1. **Modeling**
   1. How to find outlier
   2. How to handle missing value
      1. Types of missing value
      2. Method
   3. How to handle imbalanced data
      1. Definition:
      2. Methods
         1. Proper Evaluation Metric
   4. Model evaluation to understand the characteristics and application of each metrics
      1. Cross-Validation, stratified cross-validation
      2. MSE, MAE, impurity function, cross-entropy, precision, recall, AUC, ROC, F1
   5. False positive and false negative: give example where false positive is more important than false negative
      1. The patient may be diagnosed with diabetes when they actually do not have the disease. This is a false positive. This can lead to unnecessary medical treatment.
   6. How to choose a feature
      1. PCA
   7. Overfitting, underfitting respective performance and solution
      1. Overfitting
         1. model tries too hard to capture the noise in the training set
   8. Variance/bias trade-off
   9. Out-of-bag sample(RF有)
   10. Explain gradient descent, stochastic gradient descent, mini-batch gradient descent
   11. Difference between statistical learning and machine learning
   12. Spherical hashing
       1. Parametric/ Non-parametric model
          1. Parametric
          2. Non-parametric
             1. Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric ML algo
             2. Examples

K-Nearest Neighbors

Decision Trees

SVM

* + - * 1. Pros

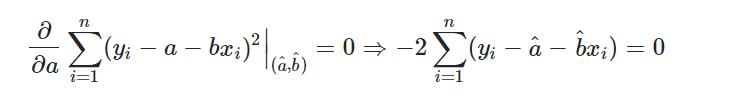
Flexibility, Power (No assumptions), Performance

* + - * 1. Cons

More Data, Slower, Overfitting

* + 1. Generative/ Discriminant model
       1. Generative Model
       2. Discriminant model
    2. Curse of dimension
       1. Definition:.
       2. Sparsity
       3. Solution

1. **Regression**
   1. The basic assumptions of linear regression, what to do when the basic assumptions are violated
      1. Linearity: The relationship between X and the mean of Y is linear
         1. If not linear: apply a nonlinear transformation to independent and dependent variable. Like taking the log, sqrt, reciprocal.
      2. Homoscedasticity: the variance of residual is the same for any value of X
         1. How to test: plot the fitted value vs. residual plot
         2. Simply take the log of the dependent variable
         3. Use weighted regression
      3. Independence: Observation are independent of each other
         1. Mostly relevant when working with time series data
         2. How to test: plot the residuals vs. time
         3. Adding lags, adding seasonal dummy variables
      4. Normality: The residuals of the model is normally distributed
         1. How to test: QQ plots, Shapiro-Wilk test
         2. Verify any outliers aren’t having a huge impact on the distribution
      5. If one of more of these assumptions are violated, then the results of our linear regression may be unreliable or misleading
   2. How to measure covariance, VIF
      1. Covariance: a measure of how changes in one variable are associated with changed in a second variable
   3. Comparison of correlation and causation, how to measure each
      1. Correlation
         1. The statistical indicator of the relationship between variables
      2. Causation
         1. means that changes in one variable brings about changes in the other; there is a cause-and-effect relationship between variables.
         2. Hypothesis testing and A/B testing
   4. Linear regression, how to change the model when performing various linear transformation on the data, how to change the predictive value, R-squared, coefficients, etc.
   5. Why the sum of residuals is zero under OLS
      1. Ordinary Least Square estimator minimizes the sum of squared residuals.



* 1. How to determine how well the model fits based on residual plot and QQ-plot
  2. The potential points that I can think of that have not been tested
     1. How to estimate the parameters of logistic regression
     2. The form of the LOSS function of logistic regression
     3. Why OLS estimation is used in linear regression, and some properties of OLS estimators

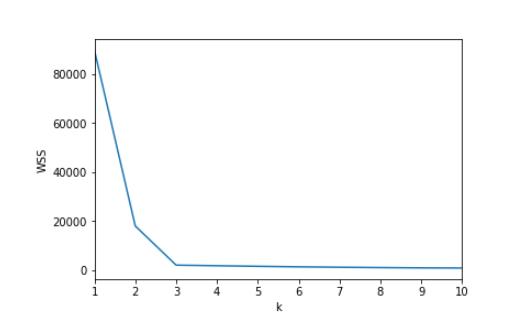
1. **Regularization**
   1. Definition and main function
      1. Avoid overfitting by constraining or shrinking the coefficient estimates towards zero (penalize the flexibility of our model)
   2. Comparing Lasso and Ridge
      1. Lasso is more likely to set parameter lamda to 0, which means we can filter out some unnecessary features, so we can get a sparse model.
      2. While Ridge will be more likely to yield a smoothing model.
      3. Why? On L1, w\* is more likely to be tangent on the vertex of the square, which means that some of the w is 0.

Diagram

Description automatically generated

* 1. Are the results of Lasso the same for different programming languages?
     1. No, because the grid is not the same
  2. L1 norm and L2 norm
     1. :
     2. :
     3. :
  3. Are the estimated coefficients of Regularization unbiased?
     1. ??

1. **Tree & Ensemble**
   1. Explain the tree model
      1. It is a supervised machine learning which performs classification and regression tasks by building tree-like structure for deciding the target variable class or value according to the features
      2. Entropy:
      3. ID3 Tree
         1. Gain:
            1. V: 对attribute a进行划分后有V个分支
            2. Choose the a with the greatest Gain
      4. C4.5 Tree
         1. Gain Ratio:
         2. IV(a):
         3. Using Gain as measure will tend to choose an attribute with large V, for example ID.
         4. Still when using C4.5 tree, choose attributes with Gain higher than average Gain then choose the one with the highest Gain Ratio.
      5. CART Tree
         1. Gini Index:
         2. We choose the attribute which yields the least Gini Index after splitting.
   2. Explain the random forest model and compare it with the boosting model (GBT is more commonly tested)
      1. Bagging (Random Forest)
         1. Definition:
            1. Models run in parallel and are independent of each other.
            2. Training set is generating with selecting observations with replacement.
            3. Every model has the same weight.
      2. Boosting
         1. Definition:
            1. Models have different weight based on their performance.
         2. Gradient Boosting Decision Tree
            1. Building tree based on the residual calculated by subtracting predicted value from true value in the last decision tree.
   3. The adjustable parameters of the random forest and GBT models in the programming language
      1. Random Forest
         1. n-estimators: number of trees
         2. criterion: splitting criterion (Gini, Entropy, log\_loss)
         3. max\_depth: maximum depth of the tree
         4. bootstrap: If set to False, the whole dataset is used to build each tree
         5. min\_samples\_leaf: The minimum number of samples required to be a leaf node.
      2. GBT (Gradient Boosting Regressor)
         1. Loss: loss function (squared\_error, absolute\_error)
         2. Learning\_rate:
         3. n\_estimators: number of boosting stages
         4. criterion: splitting criterion (friedman\_mse, squared\_error, mse)
         5. max\_depth: maximum depth of the individual regression estimators
   4. You should know that each tree of random forest is better to make deeper because random forest is more suitable for low bias high variance; each tree of boosting model should not be too deep
   5. What is the most preferred model? Why?
   6. Know the advantages and disadvantages of each model, what is applicable, what data, complexity and computational effort related to what.
      1. Adaboost
         1. Definition:
            1. Every Leaner uses the same dataset
            2. Leaners are connected sequentially and are dependent on each other (same algo mostly)
            3. Focus on decreasing the Bias but sensitive variance (数据扰动造成的影响)
         2. Pros:
            1. Good at solving hard problems (high ceilings)
         3. Cons:
            1. Overfitting
            2. Too sensitive to outliers
            3. Slow
      2. Random Forest
         1. Definition
            1. Using Bootstrap Sampling (取出又放回) to get the training set for each learner. The size of training set is equal to that of dataset.
            2. Randomly choose k attributes from all the attributes and choose the best one (Gain/ Gain Ratio/ Gini Index) for the root node.
         2. Pros:
            1. Not likely to overfit because every learn uses different training set (not sensitive to outliers)
            2. Quicker when dealing with high dimensional data because only using some attributes
            3. Tree-structure is more easily to explain
         3. Cons:
            1. Too general (Low celling)
2. **KNN**
   1. Please explain KNN and then write out the code for its implementation
      1. The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group, or another based on what group the data points nearest to it belong to.
      2. K:
         1. Small k will result the noise will have a higher influence on the result.
         2. A large value makes it computationally expensive.
         3. Simple approach k = n^ (1/2)
3. **K-MEANS**
   1. Please explain K-means and then write out the code for its implementation
      1. K-means clustering is one of the simplest unsupervised machine learning algorithms
      2. Steps:
         1. Choose K
         2. Select K random points from dataset as centroids
         3. Assign each observation to the cluster with nearest mean (least squared Euclidean Distance).
         4. Update: Recalculate means for observations assigned to each cluster.
   2. How to choose k
      1. The Elbow Method
         1. Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.



* + - 1. The Silhouette Method
         1. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
         2. A high Silhouette Score is desirable and it reaches its global maximum at the optimal k
  1. How to measure the results (unsupervised learning, I guess interviews tend to want to hear some collaboration with domain people)
     1. Drawbacks
        1. Can’t handle complicated geometric shapes.

1. **SVM**
   1. Please explain SVM, (it seems that nay model may “explain the model”)
      1. The objective of SVM is to find a hyperplane that distinctly classifies the data points
      2. To find a plane that has the maximum margin
   2. What is support vector
      1. Data points that are closer to the hyperplane and influence the position and orientation of the hyperplane
   3. Explain the kernel trick, why it kernels matrix is positive definite
      1. Not all the data are linearly separable. Higher dimensional transformations can allow us to separate data in order to make classification predictions.
      2. However, when there are more and more dimensions, computations within the space become more and more expensive.
      3. Kernel Tricks: it allow us to operate in the original feature space without computing the coordinates of the data higher dimensional space. In essence, what the kernel trick does for us is to offer a more efficient and less expensive way to transform data into higher dimensions.
   4. What does the complexity of the SVM depend on, the sample size or the number of variables?
      1. Complexity: number of examples, number of features, type of kernel function and the regularization parameter.
   5. Explain some important parameters of SVM model
      1. C: Regularization parameter
      2. Kernel: Kernel method
      3. Tol: stopping criterion
2. **ML related algorithm implementation**
   1. Write a KNN algorithm
   2. Write a KMeans algorithm
   3. Write a mini-batch gradient descent function
3. **NLP related**
   1. According to the different interview positions, for some positions ML model is built on text data, so NLP can be added points
   2. Some basic concepts such as
      1. BOW model, N-gram model
      2. Term matrix, TFIDF
      3. Stemming, part-of-speech, NET
      4. Word2Vec, Topic modeling
   3. Computational linguistic, just ask how to design rules to extract company names given a dirty data
4. **Summary**
   1. Practice manual implementation of ML algorithms, KNN, K-Means better write and often test; Naive Byes can also write a logistic regression/ linear regression can also use gradient descent write, a friend was tested EM algorithm implementation.