ML Q & A interview prep:

ML Algorithms + Basic ML

Q: What are the various types of ML?

A: Supervised (labeled data) , Unsupervised (unlabeled data) , Reinforcement (using penalty and reward)

Q: What is your favorite algorithm? Is it used for classification or regression? (explain in under a minute)

A: Open-ended, Decision Trees

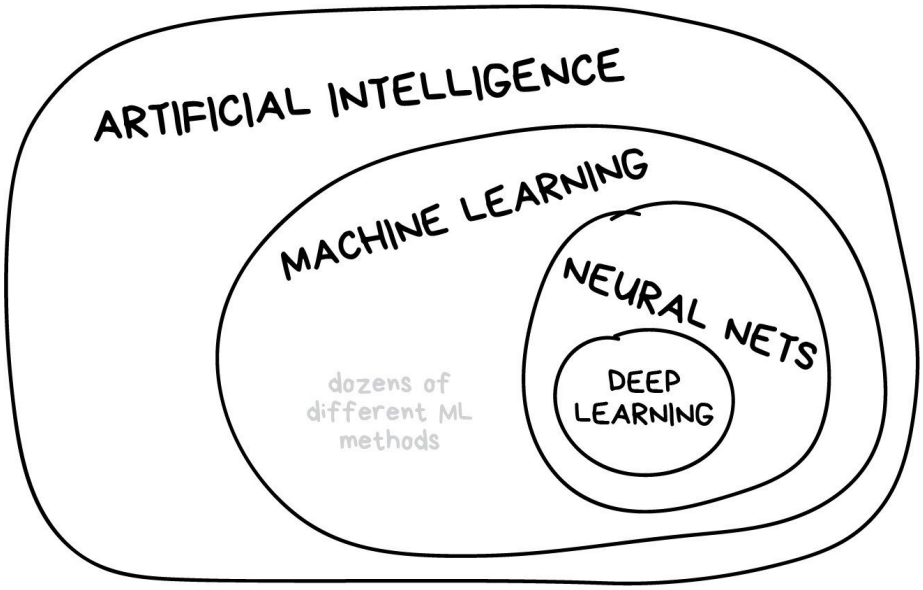
A decision tree learns from training data by creating a tree structure. Decision trees help determine feature importance because of this need to find the “best” attribute to split the data on at each node of the tree. They are considered a “greedy” algorithm because they only split by what is best at the current step without necessarily thinking steps ahead.

Decision Tree Advantages: Easy to understand model, Feature selection performed by the algorithm, Little data preparation is required

Decision Tree Disadvantages: Easy to over fit – especially as tree gets bigger, “Greedy” algorithm – may result in local optima

Q: What is the difference between ML, DL and AI?

A: - AI involves machines that can perform tasks that are characteristic of human intelligence ML is a way of achieving AI by “training” an algorithm so that it can learn data. Deep learning is one of many approaches to machine learning, which uses Neural Networks that mimic the biological structure of the brain. Another difference is the feature extraction and classification are separate steps for ML but are a single NN for DL.



Q: Talk about a recent ML paper that you've read in 2 minutes.

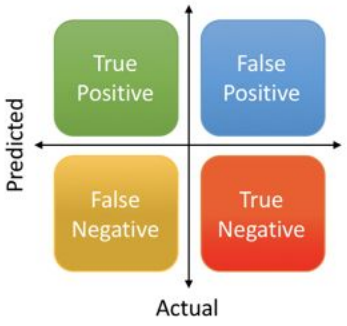
A: Open-ended

<u>Be able to explain the following metrics:</u> Accuracy, Precision, Recall, and F1, TP, TN, FP, FN

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



I asked my bf: What were the 5 presents I got him last Christmas? He couldn't exactly remember the 5 — so he randomly guessed seven times. Out of the 7 names:

- 5 were recalled correctly
- Even though he got a 100% recall (5/5) his precision was 71.4% (5/7).

Q: When should you use linear regression and when should you use logistic regression? Give an example.

A: In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values. In logistic regression, the outcome (dependent variable) has only a limited number of possible values.

For instance, if X contains the area in square feet of houses, and Y contains the corresponding sale price of those houses, you could use linear regression to predict selling price as a function of house size. While the possible selling price may not actually be any, there are so many possible values that a linear regression model would be chosen.

If, instead, you wanted to predict, based on size, whether a house would sell for more than $200K, you would use logistic regression. The possible outputs are either Yes, the house will sell for more than $200K, or No, the house will not.

Q: What are the questions for linear and logistic regression?

A: Linear regression gives an equation which is of the form Y = mX + C, means equation with degree 1. Logistic regression gives an equation which is of the form Y = eX + e-X

Q: Why shouldn't you use linear regression outputs as probabilities?

A: It's tempting to use linear regression outputs as probabilities but it's a mistake because the output can be negative, and greater than 1 whereas probability can not. As regression might actually produce probabilities that could be less than 0, or even bigger than 1, logistic regression was introduced.

Q: Explain how a ROC (Receiver operating characteristic) curve works.

A: The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs specificity (false positives). You want your model to get TPs faster than FPs, if there is the same rate of getting TP as getting FP your model is useless. (explanation)

Q: Explain TP/FP/FN/TN in a simple example:

A: - Fire alarm goes off + fire = TP Fire alarm goes off + no fire = FP Fire alarm doesn't go off + fire = FN Fire alarm doesn't go off + no fire = TN

Q: What is the difference between Type I and Type II error?

A: Type I Error leads to **False Positive** (FP). An example is a fire alarm going off when in fact there is no fire. This kind of error is synonymous to "believing a lie" or "a false alarm".

Type II Error- This is the incorrect retaining of a false Null Hypothesis (Ho). This type of error leads to **False Negative** (FN). An example of this is a fire breaking out and the fire alarm does not ring. This kind of error is synonymous to "failing to believe a truth" or "a miss".

Q: What is Bayes Theorem?

A: Essentially Bayes Theorem gives you a probability of an event given what is known as prior knowledge, P(A|B) = P(B|A)*P(A)/P(B) *(Prob = TP/ All positives(TP+FP))*. (visual explanation)

Q: Why is "Naive" Bayes naive?

A: NB makes the naive assumption that the features in a dataset are independent of each other, which isn't applicable to most real-world datasets.

Q: What is a decision tree and when would you choose to use one?

A: As the name suggests decision trees are tree-like model of decisions, they make relations between features easily interpretable. They can be used for both classification (classify passenger as survived or died) and regression (continuous values like price of a house) and don't require any assumptions of linearity in the data.

Q: How are they pruned?

A: Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning.

Reduced error pruning is perhaps the simplest version: replace each node. If it doesn't decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

Q: What is the difference between Gini Impurity and Entropy in a decision tree?

A: While both are metrics to decide how to split a tree, Gini measurement is the probability of a random sample being classified incorrectly by randomly picking a label from the branch. In information theory Entropy is the measured lack of information in a system and you calculate gain by making a split. This delta entropy tells you about how the uncertainty about the label was reduced. Gini is more common because it doesn't require the log calculations that Entropy takes.

Q: When will Entropy decrease in binary tree classification?

A: It decreases the closer we get to the leaf node.

Q: Why don't we tend to use linear regression to model binary responses?

A: Linear regression prediction output is continuous, if you want to model binary results you should use logistic regression.

Q: What is the difference between hinge loss and log loss?

A: The hinge loss is used for "maximum-margin" classification, most notably for support vector machines. Logistic loss diverges faster than hinge loss. So, in general, it will be more sensitive to outliers. Hinge loss also penalizes wrong answers, as well as correct unconfident answers.

Q: How do linear and logistic regression differ in their error minimization techniques?

A: Linear regression normally uses ordinary least squares method to minimize the errors and arrive at a best possible fit, while logistic regression uses maximum likelihood method to arrive at the solution (you can use maximum likelihood too for LR if you want).

Q: What is more important model accuracy or model performance?

A: Model accuracy is actually a subset of model performance. For example, if you wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were

fraud. However, this would be useless for a predictive model — a model designed to find fraud that asserted there was no fraud at all!

Q: What's the difference between a generative and discriminative model?

A: Discriminative models are great for classification (SVM, NN, NLPs, facial recognition), they map high dimensional sensory input into a class. A generative models care how the data was generated and will learn will learn categories of data (chatbot, GANs), unsupervised.

Q: How does SVM and logistic regression differ?

A: They only differ in the loss function — SVM minimizes hinge loss while logistic regression minimizes logistic loss.

Q: What is an SVM? What do you do if your data is not linear? (kernel trick)

A: The objective of the support vector machine algorithm is to find the hyperplane that has the maximum margin in an N-dimensional space(N — the number of features) that distinctly classifies the data points. A kernel trick allows you to map your data to a higher dimensional feature space so you can fit a hyperplane. This is done by taking the vectors in the original space and returning the dot product of the vectors in the feature space.

Q: How do you tune the regularization (C) and gamma terms in SVMs?

A: High gamma values mean only data points close to the line are considered and a high C term means a smaller-margin around line (could overfit).

Q: Explain Dijkstra's algorithm? (Know how to use it).

A: Dijkstra's algorithm is an algorithm for finding the shortest paths between nodes in a graph.

Q: How is KNN different from K-means clustering?

A: KNN or K-Nearest Neighbors is a supervised learning method technique used from classification or regression and does not require training. K-means is an unsupervised clustering algorithm fitting to K-clusters.

Q: What is the Elbow Method used for in KMeans?

A: The Elbow Method is a fundamental step for any unsupervised algorithm to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k in KMeans.

Q: What is the difference between intra-cluster variance and inter-cluster variance?

A: Inter-cluster variance is defined almost identical to intra-cluster variance. The difference is that you don't calculate the variance between all the samples within a single cluster but you take each cluster's centroid (typically the mean of all samples within a cluster) and calculate the variance between all centroids.

Q: What is ensemble learning?

A: Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. And they typically reduce overfitting in models. Ensembling techniques are further classified into Bagging and Boosting.

Q: What is the difference between bagging and boosting?

A: Both are ensemble models that use random sampling to reduce variance. Bagging models are built independently and better solves the problem of overfitting. Boosting builds on top of old models to create models with less bias, also weights the better performing examples higher, but may overfit.

Q: How do you go from a decision tree to a random forest? To a Gradient Boosted Tree?

A: Bagging takes many uncorrelated learners to make a final model and it reduces error by reducing variance. An example of a bagging ensemble is Random Forest.

Boosting is an ensemble technique in which the predictors are not made independently, but sequentially in order to learn from the mistakes of the previous predictors. Gradient Boosted Trees are an example of boosting algorithms.

Q: Describe a hash table. What makes a hash function?

A: A hash table is a data structure like a dictionary in python. A key is mapped to certain values through the use of a hash function. They are often used for tasks such as database indexing.

Q: How do you deal with imbalanced data?

A: Collect more data, resample the dataset to correct for imbalances, try different models or algorithms.

ML Validation

Q: Name two ways to evaluate the performance of a classification algorithm.

A: 1) Confusion Matrix ([TN,FP],[FN,TP]) 2) Accuracy (also AUC, F1, MAE, MSE)

Q: What's the difference between Type I and Type II error?

A: Type I error is a false positive, while Type II error is a false negative.

Q: What is the difference between MSE and MAE?

A: MAE loss is more robust to outliers, but its derivatives are not continuous, making it inefficient to find the solution. MSE loss is sensitive to outliers, but gives a more stable and closed form solution (by setting its derivative to 0). Use MAE if you have a lot of anomalies in your dataset.

Q: What is the difference between Cost and Loss functions?

A: The loss function computes the error for a single training example, while the cost function is the average of the loss functions of the entire training set.

Q: Why do we need a cost function and which is the best cost to use in classification algorithms.

A: We need a cost function to optimize our weights for model performance and I would use the cost function Mean Squared Error (or MAE or binary CE-- this is subjective) and minimize the MSE to improve the accuracy of our classification model.

Q: How do you ensure you're not overfitting with a model?

A: 1- Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data. 2- Use cross-validation techniques such as k-folds cross-validation. 3- Use regularization techniques such as LASSO that penalize certain model parameters if they're likely to cause overfitting.

Q: Explain the cross-validation resampling procedure.

A: The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group: Take the group as a hold out or test data set.
4. Take the remaining groups as a training data set.
5. Fit a model on the training set and evaluate it on the test set.
6. Retain the evaluation score and discard the model.
7. Combine evaluation scores into single average (CV error).
8. Repeat process for different model and choose the one w/ lowest CV error.

Q: How does evaluating your model differ between using CV or bootstrapping? What is MC-CV?

A: CV tends to be less biased but K-fold CV has fairly large variance. On the other hand, bootstrapping (sampling with replacement) tends to drastically reduce the variance but gives more biased results (they tend to be pessimistic). "Monte Carlo CV" aka "leave-group-out CV" does many random splits of the data to reduce variance.

Q: What's the trade-off between bias and variance?

A: Bias is due to overly simplistic assumptions while variance is error due to too much complexity in the learning algorithm you're using. Bias leads to under-fitting your data and variance leads to overfitting your data. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance and try your best to minimize each. (In machine learning/statistics as a whole, accuracy vs. precision is analogous to bias vs. variance).

Q: How can machine learning be used for time-series analysis?

A: Several ways!

Regression - Using time-based features such as week, month, day, day of week, etc as predictors. You can also add in external predictors that may influence the target (e.g. weather and temperature may affect sales of umbrellas).

ARIMA - Autoregressive Integrated Moving Average - Using autocorrelation (lags) as predictors

GARCH - Models changing variance

Time series decomposition - Splitting out trend, seasonality, etc

Others - Deep Learning, GAM-based models (prophet) - Can also be useful

Q: What cross-validation technique would you use on a time series dataset?

A: Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data — it is inherently ordered by chronological order. If a pattern emerges in later time periods for example, your model may still pick up on it even if that effect doesn't hold in earlier years!
You'll want to do something like forward chaining where you'll be able to model on past data then look at forward-facing data.

fold 1 : training [1], test [2] fold 2 : training [1 2], test [3] fold 3 : training [1 2 3], test [4] fold 4 : training [1 2 3 4], test [5] fold 5 : training [1 2 3 4 5], test [6]

Also remember prediction horizon here, a lot of applicants test models on an horizon of 1, which is a very easy task, but rarely useful. Most models have a hard time for useful prediction horizons!

Q: What's the difference between L1 and L2 regularization? How does it solve the problem of overfitting? Which regularizer to use and when?

A: Both L1 (Lasso Regression) and L2 (Ridge Regression) regularization techniques are used to address over-fitting and feature selection, the key difference between these two is the penalty term. Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds "absolute value of magnitude" of coefficient while Ridge regression adds "squared magnitude" of coefficient as penalty term to the loss function.

The key difference between these techniques is that Lasso is more binary/sparse and shrinks the less important features coefficient to zero thus, removing some features altogether and L2 regularization tends to spread error among all the terms. L1 works well for feature selection in case we have a huge number of features.

Q: How do you determine when you are overfitting or underfitting your model? What do we want to see for a good model?

Overfitting if: training loss << validation loss —- high variance

Underfitting if: training loss >> validation loss —- high bias

Just right if training loss ~ validation loss

ML Stats

Q: What is a Fourier transform? And why do we use it.

A: Fourier transforms in ML are mainly used to extract features from audio signals by converting a signal from time to frequency domain. Note that this is just one specific example of the many applications FT has and that the independent variable could be anything and not necessarily time. The Fourier transform decomposes a function of time into the frequencies that make it up (as symmetric functions), in a way similar to how a musical chord can be expressed as the frequencies of its constituent notes.

It is very useful to represent time series information using its spectral representation, which means all kinds of transforms like the Fourier Transform, Q-Transform, DCT's are used in processing all kinds of signals, from EEG's to ECG's to music. The fourier transform is a ultimately a feature set for your data that can be plugged into a Machine Learning pipeline.

Q: What's the difference between probability and likelihood?

A: For binomial distributions: Probability is the percentage that a success occurs. Likelihood is the conditional probability, i.e. the probability that the above event will happen.

Q: What is the difference between PCA and t-SNE? What are their use cases?

A: Both methods are used for dimensionality reduction, but t-SNE tries to deconvolve relationships between neighbors in high-dimensional data to understand the underlying structure of the data. Principal component analysis first identifies the hyperplane that lies closest to the data, and then projects the data onto it. PCA preserves the maximum amount of variance, but is much less computationally expensive than t-SNE.

Q: How do eigenvalues and eigenvectors relate to PCA?

A: Eigenvectors have corresponding eigenvalues and eigenvectors that have the largest eigenvalues will be the principal components (new dimensions of our data).

Q: What is Maximum Likelihood (MLE)?

A: Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.

Q: When are Maximum Likelihood and Least Squared Error equal?

A: For least squares parameter estimation we want to find the line that minimizes the total squared distance between the data points and the regression line. In maximum likelihood estimation we want to maximize the total probability of the data. When a Gaussian distribution is assumed, the maximum probability is found when the data points get closer to the mean value. Since the Gaussian distribution is symmetric, this is equivalent to minimizing the distance between the data points and the mean value.

Data Structure Questions

NLP Interview Questions

Sources:

Forwardpropagation — ML Glossary documentation
Machine Learning Interview Questions and Answers | Machine Learning Interview Preparation | Edureka
https://www.springboard.com/blog/machine-learning-interview-questions/
https://machinelearningmastery.com/k-fold-cross-validation/
https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831chttps://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433fhttps://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d
https://en.wikipedia.org/wiki/Fourier_transform
https://www.quora.com/Data-Science-Can-machine-learning-be-used-for-time-series-analysis
Linear & logistic regression
Google specific ML Interview