Sample Questions

True/False

Pruning is a technique for reducing complexity in Decision Trees. True

In general, one will always get the best performance out of Decision Trees when one does not limit their size False

In a classification tree induction, the next attribute added is the one with the largest information gain. True

Support-Vector Machines (SVMs) approach classification problems by finding the widest possible bar that fits between points of two different classes. True

Logistic Regression and Support Vector Machines are both learned by minimizing a loss function induced on a training data set. True

In K-Nearest Neighbors and K-Means Clustering, the parameter “K” represents the same concept. False

Selection Bias exists when P(In Sample | X = x) = P(In Sample), for any arbitrary X = x. False

Regularization in SVMs and Logistic Regression limits the size of the weight vector that can be learned. True

Engineering features by “binning” numerical variables into categorical bins is always best when one has a small sample size. False

L1 – Regularization is an effective way to reduce the feature size in a linear model. True

When comparing algorithms and/or hyper-parameters to perform model selection on a dataset, one should always use different validation sets to avoid biasing the results. False

Given the same data set, a model with an accuracy score of 80% is always better than a model with an AUC of 0.70. False

Any model that produces an accuracy of 50% on any given dataset is no better than choosing labels randomly. False

A model with a lot of regularization is likely to under-fit the data. True

L2 Regularization assumes a Gaussian prior distribution for the parameters in the weight vector being learned. True

Naïve Bayes can not be considered a linear model. False

**(BD Need 4 more)**

**Multiple Choice**

In order for data points to be taken as input to most data mining programs, they must be represented as:

* + a)  text documents
  + b)  feature vectors
  + c)  dependent variables
  + d)  targets

Regression is distinguished from classification by:

* + a)  discrete class probability estimation
  + b)  numerical attributes
  + c)  numerical target variable
  + d)  hypothesis testing

Entropy

* + a)  is a measure of information gain
  + b)  is used to calculate information gain
  + c)  is a measure of correlation between numeric variables
  + d)  has a strong odor

Which of the following does not describe SVM (support vector machine)?

* + a)  SVM can estimate class membership probability
  + b)  SVMs are based on supervised learning
  + c)  SVM chooses the line to maximize the margin between two classes
  + d)  SVM can be applied when the data are not linearly separable

Which of the following is not true about logistic regression:

* a)  Logistic regression predicts probability of membership in a certain class.
* b)  Logistic regression takes a categorical target variable in training data.
* c)  A logistic regression represents the odds of class membership as a linear function of the attributes.
* d)  Logistic regression requires numeric attributes and categorical attributes should be converted to numeric attributes.

A learning curve aims to test:

* a)  True positive rate vs. false positive rate
* b)  True positive rate vs. false negative rate
* c)  Generalization performance vs. size of training set
* d)  Generalization performance vs. alternate feature transformations

The area under the ROC curve is not

* + a)  equal to the Mann-Whitney-Wilcoxon statistic
  + b)  a measure of the quality of a model’s probability estimates
  + c)  likely to be at least 0.5
  + d)  larger when false positive errors cost more

The points on a model’s ROC curve

* + a)  represent the performance at different thresholds
  + b)  represent different rankings of examples
  + c)  represent the cost of different classifications

Unsupervised data mining

* + a)  Usually involves some form of Regression estimation
  + b)  Is easier to evaluate than supervised data mining
  + c)  Cannot be applied if we have a well-defined target variable
  + d)  Is best evaluated by incorporating domain knowledge

Similarity measures are most essential for

* + a)  Naïve Bayes
  + b)  Tree Induction
  + c)  Hierarchical Clustering
  + d)  Logistic Regression

Which is not true of clustering?

* + a)  Centroid-based clustering is the procedure that all observations start in one cluster, and  splits are performed recursively as one moves down the hierarchy
  + b)  Domain knowledge can be incorporated
  + c)  The “K-means” algorithm is an iterative center-based algorithm
  + d)  It is used to group objects represented by multiple attributes

Which is not true of k-Nearest Neighbor (k-NN)?

* + a)  The choice of k needs to be carefully tuned
  + b)  It builds a simple induction model
  + c)  It is fit by minimizing a loss function
  + d)  It is easy to explain how it works

Which technique divides the population into disjoint segments described by conjunctions of feature conditions?

* + a)  Naïve Bayes
  + b)  Tree induction
  + c)  Logistic Regression
  + d)  k-NN

Which data mining technology would be most useful in answering the following business question? “Of all my accounts, which are the most likely to exhibit fraud, based on my experience with prior cases of accounts that have and have not been defrauded?”

* a)  Classification tree induction
* b)  Hierarchical clustering
* c)  k-Means
* d)  Linear regression

Which business intelligence technology would be most useful in answering the following business question? “If this customer responds to my offer, how much will she spend?”

* + a)  Classification tree induction
  + b)  Hierarchical clustering
  + c)  k-Means
  + d)  Linear regression

**Short Answer**

**Please answer each of the following in at least 2 – 3 sentences.**

* 1)  You would like to build a model for predicting the likelihood of default on student loans. You are given a large number of categorical attributes of each loan such as the type of the school that a student is going to attend, the state where it is located etc., as well as numerical attributes such as outstanding loan amount, student’s age, loan interest rate and so on. Your client asks that your model must provide a clear explanation of the reason for its predictions, since the final judgment on whether to give a loan or not will be made by a human agent. What data mining technique would you suggest using? Explain why briefly (one or two sentences).

The key consideration here is that the models need to be interpretable and provide probability estimates. Our best option might be Decision Trees, which are both interpretable and can natively handle categorical features. We could also use a logistic regression after the categorical features have been converted to numeric attributes.

* 2)  I want to build a logistic regression, but I have a very large number of possible attributes describing my instances. List two reasons why might this be a problem? How should I deal with the problem? (Two sentences for each question)

This might be a problem because the likelihood to overfit increases with more features, and training time increases with more features.

For the overfitting problem, either feature selection or regularization can help reduce the likelihood to overfit. For the scalability issue, feature selection would be our best bet.

* 3)  What exactly does the area under the ROC curve represent? Be as precise as possible.

The area under the ROC represents the probability that a randomly selected positive instance will have a higher model score than a randomly selected negative instance.

* 4)  In order to apply data mining to text documents, we need to represent text documents by some term (word) based features/attributes. Please describe three common approaches to create attributes for text documents. Please give a clear explanation of all your terminologies.

Three common approaches that we can use here are: 1). Represent each word as its own binary categorical feature, 2). Extend the previous method by using the frequency of the term in a given document as the feature value, 3). Extend the previous method by weighting the per document term frequency with the inverse count of the term’s occurrence in all documents (this is called tf-idf).

We can additionally extend any of the above by using N-Grams instead of terms.

* 5)   Why does a Random Forest generally perform better classification than a single Decision Tree? Answer this within the framework of Bias and Variance.

Decision Trees can suffer from high variance because they use a greedy algorithm that often settles on local optima. The Random Forest fixes this problem by using a technique called Bagging. Bagging is well known to reduce the variance of an estimator. With this decrease in variance though comes a slight increase in bias. The RF bias is caused by the random feature selection in individual trees. Usually though the decrease in variance outweighs the increase in bias, which leads to better performance.

* 6) Distance is a key notion underlying many data mining algorithms, such as k-nearest neighbor (k-NN) and clustering. When using Euclidean distance, what problems arise from the following two scenarios: 1). Different features on are different scales (i.e., income vs. age), 2). The number of features increases.

Euclidean distance is not invariant to scale, so any feature that has a higher scale (which we can define as being proportional to the variance), will dominate the distance measure. As an example, ED gives equal distance to Age = 20 vs 40 and Income = 100, 000 vs 100, 020. We can assume generally that the age difference here is more meaningful than the income difference. Rescaling the variables would solve this issue.

Euclidean Distance also suffers from the curse of dimensionality. Average pairwise distances tend to increase with the number of features, meaning that all instances are far away from each other in a P-dimensional space. This could reduce the efficacy of kNN classifiers or make clusters less meaningful and difficult to interpret.

* 7)   For kNN, explain the effects of choosing a high vs low value for k using the bias-variance tradeoff framework. How would you choose k?

A low value of k means that only a few instances contribute to a particular instance’s classification. Such a case would be more sensitive to noise and thus would likely be high variance. Conversely, large values of k use more instances, most of which would be farther away from the instance being classified. We’d expect a large k to produce a low variance but high bias estimator. Choosing the right value of k becomes an empirical question. Just like any in other hyper-parameter selection, we could use any type of out-of-sample validation technique to test the out-of-sample error for a range of values of k. We would then choose the k with the best out-of-sample error.

* 8) If you were building a classifier for a cancer screening application, why might you choose Recall as your model selection and evaluation metric?

Many classification tasks have different costs for false positives and false negatives. In medical screening, a false negative might mean we call a person healthy when they are indeed sick. The consequence of such a mis-classification could be an escalation of the illness, with the possibility of death. If a false positive exists, we might just refer the patient to get a more in-depth, secondary screening. With this scenario, we want to make as few false negatives as possible (despite allowing for more false positives). A high recall means a low false negative rate, so Recall would be an appropriate metric here.

**Longer form questions**

**Please answer each of the following in 2 – 3 paragraphs.**

What are two reasons why we might care that selection bias exists in our modeling data set? Describe a way that a production system that deploys a model might induce selection bias in the firm’s data stream. As a data scientist working on such a system, how might we reduce the effects of selection bias or avoid it altogether? When answering this last question, use the example you provided in the middle question as a working example.

The two concerns with selection bias are: 1) Generalizability – if our sample is biased we can’t prove that our results generalize to the population at large, 2). Identifiability – If we want to estimate E[Y|X = x] for a given X = x, but the data does not contain any support for X = x, then we can not estimate this parameter.

In general, selection bias often appears in production systems because we intentionally apply the results of predictions to a non-random subset of the population. Examples we discussed in class include: 1). Credit applications – we only accept people above a certain credit score, 2). Spam detection – we only allow emails with low probability of spam into the inbox. The main issue with both of these systems, is we do not often observe the outcome for examples that do not pass our production ‘filter.’ Thus, we can not estimate E[Y|X=x] for these examples.

The simplest way to alleviate the problems induced by selection bias would be to allow for some randomness in our application filter. For instance, we can randomly accept a portion of credit applicants below the threshold to be able to learn what their default rates actually would be. For SPAM, we could randomly allow some classified SPAM into the inbox so that we can prompt users to add labels. A smarter way to do the randomization would be to allow more instances past the filter that are closer to the decision threshold.

Let’s assume we have 3 time period T1-T3, ordered as follows:

Potential Training Data Available

**T1 (w/ TV Campaign)**

**T2 (No Campaign)**

**T3 (No Campaign)**

The end of T2 marks the present. You have been tasked to build a model that predicts P(Y|X) (don’t worry about what Y and X are right now), which will be applied in T3. You know from your marketing department that a major TV campaign ran for your company in T1 but not in T2. There is no plan to run one in T3. As the data scientist, you need to decide how much data to pull as a training set.

Knowing that the data from T1 and T2 differ because of the TV campaign, how will you decided what data to use? Propose a test/experimental design so that you can learn whether or not to use data from T1 (note: do this without using cross validation). Be specific about which data will be used for training and validation.

Explain the pros and cons of using just T2 or using both T1+T2 as your training set in terms of bias-variance tradeoffs.

We know that in general we want to use as much data to train as possible. However, in our case here, we know that data from T1 is potentially biased because of the TV campaign. We can assume that data from the period T2 is distributed more similarly to the application period T3. The tradeoff here is that by using T1, we have more data and more data reduces model estimation variance. However, using T1 might induce a bias in our model, as the relationship between Y and X in T1 will likely be different than the relationship between Y and X in T2. The only way to know how this bias-variance tradeoff will balance will be to experiment.

For our experiment, we can set up our data as follows:

Validation – we’ll use the last 10-20% of T2 where ordered by time. At the very least, our validation set must come from T2, because we want out-of-sample data that is as similar in distribution to T3 as possible.

Training set 1 – we’ll use the remainder of T2 (after taking out validation data)

Training set 2 – the remainder of T2 and all of T2.

We can then build models using the above two training sets and then evaluate on the validation set. We can then select the training set which leads to the best error on the validation set.

The key here is that our validation data needs to represent the application data (T3) as best as possible.