Machine Learning

Ensemble
Ensemble means adding multiple models to create a stronger model

- GBDT
  - AdaBoost
  - GradientBoost
  - Random Forest
- Boosting(parellel)
- Baggin(sequential)

Regularization
Techniques to reduce errors by fitting fuctions on training set

- L1 Regularization — penalizes the sum of absolute values of the weights
- L2 Regularization — penalizes the sum of squares of the weights
- Dropout Regularization

Merit
Ways to measure model performance

- Confusion Matrix
  - True Positive
  - True Negative
  - False Positive
  - False Negative
- Accuracy — TP+TN/(TP+FP+FN+TN)
- Precision — TP/(TP+FP)
- Recall — TP/(TP+FN)
- Specificity — TN/(TN+FP)
- F1 — 2*(precision*recall)/(precision+recall)
- AUC — Recall vs 1-Specificity

Regression
technique for investigating the relationship between independent variables or features and a dependent variable or outcome

- Linear Regression
- Logistic Regression
- Stepwise Regression

Clustering
dividing the population or data points into a number of groups

- K-means
- K-median
- DBSCAN
- Expectation Maximization

Neural Networks
Need to understand math of forward/backward propagation

- Perception
- Back-propagation
- Forward-propagation

Classification
categorizes a set of data into classes.

- SVM
- Naive Bayes
- Decision Trees
- KNN

Dimensionality Reduction
the task of reducing the number of features in a dataset

- Principle Compoenent Analysis
- Principle Component Regresson
- Projection Pursuit

Ensemble Questions:
1. How does Random Forest avoids overfitting?
2. What are some differences betwwen boosting and bagging?
3. Describe the training process of GBDT.

Regularization Questions:
1. Explain L1 regularization and L2 regularization. What are the differences?
2. What is overfitting, and underfitting? How do we solve it?

Merit Questions:
1. Given a specific scenairo, what merits will you use and why?
2. What does recall and speicificity describe?
3. How to plot ROC curve?
4. What is confusion matrix?

Regression Questions:
1. What is the basic idea oflinear regression?
2. What are some loss functions commonly used in linear regression?
3. What is the difference between logistic regression and linear regression?
4. How does logistic regression handles multiple labels?

Clustering Questions:
1. Descibe how K-means build the model.
2. How to choose start point for clustering?
3. Are k-means sentitive to outliers? Why?
4. How to evaluate clustering effect?
5. When to use K-median instead of K-means?

Neural Network Questions:
1. List some activation functions used in neural network.
2. Does batch size effect converge? How?
3. How do you understand fine-tuning?

Classification Questions:
1. What are some advantages and disadvantages for KNN?
2. Why SVM can handle non-linear problems?
3. Describe the learning process of Naive Bayes.
4. How do we aviod overfitting in decison trees?
5. What are some commonly used decision tree methods?

Dimensionality Reduction Questions:
1. Why do we want to reduce dimensions? What problem does it solve?
2. How do we define principle components?
3. What are some limitations of PCA? How do we optimize it?