Introduction to Machine Learning

Day 0

1. Explain the difference between supervised and unsupervised machine learning?

   In supervised machine learning algorithms, we have to provide labeled data, for example, prediction of stock market prices, whereas in unsupervised we need not have labeled data, for example, classification of emails into spam and non-spam.

2. What Are the Applications of Supervised Machine Learning in Modern Businesses?

   Applications of supervised machine learning include:

   - **Email Spam Detection**
     Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.
   - **Healthcare Diagnosis**
     By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.
   - **Sentiment Analysis**
     This refers to the process of using algorithms to mine documents and determine whether they're positive, neutral, or negative in sentiment.
   - **Fraud Detection**
     Training the model to identify suspicious patterns, we can detect instances of possible fraud.

3. What Is Semi-supervised Machine Learning?

   Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data. In the case of semi-supervised learning, the training data contains a small amount of labeled data and a large amount of unlabeled data.

4. What are the different Algorithm techniques in Machine Learning?

   The different types of techniques in Machine Learning are:

   - Supervised Learning
   - Unsupervised Learning
   - Semi-supervised Learning
   - Reinforcement Learning
   - Transduction
   - Learning to Learn

5. What Are Unsupervised Machine Learning Techniques?

   There are two techniques used in unsupervised learning: clustering and association.

   - Clustering
     - Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.
   - Association
     - In an association problem, we identify patterns of associations between different variables or items.
     - For example, an eCommerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

Day 1

1. What is Jupyter Notebook?

   The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

2. Write down the code for displaying $\sin x$ between $[0, 3\pi]$

```
import numpy as np
import matplotlib.pyplot as plt

# Compute the x and y coordinates for points on a sine curve
x = np.arange(0, 3 * np.pi, 0.1)
y = np.sin(x)
plt.title("sine wave form")

# Plot the points using matplotlib
plt.plot(x, y)
plt.show()
```

3. Explain NumPy broadcasting.

   Broadcasting is the methodology adopted in NumPy used to perform arithmetic operations on arrays with differing dimensions. General arithmetic operations such as addition, multiplication, subtraction, etc. tend to broadcast arrays before performing the operations on arrays with variations in size.

4. What method would you use for creating a 3x5x6 random array of doubles?

```
numpy.random.randn(3,5,6)
```

5. How would you create a 3x6 matrix of zeros and ones?

```
numpy.zeros((3, 6), dtype=int)
numpy.ones((3, 6), dtype=int)
```

Day 2

1. What are the parametric models? Give an example.

   Parametric models are those with a finite number of parameters. To predict new data, you only need to know the parameters of the model. Examples include linear regression, logistic regression, and linear SVMs.

   Non-parametric models are those with an unbounded number of parameters, allowing for more flexibility. To predict new data, you need to know the parameters of the model and the state of the data that has been observed. Examples include decision trees, k-nearest neighbors, and topic models using latent Dirichlet analysis.

2. What is the difference between classification and regression?

   Classification is used to produce discrete results, classification is used to classify data into some specific categories. For example, classifying emails into spam and non-spam categories. Whereas, We use regression analysis when we are dealing with continuous data, for example predicting stock prices at a certain point in time.

3. What is Linear Regression?

   Linear Regression is a supervised Machine Learning algorithm. It is used to find the linear relationship between the dependent and the independent variables for predictive analysis.

4. What are the drawbacks of a linear model?

   There are a couple of drawbacks of a linear model:

   - A linear model holds some strong assumptions that may not be true in the application. It assumes a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity
   - A linear model can't be used for discrete or binary outcomes.
   - You can't vary the model flexibility of a linear model.

5. When does the linear regression line stop rotating or finds an optimal spot where it is fitted on data?

   A place where the highest RSquared value is found, is the place where the line comes to rest. RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

Day 3

1. What Is Overfitting, and How Can You Avoid It?

   Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data. When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting. There are multiple ways of avoiding overfitting, such as:

   - Regularization. It involves a cost term for the features involved with the objective function
   - Making a simple model. With lesser variables and parameters, the variance can be reduced
   - Cross-validation methods like k-folds can also be used
   - If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

2. When does regularization becomes necessary in Machine Learning?

   Regularization becomes necessary when the model begins to overfit/underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce the cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

3. What does the term decision boundary mean?

   A decision boundary or a decision surface is a hypersurface which divides the underlying feature space into two subspaces, one for each class. If the decision boundary is a hyperplane, then the classes are linearly separable.

4. How can learning curves help create a better model?

   Learning curves give the indication of the presence of overfitting or underfitting.

   In a learning curve, the training error and cross-validating error are plotted against the number of training data points.

5. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

   Both algorithms are methods for finding a set of parameters that minimize a loss function by evaluating parameters against data and then making adjustments.

   In standard gradient descent, you'll evaluate all training samples for each set of parameters. This is akin to taking big, slow steps toward the solution.

   In stochastic gradient descent, you'll evaluate only 1 training sample for the set of parameters before updating them. This is akin to taking small, quick steps toward the solution.

Day 4

1. What Is 'naive' in the Naive Bayes Classifier?

   The classifier is called 'naive' because it makes assumptions that may or may not turn out to be correct.

   The algorithm assumes that the presence of one feature of a class is not related to the presence of any other feature (absolute independence of features), given the class variable.

   For instance, a fruit may be considered to be a cherry if it is red in color and round in shape, regardless of other features. This assumption may or may not be right (as an apple also matches the description).

2. What Is Decision Tree Classification?

   A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

3. What Is Pruning in Decision Trees, and How Is It Done?

   Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

4. In k-means or KNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance?

   We don't use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, the euclidean metric can be used in any space to calculate distance. Since the data points can be present in any dimension, euclidean distance is a more viable option.

5. What is a Decision Tree?

   A decision tree is used to explain the sequence of actions that must be performed to get the desired output. It is a hierarchical diagram that shows the actions.

Day 5

1. Define Precision and Recall.

   Precision

   - Precision is the ratio of several events you can correctly recall to the total number of events you recall (mix of correct and wrong recalls).
   - Precision = (True Positive) / (True Positive + False Positive)

   Recall

   - A recall is the ratio of a number of events you can recall the number of total events.
   - Recall = (True Positive) / (True Positive + False Negative)

2. What is meant by 'Training set' and 'Test Set'?

   We split the given data set into two different sections namely,' Training set' and 'Test Set'. 'Training set' is the portion of the dataset used to train the model.

   'Testing set' is the portion of the dataset used to test the trained model.

3. Explain the Bias-Variance Tradeoff.

   Predictive models have a tradeoff between bias (how well the model fits the data) and variance (how much the model changes based on changes in the inputs).

   Simpler models are stable (low variance) but they don't get close to the truth (high bias).

   More complex models are more prone to overfitting (high variance) but they are expressive enough to get close to the truth (low bias). The best model for a given problem usually lies somewhere in the middle.

4. How much data should you allocate for your training, validation, and test sets?

   You have to find a balance, and there's no right answer for every problem.

   If your test set is too small, you'll have an unreliable estimation of model performance (performance statistic will have high variance). If your training set is too small, your actual model parameters will have a high variance.

   A good rule of thumb is to use an 80/20 train/test split. Then, your train set can be further split into train/validation or into partitions for cross-validation.

5. What Is a False Positive and False Negative and How Are They Significant?

   False positives are those cases which wrongly get classified as True but are False. False negatives are those cases which wrongly get classified as False but are True.

   In the term 'False Positive,' the word 'Positive' refers to the 'Yes' row of the predicted value in the confusion matrix. The complete term indicates that the system has predicted it as a positive, but the actual value is negative.

Day 6

1. What Is Kernel SVM?

   Kernel SVM is the abbreviated version of the kernel support vector machine. Kernel methods are a class of algorithms for pattern analysis, and the most common one is the kernel SVM.

2. What are support vector machines?

   Support vector machines are supervised learning algorithms used for classification and regression analysis.

3. When would you use random forests Vs SVM and why?

   There are a couple of reasons why a random forest is a better choice of the model than a support vector machine:

   - Random forests allow you to determine the feature importance. SVM's can't do this.
   - Random forests are much quicker and simpler to build than an SVM.
   - For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

4. What is a kernel? Explain the kernel trick

   A kernel is a way of computing the dot product of two vectors xx and yy in some (possibly very high dimensional) feature space, which is why kernel functions are sometimes called "generalized dot product"

   The kernel trick is a method of using a linear classifier to solve a non-linear problem by transforming linearly inseparable data to linearly separable ones in a higher dimension.

5. How does the SVM algorithm deal with self-learning?

   SVM has a learning rate and expansion rate which takes care of this. The learning rate compensates or penalises the hyperplanes for making all the wrong moves and expansion rate deals with finding the maximum separation area between classes.

Day 7

1. Explain Principle Component Analysis (PCA).

   PCA is a method for transforming features in a dataset by combining them into uncorrelated linear combinations.

   These new features, or principal components, sequentially maximize the variance represented (i.e. the first principal component has the most variance, the second principal component has the second most, and so on).

2. What Are Some Methods of Reducing Dimensionality?

   You can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.

   Now that you have gone through these machine learning interview questions, you must have got an idea of your strengths and weaknesses in this domain.

3. How is KNN different from k-means clustering?

   K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part). K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.

4. Why is the rotation of components so important in Principle Component Analysis(PCA)?

   Rotation in PCA is very important as it maximizes the separation within the variance obtained by all the components because of which interpretation of components would become easier. If the components are not rotated, then we need extended components to describe the variance of the components.

5. What are PCA, KPCA, and ICA used for?

   PCA (Principal Components Analysis), KPCA (Kernel-based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.