the dataset.