TA: Abtin Shahidi **Final exam**

PHYS 243

06/08/2019, 8:00 am

Student: Henry Yung

# **Part 1: (10 points)**

1. Explain Monte Carlo simulation with an example for it’s application. (5)

The Monte Carlo is a statistical technique used to model stochastic or probabilistic systems and to find the probability of the outcome. We have used several examples throughout the class. My favorite example of the Monte Carlo simulation was the class assignment of the buffon needle. We created a Monte Carlo simulation where we wanted to find the probability of a long or short needle crossing the line by throwing the respective needle N-times.

1. How can you sample from a given probability density function *pdf* ? (5)

There are two ways to sample a pdf.

Simple random sampling, in which every member of the population has an equal chance of being selected, and selecting one member must not increase or decrease the chance of another member. This does not work well for small populations.

Another approach is stratified random sampling, in which the population is divided into different strata and a random sample is taken from each strata in proportion to the total population. This ensures a representative sample.

# **Part 2: (25 points)**

1. Explain K-means clustering, Hierarchical clustering and explain a scenario in which you prefer Hierarchical to K-means and vice versa. (10)

K-Means is a clustering method where the dataset is divided into K-clusters. The users specify the number of clusters that need to be grouped into a dataset. The goal is to find the centroid for each cluster where all the data points are assigned to it. The centroid represents the characteristic of all the data within the cluster.

Hierarchical clustering is a clustering method where the order of clusters is built by two common methods. The two commons are Agglomerative and Divisive. Agglomerative clusters start with each datapoint starting in its own clusters and then pair of clusters are merged as one as it moves up the hierarchy. Divisive is the opposite where all datapoints start as one cluster and then divided and split as it moves down the hierarchy. The ending results create a dendrogram where the user can choose the number of clusters based on preference.

Hierarchical is a greedy algorithm and can be very slow and take a long time to implement. Hierarchical clustering is preferred when you have absolutely no preexisting knowledge of the data and wish to find similarities. K-means is preferred when you know how many clusters you want and is better since it is faster.

1. Explain K-Nearest Neighbors and how would you make decision on what K to choose.(5)

The k-nearest Neighbors is a non-parametric learning algorithm used for classification and regression. It uses a data set in which the data are separated into different classes and predict the class of a new point. The classes are usually identified by a distance calculation. (Ex: Minkowski or Euclidean)

We can use an evaluation technique to validate the appropriate number of Ks to choose because increasing the number of neighbors does not always increase our performance. It can actually make the prediction worse! We would choose the appropriate number of K’s that gives us the optimal performance on our model.

1. Explain Decision Trees, and Random Forest. (5)

Decision Tree is a classification technique. It is a flowchart-like structure is based on information theory where the tree is split recursively until there is a single node.

Random Forest – Random forest is ensemble learning method where it generates a series of decision tree models and then making a decision by vote or average to determine the best output.

1. Explain Kernel methods, Support Vector Machines and explain the difference between linear and Gaussian kernels. (5)

Kernal method - A function that takes low dimension input space and transfer them to a higher dimension space. The advantage of this function is to take problems that were not separable in linear space and take them to higher dimension where they can separated and grouped.

Support Vector Machine uses to the kernel function in its algorithm to classify problems.

Linear Kernels – It is measures two features and their closeness by taking their inner product with some constant.

Gaussian Kernel - It is measure that can take two examples of features and multiply the features of matrixes together through dot product. The features are similar if the value will be close to 1 and they are not if they are close to zero.

# **Part 3: (15 points)**

1. Explain ways you can evaluate the performance of a certain model. (5)

You can evaluate the performance through several methods either through cross validation, confusion matrix, and accuracy and precision score. Ultimately, it to find how well your training set performs to the test set. Hence, the training set is always larger.

1. Explain overfitting and how can you avoid that. (5)

Overfitting happens when a model picks up on the noise and random fluctuations in a dataset and attempts to apply that as a concept of the model when it does not actually apply and thus negatively impacts the model’s ability to generalize to new data. This is more likely with models that have more flexibility in learning a target function, such as nonparametric and nonlinear models. Overfitting can happen if the problem is insufficiently constrained or if the model is so powerful it fits the data as well as the noise in the data.

In order to avoid overfitting. We need to add prior knowledge and be able to handle uncertainty. There is usually prior knowledge we can leverage, such as assuming a function to be smooth, which reduces the model complexity. Regularization can also be added which prefers smooth models. To reduce uncertainty, we can use a resampling technique to estimate the model accuracy or hold back a validation dataset.

1. Explain cross-validation. (5)

Cross-validation is the process of evaluating a model by testing it on a new data set, also known as a test set. We do this since fitting the data as closely as possible does not mean that it will be able to generalize to new data well, and a useful model should be predictive in new situations. In addition, empirical estimates can be obtained which are useful in identifying errors in implementation, comparing competing models, as well as detecting overfitting or underfitting

# **Part 4: (50 points)**

Imagine that you have a dataset (100 features, 1000 instances) with both categorical and real valued data and they have been labeled with a categorical value. Also, your dataset contains few missing data. Your task here is to classify the labels. Explain with details the procedures you take to find a classifier with two different algorithms of your choice.

This is a classification issue.

1. Step 1: Gather Data - I would gather the dataset and load them into the program.
2. Step 2: Discover and Visualize the dataset for data analysis.

* In this step, I would try to find some basic statistic to understand the dataset better. I would look for things like mean, mode, standard deviations and especially outliers to identify gaps in the dataset. I would also take this time to understand correlation within the features.

1. Step 3: Prepare the data

* This step is crucial as there categorial data and missing dataset. The first step is addressing the missing data points. Based on Step 2 of the Dataset set, I would make a determination on how to handle the missing data. Can I use an appropriate replacement for the dataset such as a median or average value? Would it be appropriate to remove these datasets and how will they affect my results? After, I handle the missing values, I must normalize the dataset. This includes converting the categorical data into values that can used in the machine learning algorithms. Afterwards, I would try to reduce the number of features through PCA. The goal of PCA is to reduce the dimensionality while losing little information. Of course, in these steps, I would split the training and test dataset.

1. Step 4: Select a classifier with two different algorithms

* This is a classification issue. Based on the number of features and issues to resolve, I would choose random forest algorithm and SVM. The random forest algorithm is great classification technique because it is ensemble technique that on bunch of random forest to give a classification. SVM is always a good algorithm as it can handle numerous features efficiently. Finally, these algorithms tend to avoid overfitting.

1. Evaluating and fine tuning the model.

* In this stage, we evaluate and fine tune the model. In this step, I would run some evaluation measure to see how the model is performing. After, I would run cross-validation or grid search to find the optimal values for fine tune my algorithms. Finally, I will run my evaluation scores to ensure I classified the labels correctly and then summarize my results.

1