**Hamza Farooq 2016-CS-122**

**Machine Learning Quiz**

**Video 1**

**Question no 1: What is Machine Learning?**

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

**Question no 2: What is difference between Training set and training instance?**

The examples that the system uses to learn are called the training set. Each training example is called a training instance or sample.

**Question no 3: How does we make our Machine Learning Algorithm to perform better.**

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

* Experience E is the training data
* Particular performance measure is called Accuracy

We can make our ML algorithm perform better by

* Increasing large amount of training data
* Providing high quality training data

**Video 1.1**

**Question no 1: Difference between Instance based learning vs Model based learning?**

Instance based learning: Instead of performing explicit generalization, ML algorithm compares new problem instances with instances seen in training, which have been stored in memory.  
E.g. k-nearest neighbors

Model based learning: In this technique, a generalization is made from training data all-at-once in from of model that predicts the incoming values.  
E.g. Decision Tree, SVM and Regression problems

**Question no 2: What is supervised and unsupervised learning?**

Supervised Learning: Labels, Annotations or additional information are given with the data.  
E.g. Classification, regression problems

Unsupervised Learning: A type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.  
E.g. Cluster analysis method

**Question no 3: Difference between online and offline learning?**

Online Learning: Learning on-the-fly based on each pattern as it is observed.

Batch or offline Learning: Learning over groups of patterns or training data. Most algorithms are batch

**Video 1.2**

**Question no 1: write name of different algorithm used in supervised learning.**

* k-Nearest Neighbors
* Linear Regression
* Logistic Regression
* Support Vector Machines
* Decision Trees and Random Forests
* Neural Networks

**Question no 2: Is supervised learning only used in classification problems?**

No, there are several other algorithms that implement supervised learning such as

* Clustering
* Regression Problems

**Video 1.3**

**Question no 1: What is Clustering?**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.  
E.g. Cluster of students with respect to their hometown location.

**Question no 2: What do you know about Visualization dimensionality reduction?**

Dimensionality reduction: The process of reducing the number of parameters under consideration by removing least important parameters from the data. It reduces the time and storage space required.

Visualization: A technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data.

**Video 1.4**

**Question no 1: What is semi supervised learning?**

Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.  
E.g. Restricted Boltzmann machine(RBM), Speech Analysis, Internet Content Classification

**Question no 2: Write any advantage of semi supervised learning.**

The number of unlabeled or partially labeled samples is often larger than the number of labeled samples, since the former are less expensive and easier to obtain

**Video 1.5**

**Question no 1: What is reinforcement learning?**

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize some notion of cumulative reward.  
E.g. A cat that is exposed to environment.

**Question no 2: Why we say that reinforcement learning is type of Supervised learning?**

In reinforcement learning, the parameters of learning are supervised by assigned to be higher in priority that contribute positively in learning. Important parameters contribute in score while other parameters contribute in penalty.

**Video 1.6**

**Question no 1: What do we mean by term Parameter and feature in the Machine Learning Algorithm?**

Attributes: The characteristics of input data that are defined in problem domain.

Parameters: The variables that ML algorithm is trying to tune to build an accurate model. Several parameters may form an individual attribute.

**Question no 2: What is Batch learning?**

Batch Learning: First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned.

**Video 1.7**

**Question no 1: What is the term ‘Memorize’ mean? keeping in view concept of Machine Learning.**

Memorize: The system learns the examples by heart, then predict to new cases using a similarity measure.

**Question no 2: Give an example of instance-based learning.**

An email spam filter in which incoming emails are classified as spam or not spam on the basis of similarity measure.

**Video 1.8**

**Question no 2: What do you think which of the step is most important MLDLC and why?**

Model Selection is the most important step in MLDLC. Once model is selected, it costs too much to replace the model in future because we have to restart many of our work.

**Question no 3: What do you mean by data preprocessing?**

Data preprocessing or Data cleaning: Most Machine Learning algorithms cannot work with

* missing features
* outliers
* redundant attributes

Data preprocessing include

* Set the values to some value (zero, the mean, the median, etc.).
* Get rid of the corresponding districts.
* Get rid of the whole attribute.

**Video 1.9**

**Question no 1: Why model Selection is important step?**

Once model is selected, it costs too much to replace the model in future because we have to restart many of our work such as

* Preprocess the data for corresponding model once again
* Train the newly selected model

**Question no 2: What is neural network?**

Neural networks are a set of units, modeled loosely after the human brain, that are designed to recognize patterns. Every unit contains

* Weights
* Transfer function
* Activation function

The weights are learned by passing and updating the whole dataset several times

**Question no 3: Write different models of ML**

* Support Vector Machines (SVMs)
* Decision Trees and Random Forests
* Neural networks

**Video 2**

**Question no 1: Difference between Traditional Programing and Machine Learning?**

Traditional Programming: To write a computer program in which rules are explicitly defined according to human expert knowledge. Program will likely to become a long list of complex rules and pretty hard to maintain.

Machine Learning: Problems for which existing solutions require a lot of hand-tuning or long lists of rules, one machine learning algorithm can often simplify code and perform better. A machine learning system can adapt to new data in fluctuation environment.

**Video 3**

**Question no 1: what three things we first look for in solving a computer science problem?**

Whenever we go to solve a real world problem using **Traditional Approach**, we have

* Input
  + Study the problem
* Processing
  + Write rules
  + Evaluate
  + Analyze Errors
* Output
  + Launch if no errors

**Question no 2: Can our rules which we describe initially in the program to solve the problem automated in Machine learning? Explain**

Yes, using large and high-quality data, we train the ML classifier to write the rules on its own to solve the problem and no doubt the rules written by machine will be more appropriate.

**Video 4**

**Question no 1: write difference applications of machine learning.**

* Virtual Personal Assistants (Siri, Alexa, Google)
* Video Surveillance (ASSVS)
* Social Media Services
* Email Spam and Malware Filtering
* Product Recommendations (Amazon, Alibaba)
* Online Fraud Detection (Paypal using ML for protection against money laundering)

**Question no 2: Write difference between ML, Data sciences and AI**

Machine Learning: Machine learning is the ability of a computer system to learn from the environment and improve itself from experience without the need for any explicit programming.

Data Science: Data science is the extraction of relevant insights from many structural and unstructured data.

Artificial Intelligence: Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions

**Video 5:**

**Question no 1: What do you mean by deployment in Machine learning?**

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data.

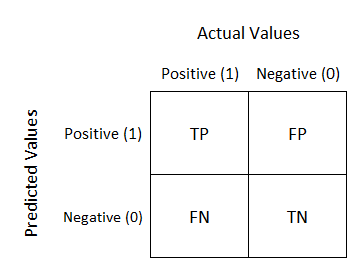
**Question no 2: What is validation data and test set?**

Validation data : Validation data is subset if test set data used to give an estimate of model skill while tuning model’s hyperparameters.

Test Set: A set of examples used only to assess the performance of a fully-specified classifier. It is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

**Question no 3: What is confusion matrix? Explain in term of true positive, false positive. ( You can get help from google)**

Confusion Matrix: A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known.



**Video 6**

**Question no 1: How we perform error analysis in ML?**

1. Check
   1. Failure points
   2. Data leading to wrong predictions
   3. Model testing with noisy data
2. Construct confusion matrix
3. Check if the model
   1. Underfit (Low quantity of data)
   2. Overfit (Repetition in data)

**Question no 2: What are the factors through which error occur? Explain what is noisy data?**

Error occur due to

* Failure points
* Data leading to wrong predictions
* Model testing with noisy data

Noisy data: Data containing too much unwanted information and inaccurate desired data definitions.

**Question no 3: What is overfit model and what is underfit model?**

Overfit model: Classifer perform well on training data but not on testing data. Generally occur due to repetition in training data.

Underfit model: Classifer doesn’t perform well on training data and testing data as well. Generally occur due to low quantity of training data.

**Video 7**

**Question no 1: What do the term “Three Folds” mean regarding dataset?**

Three folds mean three phases given below

1. Training Set: Data used to train the classifier
2. Validation Set: Data used to measure how well the classifier has learned unseen data.
   1. Used during training of the data for tuning the parameters
   2. Validation and test sets are to direct you towards most important changes to your machine learning model.
3. Test Set: Data used to measure the classification accuracy of trained model.

**Video 8**

**Question no 1: Describe different challenges machine learning application has to face? Explain each challenge in one line.**

Machine Learning application challenges are given below

1. Insufficient quantity of training data set
2. Nonrepresentative training data. Data not related to problem domain
3. Poor-Quality data, full of errors, noise and outliers
4. Irrelevant features. Garbage in, Garbage out.
5. Overfitting the training data. Repetition in training data.
6. Underfitting the training data. Not enough amount of training data
7. Stepping back:
   1. Learning from data instead of explicit rules
   2. Too many different types of ML systems:
      1. Supervised / Unsupervised
      2. Batch / Online
      3. Instance based / Model based

**Question no 2: What are outliers?**

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error.

**Video 9**

**Question no 1: Why Dataset is important in building any machine learning model?**

Dataset is important in various aspects such as

* Data is used to train the model that will solve the problem specified tasks
* Data is used to memorize the information for future prediction.
* Data is the key input to ML algorithm without which machine learning is just a hollow structure

**Question no 2: What are 8 steps in building a machine learning model?**

1. Frame the problem and look at the big picture
2. Get the data
3. Explore the data to gain insights
4. Prepare the data to better expose the underlying data patterns to machine learning algorithms
5. Explore many different models and short-list

**Question no 3: what is fine tune? How it done and what are hyperparameter?**

Finetuning means taking weights of a trained neural network and use it as initialization for a new model being trained on data from the same domain (often e.g. images). It is used to:

* speed up the training
* overcome small dataset size

Hyperparameters: A parameter whose value is set before the learning process begins. By contrast, the values of other parameters are derived via training.

**Video 10 (Both Frame and data acquisition)**

**Question no 1: What Is the meaning of the phrase “Frame the problem”**

Frame the Problem include following:

1. Define the objective in business terms.

2. How will your solution be used?

3. What are the current solutions/workarounds (if any)?

4. How should you frame this problem (supervised/unsupervised, online/offline,

etc.)?

5. How should performance be measured?

6. Is the performance measure aligned with the business objective?  
7. What would be the minimum performance needed to reach the business objective?  
8. What are comparable problems? Can you reuse experience or tools?  
9. Is human expertise available?  
10. How would you solve the problem manually?  
11. List the assumptions you (or others) have made so far.  
12. Verify assumptions if possible.

**Question no 2: What is data acquisition?**

1. List the data you need and how much you need.

2. Find and document where you can get that data.

3. Check how much space it will take.

4. Check legal obligations, and get authorization if necessary.

5. Get access authorizations.

6. Create a workspace (with enough storage space).

7. Get the data.

8. Convert the data to a format you can easily manipulate (without changing the

data itself).

9. Ensure sensitive information is deleted or protected (e.g., anonymized).

10. Check the size and type of data (time series, sample, geographical, etc.).

11. Sample a test set, put it aside, and never look at it (no data snooping!).

**Video 11 (Both Data-pipeline and Preprocessing)**

**Question no 1: What is data pipeline?**

A sequence of data processing components is called a data pipeline. Pipelines are very common in Machine Learning systems, since there is a lot of data to manipulate and many data transformations to apply.

**Question no 2: What is regression**

A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. Two major types of regression are

* Linear regression
* Logistic regression

**Question no 3: Explain data preprocessing**

Data preprocessing or Data cleaning: Most Machine Learning algorithms cannot work with

* missing features
* outliers
* redundant attributes

Data preprocessing include

* Set the values to some value (zero, the mean, the median, etc.).
* Get rid of the corresponding districts.
* Get rid the whole attribute.

**Video 12**

**Question no 1: How can we create training and testing dataset (Through code)**

**Explain each line of code.**

**import numpy as np**

**def** split\_train\_test(data, test\_ratio):

shuffled\_indices = np.random.permutation(len(data))

test\_set\_size = int(len(data) \* test\_ratio)

test\_indices = shuffled\_indices[:test\_set\_size]

train\_indices = shuffled\_indices[test\_set\_size:]

**return** data.iloc[train\_indices], data.iloc[test\_indices]

1. Import numeric array operation library
2. Define function to split data set
   1. Data: data set itself
   2. Test\_ratio: training set to test set ratio
3. Get an array of randomly shuffled indices
4. Set size of test \_set array depending upon test\_ratio
5. Separate indices of test\_set from shuffled\_indices
6. Separate indices of train\_set from shuffled\_indices
7. Return a tuple of instances of training set and test set by aiding train\_indices and test\_indices respectively.

**Question no 2: How can we do sampling?**

Two sampling methods are used

1. Purely random sampling
   * Random sample is a subset of a statistical population in which each number of the subset has an equal probability of being chosen
   * Random samples involve the random selection of data from the entire population so each possible sample is equally likely to occur
2. Stratified sampling
   * The population is divided into homogeneous subgroups called strata, and the right number of instances is sampled form each stratum to guarantee that the test set is representative of the overall population
   * Random samples are then selected from each stratum