Group 1 - Deep Learning

- What is gradient descent and how is it applied to deep learning
  -
- Describe how the gradient descent process works
  - Current – gradient X learning rate (size of the features)
- What is stochastic gradient descent
  - Use a single random sample without replacement to calculate the gradient
- What is mini-batch stochastic gradient descent
  - A small number of batch records (random without replacement) 32 6*4*
- What are some advantages of stochastic gradient descent over non stochastic gradient descent
  - Computationally faster than the non stochastic, and easier to fit for the memory.
- How are partial derivatives used in gradient descent algorithm
  - Holding all other variables as constant and derivate one variable one at a time, so we update the slop of the variable one batch at a time…..
- What does it mean to flatten a matrix and how is this concept applied to deep learning
  - Turn the matrix of the original data space into one dimensional vector. We can apply it when processing the image or the voice data
- What is a hidden layer
  - Between the input and last layer,
  - We cannot interpret the meaning of the hidden layer.
  - Grid search to optimize.
- What are the capabilities of Neural Networks
  - They can approximate any kinds of functions (Linear or Non-linear
- What is forward and back propagation
  - Forward is feed the network with the input to get the output.
  - Backward is from the output to calculate the derivative slop based on the chain rule and propagate the error back to the previous layer.
- How are partial derivatives used in gradient descent computed at run time
  - Uses the chain rule to calculate the derivative for each layer.
- What is a partial derivative.
  - Holding all other variables constant, and calculate the derivative one feature at a time.

Group 2 - PCA

- What is a loading vector
  - Loading vector represents the direction that the data varies the most.
- What is a score
  - Multiply the loading vector with the record = the PCA score.
- What is a projection
  - It is a point on the loading vector
  - Multiplication of the loading vector with the training data.
- What is a scree plot
  - Amount of variance explain for each principal component
- Name 3 use cases for PCA in data science
  - Sentiment Analysis – Reduce training time.

  - Visualization the data
  - Break up the correlations among features
- Assuming a training data matrix of size 100x3, what is the size of the resulting loading vector matrix
  - 3 X 3
- How would we calculate the score for the 5th row of data using the 2nd principal component. What is the shape of the resulting score.
  - **Scaler? (Scaler + 2pca) 一层一层的**
- How can we perform feature selection using PCA
  - Look at the first pca and look at the feature with the largest coefficient
- What data pre-processing is required in order to perform PCA analysis
  - Standardize the data into the same scale.
- How can we get the original data back from PCA scores
  - Inverse transform – multiply the transpose of the loading vector times the score.
- Describe how PCA can be used to implement a recommender system
  - Reduce the dimension, and compute the Euclidean distance the make the recommender.

Group 3 - Logistic Regression / Regularization

- Describe how logistic regression works from a high level point of view
  - Use sigmoid function 0-1, uses the linear combination of the features
  - It is a sigmoid function that converts the combinations of the features into the probability, Plugin a linear regression into the sigmoid function
- What is the characteristic shape of a logistic regression function
  - S Shape
- Is logistic regression considered a linear model
  - Yes
  - It uses the linear combinations of the feature to calculate the
- How does the logistic regression training process affect the shape of the characteristic curve. In other words, how do the weights and intercept affect the shape of the curve.
  - Varying the slope of the yhat linear regression changes the slope of the sigmoid probability prediction.
  - Slope down, curve flatten, 形状
  - Varying the y intercept of the yhat linear regression changes the x axis bias location of the sigmoid probability prediction. 水平位移
- What is regularization
  - Minimize the larger weight, lower the high variance.
- How would you know when regularization is needed
  - Apply when overfitting
  - Larger accuracy in training set, lower accuracy in test set.
- How can regularization be used for feature selection in regression
  - L1 regularization can penalize the features to zero.
- How is L1 regularization calculated
  - Lasso
  - It adds the L1 term into the loss function
  - Summation of the absolute value of coefficient added to the loss function
  - Added summation times lambda back to the loss function make the function less complex

- How is L2 regularization calculated
    - Ridge
    - It adds the L2 term into the loss function
    -
- How is elastic net regularization calculated
    - Lambda times the penalty term
    - And the penalty term equals to the alpha times L1 plus 1- alpha times L2
- What is model flexibility in regression
    - How well the model fits for different training dataset.
    - Associated with the number of columns in the training data.
    - The number of the columns goes up, the flexibility of the model goes up
    - Model complexity increases, flexibility increases. (polynomial expansion)
    - Random Forest -> increase depth
- How can we increase / decrease model flexibility
    - DL -> increase the number of hidden layers or nodes in those layers.

Group 4 Assessing Model Accuracy

- What is generalization performance
    - How the model fits on the overall data(unseen test data).
- What is a loss function
    - Distance between true value and predicted values
- Describe the bias / variance trade off
    - If the model is more flexible it will have higher variance, lower bias(fits well on the training set, overfitting).
- What is bias
    - Underfitting, model not able to fit on the training dataset,
- What is variance
    - Overfitting, higher variable, fits well on the training set only.
    - Small changes on the training set leads to a larger variance on the test dataset./
- When is a model optimized in terms of bias and variance
    - Both bias and variance be minimized.
- How does model complexity / flexibility relate the the size of the training data
    - Fit the small data with a Complex model leads to overfitting
- How would you know if a model is overfitting
    - Low training error, high test error
- What is K-Fold cross validation and how does it work
    - Break the data into K fold, train the model using 1-k-1 fold and gauge the model using the k fold
- What is a confusion matrix and how is it used in the data science process
    - Evaluate the performance what we don right and wrong on the classification model.
- What is a ROC curve and how is it used in the data science process
    - It measures how well your model can address the true positive and true negative under all levels of thresholds.
    - Used for binary classification problem only
- How does a confusion matrix relate to the ROC curve
    - We use tpr and fpr from confusion matrix to construct roc curve.

Group 5 Statistical Learning

- What is statistical learning
  - Broad range of learning technic that ranges from supervised to unsupervised learning.
- What is reducible error
  - Can be reduced by optimizing the model
  - Least square error
- What is irreducible error
  - Cannot be reduced by the model
  - Error in the dataset or measuring error
  -
  - The patient's general health on a given day
  - Small manufacturing variations in the drug
  - Small measurement errors in x1, x2, x3, …, xN
- What is a parametric model
  - Estimating the function training model consists on some shape
  - Assume the estimating function in a shape
  -
  - Makes an assumption about the shape of *f*
  -
- What is a non-parametric model
  - Nonparametric methods don't make assumptions about *f*.(Linear or something)
  - Random Forest
- What is the notion of predictions
  - Prediction is use the estimation function to predict how that data would work.
- What is inference
  - Infer which feature is more important
- What is maximum likelihood estimation and how does it work
  - Estimate the probability and take the largest as the prediction.

Group 6 Other

- Describe TFIDF
  - Term frequency / Inversed ducuments frequency, reflects the frequency the word appears and idf reflects the degree of the importance of the word
- What is TF
  - Count the frequency the word shows up in the dataset, the more it appears in a document, the more important the word it is.
- What is IDF
  - The higher the word appears in the doc, the lower the idf value it will be.
- How is TFIDF used in machine learning
  - Token, natural language processing.
- Describe how TFIDF could be used to determine sentiment of text. Describe the entire process starting with a data frame containing text and walk through the steps needed to determine text sentiment.
  - Tokenizer
  - Remove the stop word
  - Standardize Scaling
  - Use logistic regression

    - The number of the columns > n of rows
- What is spark
    - A distributed application that has a machine learning ability
- What is Hadoop
    - A combination of hdfs system applies map reduce paradigm used on distributed system.
    - Distributed and replicated dataset, master slave. Computing and data tolorent
- What is a resilient distributed dataset
    - Rdd, back to python datafram
    - Immutable data type manage data type across the distributed system.
- What is a lineage graph
    - A dag
    - A set of instructions of how to build RDD
- Why does spark use lazy execution
    - Save computing resources.
    - Optimize the performance.
- What major improvement does spark have over Hadoop
    - DataFrame kit
    - Spark does computing in memory, Hadoop computes using hard drive.

Group 7 Random Forest

- Describe how random forest models work from a high level point of view.
    - Majority vote, do not use all features
- What is the difference between random forest and regular decision trees.
    - Decision Tree uses all,
    - Random Forest, select features ,average result
- How does random forest prevent high correlation between trees in the forest
    - Select features, averages the trees to decorrelation
- Describe the training process.
    - Random forest uses bagging, sample with replacement for each tree. Having low variance.
- Describe one or more similarities between random forest and GBM
    - Select Features
- How does the number of trees in the forest relate to variance in the predicted value
    - Reduce variance by increase the number of trees
- How does the tree depth relate to model bias / variance
    - Higher the tree, lesser the bias,
- What is entropy
    - Entropy entry of randomness
- What is gini index
    - Gini Level of impurity
- What is information gain and how does it relate to relate to the training process
    - The level of randomness reduced by a particular split
    - Select the cut that maximize the information gain

Group 8 GBM

- Describe boosting from a high level point of view

    - Iteratively train and combine in a sequence.
- Describe how gradient boosting machines (GBM) work from a high level point of view.
    - Iteration calculates the gradient in order to reduce the loss function, calculate it and move the coefficient on to that direction.
    - Give incorrect cases more weights
    - Passing the residual to the model, and try to reduce the residual.
- Describe some differences between random forest and GBM
    - Random forest can run parallelly, GBM is a serials process
    - Random Forest using bagging
    - Random forest vote for the majority. GBM focus on error
- Describe some similarities between random forest and GBM
    - They all select some features of the subset.
- Describe some characteristics of individual trees in a GBM forest
    - GBM uses small trees(weak learner).
    - 3-4 trees < 10 should be a good number of trees.
- Describe the GBM training process
    - Calculate the residual from the previous tree and feed it into the next tree.
- How does GBM prevent correlation between trees
    - It randomly selects features
    - N = sqrt (m)
- How does the learning rate in GBM relate to the number of trees in the forest
    - If you need a higher speed of training time, you can use a higher learning rate and smaller number of trees.
    - The learning rate represents the fraction of the residual passed between the training models
    - Should have larger number of trees if your learning rate is small.
- How does tree depth relate to model accuracy in GBM
    - Smaller the depth, the less the accuracy will be. (Only for random forest), for GBM, increases the depth of the tree will leads the model works bad. The good learner cannot agree with each other and the result falls apart 1:32:25
    - Smaller the depth, the less complexity the model it will be.

Group 9: Bootstrap Sampling / Bagging

- Describe Bootstrap sampling from a high level point of view
    - Bootstrap sample first, pick up few columns, use the sample to build one tree. And same process to build all other trees(This is for the a Random forest), and use the majority vote as the result.
    - Bagging aggregates the result which trained based on the bootstrap sampling.
- Describe bagging from a high level point of view
    - Bagging aggregates the result which trained based on the bootstrap sampling.
- What is the purpose of bootstrap sampling
    - Reduce the variance
    - Create more data
    - Maybe biased(So sample the columns also)
    -
- Is it possible to have duplicate observations in a bootstrap sample
    - Yes
- What is an out of bag sample
    - The records that have not been selected. (33%)

- Describe how out of bag samples are useful in the data science work flow
  - Used as the test dataset
- What are some important details concerning the bootstrap sampling process.
  - CLT
  - Randomly sample the columns of the sample to reduce the variance.
  - The size of the bootstrap sampling is the same with the original data.
- Describe a specific data science situation where bootstrap sampling could be useful
  - When you only have a small number of data