Data Science

p-value : probability of obtaining a sample at least as extreme as ours if the null hypothesis were true.

- Cover the basic distributions - Bayesian, Gaussian, binomial, Poisson, chi-squared, exponential (these latter two are both just special cases of gamma).
- Confidence intervals, summary statistics, p values, and moments of distributions. The central limit theorem.
- Hypothesis testing using both parametric and nonparametric approaches. Bootstrapping, permutation tests, and Monte Carlo simulations.
- Mathematical Biostatistics Bootcamp for better understanding of distributions and hypothesis testing.

Machine Learning

Supervised :

- Regression (continuous valued output) A **regression** model predicts continuous values. For example, regression models make predictions that answer questions like the following:
  What is the value of a house in California?
  What is the probability that a user will click on this ad?
- Classification (Discrete valued output) A **classification** model predicts discrete values. For example, classification models make predictions that answer questions like the following:
  Is a given email message spam or not spam?
  Is this an image of a dog, a cat, or a hamster

Training vs testing data set(the more data the better for both) -> If the training data is less, we can use cross validation.

- SVM to deal with infinite number of features.

  Linear Regression

  1. Model and cost function
  2. Gradient descent (need to choose alpha, needs many iterations, works for large n) Feature scaling (n) :
     1. (between -1 <x<1) to converge quickly for gradient descent
     2. mean normalization.
  3. Normal Equation = (X^TX)^-1 X^T y (no need of alpha ,don't need to iterate, slow if n is large < 10,000)
     1. if X^TX is non invertible -> check for redundant features, too many features -delete, use regularization.
  4. Multivariate Linear Regression
  5. Multiple features
  6. Polynomial regression

Logistic Regression

1. Higher order polynomial functions
2. multi class classification : one-vs-all. Evaluate models using
    1. accuracy = Fraction of prediction we got right = TP+TN/ TP+FP+FN+TN
    2. precision = True positive / all positive predictions = tp/tp+fp
    3. recall = True positive / all actual positives = tp/tp+fn
3. How to avoid Overfitting ?

Regularization

1. Early stopping before converging
2. penalizing the parameters (logistic and linear)
3. L1
4. L2
5. Prediction Bias ? Average of predictions = Average of observations.

Linear vs Logistic Regression

Debugging learning algorithms:

Regularized linear regression, during testing hypothesis on new set of samples, there is large errors in predictions, what to do?

1. High variance (overfit):
    1. Get more training examples
    2. try smaller sets of features
    3. try increasing lambda
2. High bias (underfit):
    1. try additional sets of features
    2. try additional polynomial features
    3. try decreasing lambda

Evaluate a hypothesis -> cross validation.

Bias(underfit) vs variance(overfit) :

Bias: Training and CV errors both will be high

Variance: Training error is low and CV error is high

Bias/variance as a function of Regularization

Learning curves

High bias: more training data won't help

High variance: more training likely to help.

Recommended approach for new models

1. start by simple algorithm, and test it on CV data
2. plot learning curves to decide if we need more data, more features
3. error analysis: manually examine examples(on CV set) and manually spot any systematic trend. e.g: stemming for spam filters.
4. error metrics for skewed classes : good algorithm has high precision & recall.
    1. Precision : TP/ TP+FP
    2. Recall : TP/TP+FN if recall is zero -> data is skewed
    3. F1 score: 2PR/P+R
5. Large data sets - with lot of parameters , if human can predict y based on x?

SVM (support vector machines)

1. large margin classifier: maximize margin between closest support vectors.
2. SVM is deterministic and is faster for kernel space.
3. kernels:
    1. Linear kernels
    2. Gaussian kernel : do perform feature scaling before using it.

Logistic Regression vs SVMs

Unsupervised :

Clustering, non-clustering.
SVD to de clutter two voices in cocktail party problem.

K-means algorithm

1. choosing k - usually lower cost function for more k, but if k=3 > k=5 -> most likely k-mean was struck at bad local minima, so re-running will help.
2. **PCA - Principal component analysis**
    1. to reduce dimensions & speed up learning process
    2. used to compress data & visualize high dimensionality data.
    3. Don't use PCA to prevent overfitting by reducing dimension, use regularization.

Anomaly detection $p(x) < E$

Questions

1. How do you choose n grams? why choose 2 vs 5 and their advantages/ disadvantages?
2. Models performs poorly on test data why? what happens when n grams are 0 for testing data? --> borrow prior data from other similar data set/ generate synthetic data.
3. What is CNN and logistic regression?