INTERVIEW QUESTIONS
For
MECHINE LEARNING

Website: www.analytixlabs.co.in
Email: info@analytixlabs.co.in

MACINE LEARNING

Q. What is cross-validation? How to do it right?
Ans: It's a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. Mainly used in settings where the goal is prediction and one wants to estimate how accurately a model will perform in practice. The goal of cross-validation is to define a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like over fitting, and get an insight on how the model will generalize to an independent data set.
Examples: leave-one-out cross validation, K-fold cross validation

How to do it right?

- the training and validation data sets have to be drawn from the same population
- predicting stock prices: trained for a certain 5-year period, it's unrealistic to treat the subsequent 5-year a draw from the same population
- common mistake: for instance the step of choosing the kernel parameters of a SVM should be cross-validated as well

Bias-variance trade-off for k-fold cross validation:

Leave-one-out cross-validation: gives approximately unbiased estimates of the test error since each training set contains almost the entire data set (n-1 observations).

We average the outputs of n fitted models, each of which is trained on an almost identical set of observations hence the outputs are highly correlated. Since the variance of a mean of quantities increases when correlation of these quantities increase, the test error estimate from a LOOCV has higher variance than the one obtained with k-fold cross validation
Typically, we choose k=5 or k=10, as these values have been shown empirically to yield test error estimates that suffer neither from excessively high bias nor high variance.

Q. Is it better to design robust or accurate algorithms?
Ans:

- The ultimate goal is to design systems with good generalization capacity, that is, systems that correctly identify patterns in data instances not seen before
- The generalization performance of a learning system strongly depends on the complexity of the model assumed
- If the model is too simple, the system can only capture the actual data regularities in a rough manner. In this case, the system has poor generalization properties and is said to suffer from under fitting
- By contrast, when the model is too complex, the system can identify accidental patterns in the training data that need not be present in the test set. These spurious patterns can be the result of random fluctuations or of measurement errors during the data collection process. In this case, the generalization capacity of the learning system is also poor. The learning system is said to be affected by over fitting
- Spurious patterns, which are only present by accident in the data, tend to have complex forms. This is the idea behind the principle of Occam's razor for avoiding over fitting: simpler models are preferred if more complex models do not significantly improve the quality of the description for the observations
- Quick response: Occam's razor. It depends on the learning task. Choose the right balance
- Ensemble learning can help balancing bias/variance (several weak learners together = strong learner)

Q. How to define/select metrics?
Ans:

- Type of task: regression? Classification?
- Business goal?
- What is the distribution of the target variable?
- What metric do we optimize for?
- Regression: RMSE (root mean squared error), MAE (mean absolute error), WMAE (weighted mean absolute error), MSLE (root mean squared logarithmic error)…
- Classification: recall, AUC, accuracy, misclassification error, Cohen's Kappa…

Common metrics in regression:

- Mean Squared Error Vs Mean Absolute Error RMSE gives a relatively high weight to large errors. The RMSE is most useful when large errors are particularly undesirable.
  The MAE is a linear score: all the individual differences are weighted equally in the average. MAE is more robust to outliers than MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Root Mean Squared Logarithmic Error
  RMSLE penalizes an under-predicted estimate greater than an over-predicted estimate (opposite to RMSE)

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1) - \log(a_i+1))^2}$$

Where p_i is the ith prediction, a_i the ith actual response, $log(b)$ the natural logarithm of b.

- Weighted Mean Absolute Error
  The weighted average of absolute errors. MAE and RMSE consider that each prediction provides equally precise information about the error variation, i.e. the standard variation of the error term is constant over all the predictions. Examples: recommender systems (differences between past and recent products)

$$WMAE = \frac{1}{\sum w_i}\sum_{i=1}^{n} w_i|y_i - \hat{y}_i|$$

Common metrics in classification:

- Recall / Sensitivity / True positive rate:
  High when FN low. Sensitive to unbalanced classes.
  $Sensitivity = \frac{TP}{TP+FN}$
- Precision / Positive Predictive Value
  High when FP low. Sensitive to unbalanced classes.
  $Precision = \frac{TP}{TP+FP}$
- Specificity / True Negative Rate
  High when FP low. Sensitive to unbalanced classes.
  $Specificity = \frac{TN}{TN+FP}$
- Accuracy
  High when FP and FN are low. Sensitive to unbalanced classes (see "Accuracy paradox")
  $Accuracy = \frac{TP+TN}{TN+TP+FP+FN}$
- ROC / AUC
  ROC is a graphical plot that illustrates the performance of a binary classifier ($Sensitivity$ Vs $1 - Specificity$ or $Sensitivity$ Vs $Specificity$). They are not sensitive to unbalanced classes.
  AUC is the area under the ROC curve. Perfect classifier: AUC=1, fall on (0,1); 100% sensitivity (no FN) and 100% specificity (no FP)
- Logarithmic loss
  Punishes infinitely the deviation from the true value! It's better to be somewhat wrong than emphatically wrong!
  $logloss = -\frac{1}{N}\sum_{i=1}^{n}(y_i \log(p_i) + (1-y_i)\log(1-p_i))$
- Misclassification Rate
  $Misclassification = \frac{1}{n}\sum_i I(y_i \neq \hat{y}_i)$
- F1-Score
  Used when the target variable is unbalanced. $F_1 Score = 2\frac{Precision \times Recall}{Precision + Recall}$

Q. Explain what regularization is and why it is useful. What are the benefits and drawbacks of specific methods, such as ridge regression and lasso?

Ans:

- Used to prevent over fitting: improve the generalization of a model
- Decreases complexity of a model
- Introducing a regularization term to a general loss function: adding a term to the minimization problem
- Impose Occam's Razor in the solution

Ridge regression:

- We use an L2 penalty when fitting the model using least squares
- We add to the minimization problem an expression (shrinkage penalty) of the form λ×∑coefficients
- λ: tuning parameter; controls the bias-variance tradeoff; accessed with cross-validation
- A bit faster than the lasso

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

The Lasso:

- We use an L1L1 penalty when fitting the model using least squares
- Can force regression coefficients to be exactly: feature selection method by itself

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} ||\beta_j|| \right\}$$

Q. What are L1 and L2 regularisation?
Ans: Mathematically speaking, it adds a **regularization**term in order to prevent the coefficients to fit so perfectly to overfit. The difference between the **L1 and L2** is just that **L2** is the sum of the square of the weights, while **L1** is just the sum of the weights

The differences of L1-norm and L2-norm as a loss function can be promptly summarized as follows:

L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection

Q. What is cross validation and grid search?
Ans: Cross-validation is a method for robustly estimating test-set performance (generalization) of a model. **Grid-search** is a way to select the best of a family of models, parameterized by a **grid** of parameters

Q. Is decision tree an ensemble learning algorithm
Ans: An **ensemble** is itself a supervised **learning algorithm**, because it can be trained and then used to make predictions. ... Although perhaps non-intuitive, more random **algorithms** (like random **decision trees**) can be used to produce a stronger **ensemble** than very deliberate **algorithms** (like entropy-reducing **decision trees**)

Q. Is correlation good?
Ans: In statistics, the **correlation** coefficient r measures the strength and direction of a linear **relationship** between two variables on a scatterplot. ... A perfect downhill (negative) linear **relationship**. –0.70. A strong downhill (negative) linear **relationship**.

Q. Can decision tree be used for clustering?
Ans: A **decision tree** algorithm **can** be **applied** to solve the problem. However, for the technique to work many important issues have to be addressed. The key issue is that the purity function **used** in **decision tree** building is not sufficient for **clustering**.

Q. Naive in Naive bayes?
Ans: In machine learning, **naive Bayes** classifiers are a family of simple "probabilistic classifiers" based on applying **Bayes**' theorem with strong (**naive**) independence assumptions between the features.

Q. Explain what is local optimum and why it is important in a specific context, such as K-means clustering. What are specific ways of determining if you have a local optimum problem? What can be done to avoid local optima?
Ans:

- A solution that is optimal in within a neighboring set of candidate solutions
- In contrast with global optimum: the optimal solution among all others
- K-means clustering context:
  It's proven that the objective cost function will always decrease until a local optimum is reached.
  Results will depend on the initial random cluster assignment

- Determining if you have a local optimum problem:
  Tendency of premature convergence
  Different initialization induces different optima
- Avoid local optima in a K-means context: repeat K-means and take the solution that has the lowest cost

Q. Assume you need to generate a predictive model using multiple regression. Explain how you intend to validate this model

Ans:

Validation using R^2:

- % of variance retained by the model
- Issue: R^2 is always increased when adding variables

$$R^2 = \frac{RSS_tot - \acute{RSS}_res}{RSS_tot} = \frac{RSS_reg}{RSS_tot} = 1 - \frac{\breve{RSS}_res}{RSS_tot}$$

Analysis of residuals:

- Heteroskedasticity (relation between the variance of the model errors and the size of an independent variable's observations)
- Scatter plots residuals Vs predictors
- Normality of errors
- Etc. : diagnostic plots

Out-of-sample evaluation: with cross-validation

Q. What is latent semantic indexing? What is it used for? What are the specific limitations of the method?

Ans:

- Indexing and retrieval method that uses singular value decomposition to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text
- Based on the principle that words that are used in the same contexts tend to have similar meanings
- "Latent": semantic associations between words is present not explicitly but only latently
- For example: two synonyms may never occur in the same passage but should nonetheless have highly associated representations

Used for:

- Learning correct word meanings
- Subject matter comprehension
- Information retrieval
- Sentiment analysis (social network analysis)

Q. Explain what resampling methods are and why they are useful

Ans:

- repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model
- example: repeatedly draw different samples from training data, fit a linear regression to each new sample, and then examine the extent to which the resulting fit differ
- most common are: cross-validation and the bootstrap
- cross-validation: random sampling with no replacement
- bootstrap: random sampling with replacement
- cross-validation: evaluating model performance, model selection (select the appropriate level of flexibility)
- bootstrap: mostly used to quantify the uncertainty associated with a given estimator or statistical learning method

Q. What is principal component analysis? Explain the sort of problems you would use PCA for. Also explain its limitations as a method

Ans: Statistical method that uses an orthogonal transformation to convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components.

Reduce the data from n to k dimensions: find the k vectors onto which to project the data so as to minimize the projection error.

Algorithm:

1) Pre-processing (standardization): PCA is sensitive to the relative scaling of the original variable
2) Compute covariance matrix Σ
3) Compute eigenvectors of Σ
4) Choose k principal components so as to retain x% of the variance (typically x=99)

Applications:

1) Compression
- Reduce disk/memory needed to store data
- Speed up learning algorithm. Warning: mapping should be defined only on training set and then applied to test set

2) Visualization: 2 or 3 principal components, so as to summarize data

Limitations:
- PCA is not scale invariant
- The directions with largest variance are assumed to be of most interest
- Only considers orthogonal transformations (rotations) of the original variables
- PCA is only based on the mean vector and covariance matrix. Some distributions (multivariate normal) are characterized by this but some are not
- If the variables are correlated, PCA can achieve dimension reduction. If not, PCA just orders them according to their variances

Q. Explain what a false positive and a false negative are. Why is it important these from each other? Provide examples when false positives are more important than false negatives, false negatives are more important than false positives and when these two types of errors are equally important

Ans:

- False positive
  improperly reporting the presence of a condition when it's not in reality. Example: HIV positive test when the patient is actually HIV negative
- False negative
  improperly reporting the absence of a condition when in reality it's the case. Example: not detecting a disease when the patient has this disease.

When false positives are more important than false negatives:

- In a non-contagious disease, where treatment delay doesn't have any long-term consequences but the treatment itself is grueling
- HIV test: psychological impact

When false negatives are more important than false positives:
- If early treatment is important for good outcomes
- In quality control: a defective item passes through the cracks!
- Software testing: a test to catch a virus has failed

Q. What is the difference between supervised learning and unsupervised learning? Give concrete examples
Ans:

- Supervised learning: inferring a function from labeled training data
- Supervised learning: predictor measurements associated with a response measurement; we wish to fit a model that relates both for better understanding the relation between them (inference) or with the aim to accurately predicting the response for future observations (prediction)
- Supervised learning: support vector machines, neural networks, linear regression, logistic regression, extreme gradient boosting
- Supervised learning examples: predict the price of a house based on the are, size; churn prediction; predict the relevance of search engine results.
- Unsupervised learning: inferring a function to describe hidden structure of unlabeled data
- Unsupervised learning: we lack a response variable that can supervise our analysis
- Unsupervised learning: clustering, principal component analysis, singular value decomposition; identify group of customers
- Unsupervised learning examples: find customer segments; image segmentation; classify US senators by their voting.

Q. What does NLP stand for?
Ans: "Natural language processing"!

- Interaction with human (natural) and computers languages
- Involves natural language understanding

Major tasks:
- Machine translation
- Question answering: "what's the capital of Canada?"
- Sentiment analysis: extract subjective information from a set of documents, identify trends or public opinions in the social media
- Information retrieval

Q. What are feature vectors?
Ans:

- n-dimensional vector of numerical features that represent some object
- term occurrences frequencies, pixels of an image etc.
- Feature space: vector space associated with these vectors

Q. When would you use random forests Vs SVM and why?
Ans:

- In a case of a multi-class classification problem: SVM will require one-against-all method (memory intensive)
- If one needs to know the variable importance (random forests can perform it as well)
- If one needs to get a model fast (SVM is long to tune, need to choose the appropriate kernel and its parameters, for instance sigma and epsilon)
- In a semi-supervised learning context (random forest and dissimilarity measure): SVM can work only in a supervised learning mode

Q. How do you take millions of users with 100's transactions each, amongst 10k's of products and group the users together in meaningful segments?
Ans:

1. Some exploratory data analysis (get a first insight)

- Transactions by date
- Count of customers Vs number of items bought

- Total items Vs total basket per customer
- Total items Vs total basket per area

2. Create new features (per customer):

Counts:

- Total baskets (unique days)
- Total items
- Total spent
- Unique product id

Distributions:

- Items per basket
- Spent per basket
- Product id per basket
- Duration between visits
- Product preferences: proportion of items per product cat per basket

3. Too many features, dimension-reduction? PCA?
4. Clustering:

- PCA

5. Interpreting model fit

- View the clustering by principal component axis pairs PC1 Vs PC2, PC2 Vs PC1.
- Interpret each principal component regarding the linear combination it's obtained from; example: PC1=spend y axis (proportion of baskets containing spend y items, raw counts of items and visits)

Q. How do you know if one algorithm is better than other?
Ans:

- In terms of performance on a given data set?
- In terms of performance on several data sets?
- In terms of efficiency?

In terms of performance on several data sets:
- "Does learning algorithm A have a higher chance of producing a better predictor than learning algorithm B in the given context?"
- "Bayesian Comparison of Machine Learning Algorithms on Single and Multiple Datasets", A. Lacoste and F. Laviolette
- "Statistical Comparisons of Classifiers over Multiple Data Sets", Janez Demsar

In terms of performance on a given data set:
- One wants to choose between two learning algorithms
- Need to compare their performances and assess the statistical significance

One approach (Not preferred in the literature):
- Multiple k-fold cross validation: run CV multiple times and take the mean and sd
- You have: algorithm A (mean and sd) and algorithm B (mean and sd)
- Is the difference meaningful? (Paired t-test)

Sign-test (classification context):
Simply counts the number of times A has a better metrics than B and assumes this comes from a binomial distribution. Then we can obtain a p-value of the Ho test: A and B are equal in terms of performance.
Wilcoxon signed rank test (classification context):
Like the sign-test, but the wins (A is better than B) are weighted and assumed coming from a symmetric distribution around a common median. Then, we obtain a p-value of the Ho test.

Other (without hypothesis testing):
- AUC
- F-Score
- See question 3

Q. How do you test whether a new credit risk scoring model works?
Ans:

- Test on a holdout set
- Kolmogorov-Smirnov test

Kolmogorov-Smirnov test:
- Non-parametric test
- Compare a sample with a reference probability distribution or compare two samples
- Quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution
- Or between the empirical distribution functions of two samples
- Null hypothesis (two-samples test): samples are drawn from the same distribution
- Can be modified as a goodness of fit test
- In our case: cumulative percentages of good, cumulative percentages of bad

Q. What is: collaborative filtering, n-grams, cosine distance?
Ans: Collaborative filtering:
- Technique used by some recommender systems
- Filtering for information or patterns using techniques involving collaboration of multiple agents: viewpoints, data sources.

1. A user expresses his/her preferences by rating items (movies, CDs.)
2. The system matches this user's ratings against other users' and finds people with most similar tastes
3. With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user

n-grams:
- Contiguous sequence of n items from a given sequence of text or speech
- "Andrew is a talented data scientist"
- Bi-gram: "Andrew is", "is a", "a talented".
- Tri-grams: "Andrew is a", "is a talented", "a talented data".
- An n-gram model models sequences using statistical properties of n-grams; see: Shannon Game
- More concisely, n-gram model: P(Xi|Xi−(n−1)...Xi−1): Markov model
- N-gram model: each word depends only on the n−1n−1 last words

Issues:
- when facing infrequent n-grams
- solution: smooth the probability distributions by assigning non-zero probabilities to unseen words or n-grams
- Methods: Good-Turing, Backoff, Kneser-Kney smoothing

Cosine distance:
- How similar are two documents?
- Perfect similarity/agreement: 1
- No agreement: 0 (orthogonality)
- Measures the orientation, not magnitude

Given two vectors A and B representing word frequencies:
cosine-similarity(A,B)=⟨A,B⟩/||A||·||B||

Q. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?
Ans:

- Good data is definitely more important than good models
- If quality of the data wasn't of importance, organizations wouldn't spend so much time cleaning and preprocessing it!
- Even for scientific purpose: good data (reflected by the design of experiments) is very important

How do you define good?
- good data: data relevant regarding the project/task to be handled
- good model: model relevant regarding the project/task
- good model: a model that generalizes on external data sets

Is there a universal good model?
- No, otherwise there wouldn't be the over fitting problem!
- Algorithm can be universal but not the model
- Model built on a specific data set in a specific organization could be ineffective in other data set of the same organization
- Models have to be updated on a somewhat regular basis

Are there any models that are definitely not so good?
- "all models are wrong but some are useful" George E.P. Box
- It depends on what you want: predictive models or explanatory power
- If both are bad: bad model

Q. Why is naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?
Ans:

- Naïve: the features are assumed independent/uncorrelated
- Assumption not feasible in many cases
- Improvement: de-correlate features (covariance matrix into identity matrix)

Q. What are the drawbacks of linear model? Are you familiar with alternatives (Lasso, ridge regression)?
Ans:

- Assumption of linearity of the errors
- Can't be used for count outcomes, binary outcomes
- Can't vary model flexibility: over fitting problems
- Alternatives: see question 4 about regularization

Q. Do you think 50 small decision trees are better than a large one? Why?
Ans:

- Yes!
- More robust model (ensemble of weak learners that come and make a strong learner)
- Better to improve a model by taking many small steps than fewer large steps
- If one tree is erroneous, it can be auto-corrected by the following
- Less prone to over fitting.

Q. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything? Are you familiar with A/B testing?
Ans:

Example with linear regression:
- F-statistic (ANOVA)

$$F = \frac{\frac{RSS_1 - RSS_2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}}$$

p1: number of parameters of model 1
p2: number of parameters of model 2
nn: number of observations
Under the null hypothesis that model 2 doesn't provide a significantly better fit than model 1, F will have an F distribution with p2−p1,n−p2) degrees of freedom. The null hypothesis is rejected if the F calculated from the data is greater than the critical value of the F distribution for some desired significance level.
Others: AIC/BIC (regression), cross-validation: assessing test error on a test/validation set

Q. What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?
Ans:

- Effect would be similar to regularization: avoid over fitting
- Used to increase robustness

Q. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with?
data reduction techniques other than PCA?:
Ans: Partial least squares: like PCR (principal component regression) but chooses the principal components in a supervised way. Gives higher weights to variables that are most strongly related to the response

step-wise regression?
- the choice of predictive variables are carried out using a systematic procedure
- Usually, it takes the form of a sequence of F-tests, t-tests, adjusted R-squared, AIC, BIC
- at any given step, the model is fit using unconstrained least squares
- can get stuck in local optima
- Better: Lasso

step-wise techniques:
- Forward-selection: begin with no variables, adding them when they improve a chosen model comparison criterion
- Backward-selection: begin with all the variables, removing them when it improves a chosen model comparison criterion

Better than reduced data:
Example 1: If all the components have a high variance: which components to discard with a guarantee that there will be no significant loss of the information?
Example 2 (classification):
- One has 2 classes; the within class variance is very high as compared to between class variance
- PCA might discard the very information that separates the two classes

Better than a sample:
- When number of variables is high relative to the number of observations

Q. How would you define and measure the predictive power of a metric?
Ans:

- Predictive power of a metric: the accuracy of a metric's success at predicting the empirical
- They are all domain specific
- Example: in field like manufacturing, failure rates of tools are easily observable. A metric can be trained and the success can be easily measured as the deviation over time from the observed
- In information security: if the metric says that an attack is coming and one should do X. Did the recommendation stop the attack or the attack never happened?

Q. Do we always need the intercept term in a regression model?
Ans:

- It guarantees that the residuals have a zero mean
- It guarantees the least squares slopes estimates are unbiased
- the regression line floats up and down, by adjusting the constant, to a point where the mean of the residuals is zero

Q. What are the assumptions required for linear regression? What if some of these assumptions are violated?
Ans:

1. The data used in fitting the model is representative of the population
2. The true underlying relation between x and y is linear
3. Variance of the residuals is constant (homoscedastic, not heteroscedastic)
4. The residuals are independent
5. The residuals are normally distributed

Predict y from x: 1) + 2)
Estimate the standard error of predictors: 1) + 2) + 3)
Get an unbiased estimation of y from x: 1) + 2) + 3) + 4)
Make probability statements, hypothesis testing involving slope and correlation, confidence intervals: 1) + 2) + 3) + 4) + 5)

Note:
- Common mythology: linear regression doesn't assume anything about the distributions of x and y
- It only makes assumptions about the distribution of the residuals
- And this is only needed for statistical tests to be valid
- Regression can be applied to many purposes, even if the errors are not normally distributed

31. What is collinearity and what to do with it? How to remove multicollinearity?

Collinearity/Multicollinearity:
- In multiple regression: when two or more variables are highly correlated
- They provide redundant information
- In case of perfect multicollinearity: $\beta=(XTX)^{-1}XTy$ doesn't exist, the design matrix isn't invertible
- It doesn't affect the model as a whole, doesn't bias results
- The standard errors of the regression coefficients of the affected variables tend to be large
- The test of hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the explanatory (Type II error)
- Leads to over fitting

Remove multicollinearity:
- Drop some of affected variables
- Principal component regression: gives uncorrelated predictors
- Combine the affected variables

- Ridge regression
- Partial least square regression

Detection of multicollinearity:
- Large changes in the individual coefficients when a predictor variable is added or deleted
- Insignificant regression coefficients for the affected predictors but a rejection of the joint hypothesis that those coefficients are all zero (F-test)
- VIF: the ratio of variances of the coefficient when fitting the full model divided by the variance of the coefficient when fitted on its own
- rule of thumb: VIF>5 indicates multicollinearity
- Correlation matrix, but correlation is a bivariate relationship whereas multicollinearity is multivariate

Q. How to check if the regression model fits the data well?
Ans: R squared/Adjusted R squared:
- Describes the percentage of the total variation described by the model
- R^2 always increases when adding new variables: adjusted R^2 incorporates the model's degrees of freedom

F test:
- Evaluate the hypothesis Ho: all regression coefficients are equal to zero Vs H1: at least one doesn't
- Indicates that R^2 is reliable

RMSE:
- Absolute measure of fit (whereas R2 is a relative measure of fit)

Q. What is a decision tree?
Ans:
1. Take the entire data set as input
2. Search for a split that maximizes the "separation" of the classes. A split is any test that divides the data in two (e.g. if variable2>10)
3. Apply the split to the input data (divide step)
4. Re-apply steps 1 to 2 to the divided data
5. Stop when you meet some stopping criteria
6. (Optional) Clean up the tree when you went too far doing splits (called pruning)

Finding a split: methods vary, from greedy search (e.g. C4.5) to randomly selecting attributes and split points (random forests)
Purity measure: information gain, Gini coefficient, Chi Squared values
Stopping criteria: methods vary from minimum size, particular confidence in prediction, purity criteria threshold
Pruning: reduced error pruning, out of bag error pruning (ensemble methods)

Q. What impurity measures do you know?
Ans:
Gini:
$Gini = 1 - \sum_j p_j^2$
Information Gain/Deviance
Information Gain= $\sum_j p_j log_2 p_j$
Better than Gini when pj are very small: multiplying very small numbers leads to rounding errors, we can instead take logs.

Q. What is random forest? Why is it good?
Ans:
Random forest? (Intuition):
- Underlying principle: several weak learners combined provide a strong learner
- Builds several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
- Rule of thumb: at each split m=sqrt(p)
- Predictions: at the majority rule

Why is it good?
- Very good performance (de-correlates the features)
- Can model non-linear class boundaries
- Generalization error for free: no cross-validation needed, gives an unbiased estimate of the generalization error as the trees is built
- Generates variable importance

Q. How do we train a logistic regression model? How do we interpret its coefficients?
Ans:

$\log(odds) = \log(\frac{P(y=1|x)}{P(y=0|x)}) =$ is a linear function of the input features

Minimization objective/Cost function:

$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\beta(x^i)) + (1-y^i)\log(1-h_\beta(x^i))$

Where: $h_\beta(x) = g(\beta^T x)$ and $g(z) = \frac{1}{1+e^{-z}}$ (sigmoid function)

Intuition:

- if $y_i = 0$, $J(\beta) = log(1-h_\beta(x)_i)$, will converge to ∞ as $h_\beta(x)_i$ becomes far from 0
- Converse: when $y_i = 1$, $J(\beta) = log(h_\beta(x)_i)$, will converge to ∞ as $h_\beta(x)_i$ becomes far from 1

Interpretation of the coefficients: the increase of log odds for the increase of one unit of a predictor, given all the other predictors are fixed

Q. What is the maximal margin classifier? How this margin can be achieved?
Ans:

- When the data can be perfectly separated using a hype rplane, there actually exists an infinite number of these hyper planes
- Intuition: a hyper plane can usually be shifted a tiny bit up, or down, or rotated, without coming into contact with any of the observations
- Large margin classifier: choosing the hyper plane that is farthest from the training observations
- This margin can be achieves using support vectors

Q. Which kernels do you know? How to choose a kernel?
Ans:

- Gaussian kernel
- Linear kernel
- Polynomial kernel
- Laplace kernel
- Esoteric kernels: string kernels, chi-square kernels

- If number of features is large (relative to number of observations): SVM with linear kernel ; e.g. text classification with lots of words, small training example
- If number of features is small, number of observations is intermediate: Gaussian kernel
- If number of features is small, number of observations is small: linear kernel

Q. Is it beneficial to perform dimensionality reduction before fitting an SVM? Why or why not?
Ans:

- When the number of features is large comparing to the number of observations (e.g. document-term matrix)
- SVM will perform better in this reduced space

Q. What is curse of dimensionality? How does it affect distance and similarity measures?
Ans:

- Refers to various phenomena that arise when analyzing and organizing data in high dimensional spaces
- Common theme: when number of dimensions increases, the volume of the space increases so fast that the available data becomes sparse
- Issue with any method that requires statistical significance: the amount of data needed to support the result grows exponentially with the dimensionality
- Issue when algorithms don't scale well on high dimensions typically when O(n^kn)
- Everything becomes far and difficult to organize
-

Illustrative example: compare the proportion of an inscribed hypersphere with radius r and dimension d to that of a hypercube with edges of length $2r$

- Volume of such a sphere is $V_{sphere} = \frac{2r^d \pi^{d/2}}{d\Gamma(d/2)}$

- The volume of the cube is: $V_{cube} = 2r^d$

As d increases (space dimension), the volume of hypersphere becomes insignificant relative to the volume of the hypercube:

$$\lim_{d \to \infty} \frac{V_{sphere}}{V_{cube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} = 0$$

- Nearly all of the dimensional space is far away from the center
- It consists almost entirely of the corners of the hypercube, with no middle!

Q. Which Random Forest parameters can be tuned to enhance the predictive power of the model? - Programming

Ans: A very good thing about the 'Random forests' algorithm is that it works usually good with default parameters, unlike other techniques such as SVM.
But it doesn't imply that tuning is not needed altogether. The parameters to tune is the number of trees in the ensemble (important), the depth of the trees in the ensemble, number of features used for splitting (important), the minimum size of the parent node and minimum size of the leaf node in a tree.

Also, one way to estimate the optimal parameters is through genetic algorithm with cross validation.

Q. What is Machine learning?
Ans: Do you have any experience in building ontologies Machine learning is a branch of computer science which deals with system programming in order to automatically learn and improve with experience. For example: Robots are programed so that they can perform the task based on data they gather from sensors. It automatically learns programs from data.

Q. Mention the difference between Data Mining and Machine learning?
Ans: Do you have any experience in building ontologies Machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. While, data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns. During this process machine, learning algorithms are used.

Q. What is 'Overfitting' in Machine learning?
Ans: In machine learning, when a statistical model describes random error or noise instead of underlying relationship 'overfitting' occurs. When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.

Q. Why overfitting happens?
Ans: The possibility of overfitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.

Q. How can you avoid overfitting ?
Ans: By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as cross validation. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.

In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to "test" the model in the training phase.

Q. What is inductive machine learning?
Ans: The inductive machine learning involves the process of learning by examples, where a system, from a set of observed instances tries to induce a general rule.

Q. What are the five popular algorithms of Machine Learning?
Ans:

a. Decision Trees
b. Neural Networks (back propagation)
c. Probabilistic networks
d. Nearest Neighbor
e. Support vector machines

Q. What are the different Algorithm techniques in Machine Learning?
Ans:

The different types of techniques in Machine Learning are

a. Supervised Learning
b. Unsupervised Learning
c. Semi-supervised Learning
d. Reinforcement Learning
e. Transduction
f. Learning to Learn

Q. What are the three stages to build the hypotheses or model in machine learning?
Ans:

a) Model building
b) Model testing
c) Applying the model

Q. What is the standard approach to supervised learning?
Ans: The standard approach to supervised learning is to split the set of example into the training set and the test.

Q. What is 'Training set' and 'Test set'?
Ans: In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as 'Training Set'. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

Q. List down various approaches for machine learning?
Ans: The different approaches in Machine Learning are

a. Concept Vs Classification Learning
b. Symbolic Vs Statistical Learning
c. Inductive Vs Analytical Learning

Q. What is not Machine Learning?
Ans:

a. Artificial Intelligence
b. Rule based inference

Q. Explain what is the function of 'Unsupervised Learning'?
Ans:

a. Find clusters of the data
b. Find low-dimensional representations of the data
c. Find interesting directions in data
d. Interesting coordinates and correlations
e. Find novel observations/ database cleaning

Q. Explain what is the function of 'Supervised Learning'?
Ans:

a. Classifications
b. Speech recognition
c. Regression

d. Predict time series
e. Annotate strings

Q. What is algorithm independent machine learning?
Ans: Machine learning in where mathematical foundations is independent of any particular classifier or learning algorithm is referred as algorithm independent machine learning?

Q. What is the difference between artificial learning and machine learning?
Ans: Designing and developing algorithms according to the behaviours based on empirical data are known as Machine Learning. While artificial intelligence in addition to machine learning, it also covers other aspects like knowledge representation, natural language processing, planning, robotics etc.
Q. What is classifier in machine learning?
Ans: A classifier in a Machine Learning is a system that inputs a vector of discrete or continuous feature values and outputs a single discrete value, the class.
Q. What are the advantages of Naive Bayes?
Ans: In Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. The main advantage is that it can't learn interactions between features.

Q. In what areas Pattern Recognition is used?
Ans: Pattern Recognition can be used in

a. Computer Vision
b. Speech Recognition
c. Data Mining
d. Statistics
e. Informal Retrieval
f. Bio-Informatics

Q. What is Genetic Programming?
Ans: Genetic programming is one of the two techniques used in machine learning. The model is based on the testing and selecting the best choice among a set of results.

Q. What is Inductive Logic Programming in Machine Learning?
Ans: Inductive Logic Programming (ILP) is a subfield of machine learning which uses logical programming representing background knowledge and examples.

Q. What is Model Selection in Machine Learning?
Ans: The process of selecting models among different mathematical models, which are used to describe the same data set is known as Model Selection. Model selection is applied to the fields of statistics, machine learning and data mining.

Q. What are the two methods used for the calibration in Supervised Learning?
The two methods used for predicting good probabilities in Supervised Learning are

a. Platt Calibration
b. Isotonic Regression

These methods are designed for binary classification, and it is not trivial.
Q. Which method is frequently used to prevent overfitting?
Ans: When there is sufficient data 'Isotonic Regression' is used to prevent an overfitting issue.

Q. What is the difference between heuristic for rule learning and heuristics for decision trees?

Ans: The difference is that the heuristics for decision trees evaluate the average quality of a number of disjointed sets while rule learners only evaluate the quality of the set of instances that is covered with the candidate rule.

Q. What is Perceptron in Machine Learning?
Ans: In Machine Learning, Perceptron is an algorithm for supervised classification of the input into one of several possible non-binary outputs.

Q. Explain the two components of Bayesian logic program?
Ans: Bayesian logic program consists of two components. The first component is a logical one; it consists of a set of Bayesian Clauses, which captures the qualitative structure of the domain. The second component is a quantitative one, it encodes the quantitative information about the domain.

Q. What are Bayesian Networks (BN)?
Ans: Bayesian Network is used to represent the graphical model for probability relationship among a set of variables.

Q. Why instance based learning algorithm sometimes referred as Lazy learning algorithm?
Ans: Instance based learning algorithm is also referred as Lazy learning algorithm as they delay the induction or generalization process until classification is performed.

Q. What are the two classification methods that SVM (Support Vector Machine) can handle?

a. Combining binary classifiers
b. Modifying binary to incorporate multiclass learning

Q. What is ensemble learning?
Ans: To solve a particular computational program, multiple models such as classifiers or experts are strategically generated and combined. This process is known as ensemble learning.

Q. Why ensemble learning is used?
Ans: Ensemble learning is used to improve the classification, prediction, and function approximation etc. of a model.

Q. When to use ensemble learning?
Ans: Ensemble learning is used when you build component classifiers that are more accurate and independent from each other.

Q. What are the two paradigms of ensemble methods?
Ans: The two paradigms of ensemble methods are

a. Sequential ensemble methods
b. Parallel ensemble methods

Q. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?
Ans: The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are

used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

Q. What is bias-variance decomposition of classification error in ensemble method?
Ans: The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm's prediction fluctuates for different training sets.

Q. What is an Incremental Learning algorithm in ensemble?
Incremental learning method is the ability of an algorithm to learn from new data that may be available after classifier has already been generated from already available dataset.

Q. What is PCA, KPCA and ICA used for?
Ans: PCA (Principal Components Analysis), KPCA (Kernel based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.

Q. What is dimension reduction in Machine Learning?
Ans: In Machine Learning and statistics, dimension reduction is the process of reducing the number of random variables under considerations and can be divided into feature selection and feature extraction

Q. What are support vector machines?
Ans: Support vector machines are supervised learning algorithms used for classification and regression analysis.

Q. What are the components of relational evaluation techniques?
Ans: The important components of relational evaluation techniques are

a. Data Acquisition
b. Ground Truth Acquisition
c. Cross Validation Technique
d. Query Type
e. Scoring Metric
f. Significance Test

Q. What are the different methods for Sequential Supervised Learning?
Ans: The different methods to solve Sequential Supervised Learning problems are

a. Sliding-window methods
b. Recurrent sliding windows
c. Hidden Markow models
d. Maximum entropy Markow models
e. Conditional random fields
f. Graph transformer networks

Q. What are the areas in robotics and information processing where sequential prediction problem arises?
Ans: The areas in robotics and information processing where sequential prediction problem arises are

a. Imitation Learning

b. Structured prediction
c. Model based reinforcement learning

Q. What is batch statistical learning?
Ans: Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

Q. What is PAC Learning?
Ans: PAC (Probably Approximately Correct) learning is a learning framework that has been introduced to analyze learning algorithms and their statistical efficiency.

Q. What are the different categories you can categorized the sequence learning process?
Ans:

a. Sequence prediction
b. Sequence generation
c. Sequence recognition
d. Sequential decision

Q. What is sequence learning?
Ans: Sequence learning is a method of teaching and learning in a logical manner.

Q. What are two techniques of Machine Learning?
Ans: The two techniques of Machine Learning are

a. Genetic Programming
b. Inductive Learning

Q. Give a popular application of machine learning that you see on day to day basis?
Ans: The recommendation engine implemented by major ecommerce websites uses Machine Learning