Abhishek Deo GTID: 902866240

CS 4641

Homework 1

Supervised Learning Analysis

Datasets:

- UCI Seismic-Bumps Dataset:
    - The seismic-bump dataset contains information regarding seismic hazard frequency in underground coal mining facilities. There are 19 attributes that determine a seismic occurrence which all relate to the intrinsic properties of the hazard. The attributes consist of binary values and numerical values. There are 2584 instances in the dataset that represent seismic hazard occurrences. The data set is characterized by unbalanced distribution of positive and negative examples where there are only 170 positive instances and the rest are negative. The distinction between positive and negative data in this context is whether or not seismic activity occurred based on the values of the attributes.
- UCI Solar Flare Dataset:
    - The solar flare dataset contains information regarding the number of solar flares for each class that occur in a 24 hour period. The database contains 3 potential classes of solar flares where each one is for the number of times a certain type of solar flare occurred in a 24 hour period. The three classes of solar flares are called C-class, M-class, and X-class solar flares which each represent the number of common flares, moderate flares, and severe flares respectively. Each instance of a solar flare represents captured features for one active region on the sun. There are 10 attributes that determine the occurrence of a flare. The attributes consist of nominal values and binary values. There are 323 instances that represents solar flare occurrences within the three classes.

Interest: Both datasets are interesting because they both contain recordings of natural instances. Both seismic activity and solar flares can potentially be unpredictable which makes it difficult to provide insight on predicting their next occurrence. The datasets are split 70% for training and 30% for testing respectively after being randomized. Finally, both datasets have three classifications and a binary value as an end result that determines whether or not the event occurred.

- UCI Seismic-Bumps Dataset:
    - The seismic-bumps dataset provides knowledge on various scenarios of possible seismic activity. The data is based on 19 common attributes which measure the consequential impact of seismic activity. Even though they happen frequently, any occurrence of a moderate to large event is sudden and unpredictable. Most

seismic activity are natural, dynamic processes that shape Earth's landscape and are caused by events such as sudden slips on fault lines. This dataset is interesting because it helps narrow done and predict when even moderate to large level events of seismic activity could possibly occur based on the attributes.

- UCI Solar Flare Dataset:
  - The solar flare dataset is more or less similar to the seismic dataset in the sense that it is naturally occurring recordings. While solar flares occur on the sun, they emit many forms of radiation which are harmful to life and can easily reach Earth. Solar flares are also more difficult and unpredictable compared to seismic activity which makes it an even more interesting dataset. Based on the attributes, this dataset helps narrow down future predictions of moderate to severe level solar flares. While larger flares are less frequent than smaller solar flares, they are more unpredictable and dangerous.
- Machine Learning Viewpoint:
  - From a machine learning perspective both datasets are interesting for the same reason but have subtle differences based on the impact of the natural activity. Seismic activity, while more direct and frequent are less harmful than solar flares which is more indirect and infrequent. Because of the number of instances on the solar flares due to how infrequent they are or how difficult they are to record, the solar flare dataset had a relatively easy time on the algorithms compared to seismic activity. **<Talk about over-fitting and conclusion later>**

Decision Tress (J48 Algorithm):

Seismic Activity:

Pruned	Confidence	Training	Testing	Leaves	Size	Time
Yes	0.1	93.3665%	93.5484%	1	1	0.01 s
Yes	0.2	93.3665%	93.2903%	1	1	0.01 s
Yes	0.3	91.9292%	92.6452%	13	23	0.01 s
Yes	0.4	91.7081%	92.1290%	20	35	0.01 s
Yes	0.5	91.9845%	91.8710%	54	93	0.03 s
No	x	91.2106%	90.4516%	64	107	0.03 s

Solar Flares:

Pruned	Confidence	Training	Testing	Leaves	Size	Time
Yes	0.1	95.5882%	96.9072 %	1	1	0 s
Yes	0.2	95.5882%	96.9072 %	1	1	0 s
Yes	0.3	95.5882%	96.9072 %	1	1	0 s
Yes	0.4	95.5882%	96.9072 %	1	1	0 s
Yes	0.5	95.5882%	96.9072 %	1	1	0 s
No	x	95.5882%	96.9072 %	5	7	0 s

In the seismic activity dataset, both the testing and training data percentage decrease slowly where the training data decreases at a faster rate. Since the training data is decreasing at a slightly faster rate than the testing data there are some signs of overfitting. It's possible that this is a consequence of increased pruning working to remove some of the overfitting that occurred in the algorithm by removing the branches that memorized the data. Compared to the solar flare dataset, it performed significantly worst. The seismic activity dataset has twice as many attributes which contributes to the performance quality and overfitting quantity. In addition, since decision trees rely on information gain per attribute, there is not much information gained for the seismic activity because the attributes all influence each. Considering that the attributes consist of mostly binary values and numeric values that all are dependent of each other as well as the fact that the data is unbalanced, it is no surprise that the seismic dataset performed poorly. (Unbalanced distribution of positive and negative instances)

Consequently, the solar flare dataset performed the same on both the training and testing dataset. The size of the tree and pruning remained to be one is no surprise. There are fewer instances and most or all of them can be determined right from the start. In addition, some of the attributes that determine the occurrence and classification of a solar flare are time based. Specifically if a subsequent solar flare occurs within 24 hours of the previous solar flare in the same active region on the sun of which there are many. This attribute is very influential in determining whether or not a solar flare will occur or not. There is no overfitting as both the testing and training data are constant. Overall, the seismic activity was more difficult on the algorithms compared to solar flares because of the intrinsic properties of the dataset despite the end result.

Boosting (Adaboost with J48):

Seismic Activity:

Pruned	Confidence	Testing	Training	Leaves	Size	Iterations	Time
Yes	0.1	92.6471%	95.8763%	7	9	10	0 s
Yes	0.2	94.1176%	94.8454%	5	7	10	0 s
Yes	0.3	94.1176%	94.8454%	5	7	10	0 s
Yes	0.4	94.1176%	94.8454%	5	7	10	0 s
Yes	0.5	94.1176%	94.8454%	7	10	10	0 s
No	x	92.6471%	96.9072%	8	12	10	0 s

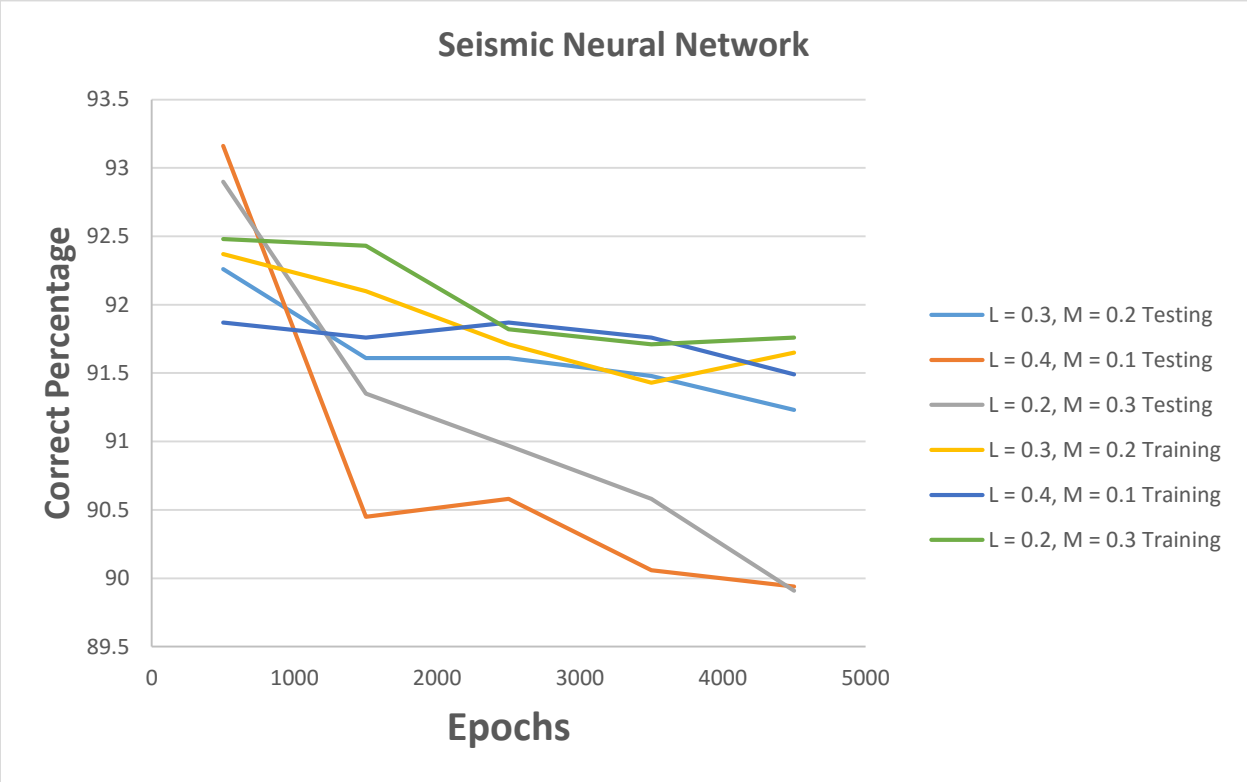
Solar Flares:

Pruned	Confidence	Testing	Training	Leaves	Size	Iterations	Time
Yes	0.1	92.6471%	95.8763%	4	6	10	0 s
Yes	0.2	94.1176%	94.8454%	12	16	10	0 s
Yes	0.3	94.1176%	94.8454%	12	16	10	0 s
Yes	0.4	94.1176%	94.8454%	12	16	10	0 s
Yes	0.5	94.1176%	94.8454%	12	16	10	0 s
No	x	92.6471%	96.9072%	17	23	10	0 s

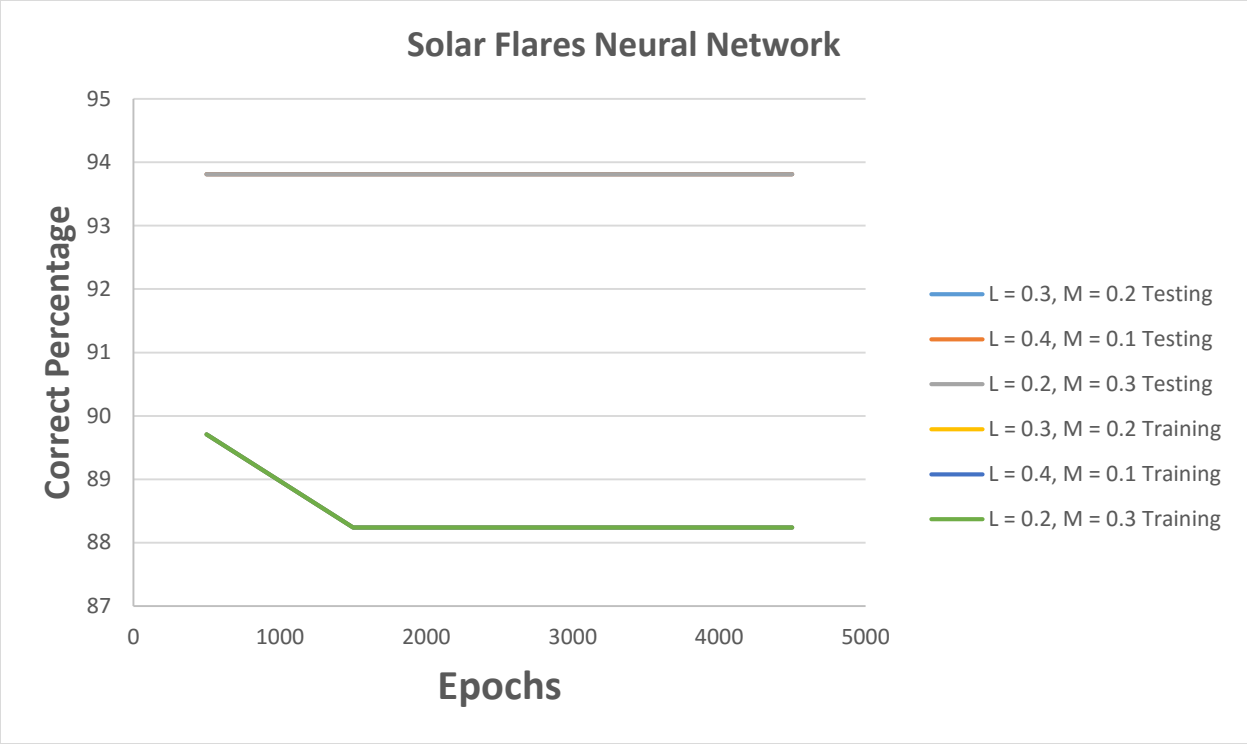
Boosting algorithm was run by using Adaboost with the J48 algorithm. The results from boosting for both datasets improved. Both the testing and training results for seismic activity have a pretty similar pattern and overfitting has decreased. Boosting allowed the number of attributes for the seismic activity dataset allowed the algorithm to separate the data points in a relatively simpler matter. Consequently, solar flares also improved a decent amount albeit its original performance on the decision tree without boosting. Solar flare results for the training portion increased by roughly 1 percent across the board. The training data became more consistent across variations in confidence levels and improved by roughly 4 to 5 percent. For both cases, there were many overlap occurrences in the dataset itself which caused most of these changes. Specifically for seismic activity where many of the attributes are similar to each other so they overlap quite easily. Overlapping causes some fluctuations with the results but that is all dependent on the attributes.

I believe for boosting significantly helped the seismic activity dataset over the solar flares dataset. This is to be expected because of the number of attributes the seismic activity dataset has to begin with. The original decision tree results without boosting ended up being poor due to overfitting. Even though there were binary values among the attributes, boosting the classifier algorithm removed some of the error. Error being decisions in the tree that may have not been picked up from the original classifier algorithm.

Neural Networks (Multilayer Perception):

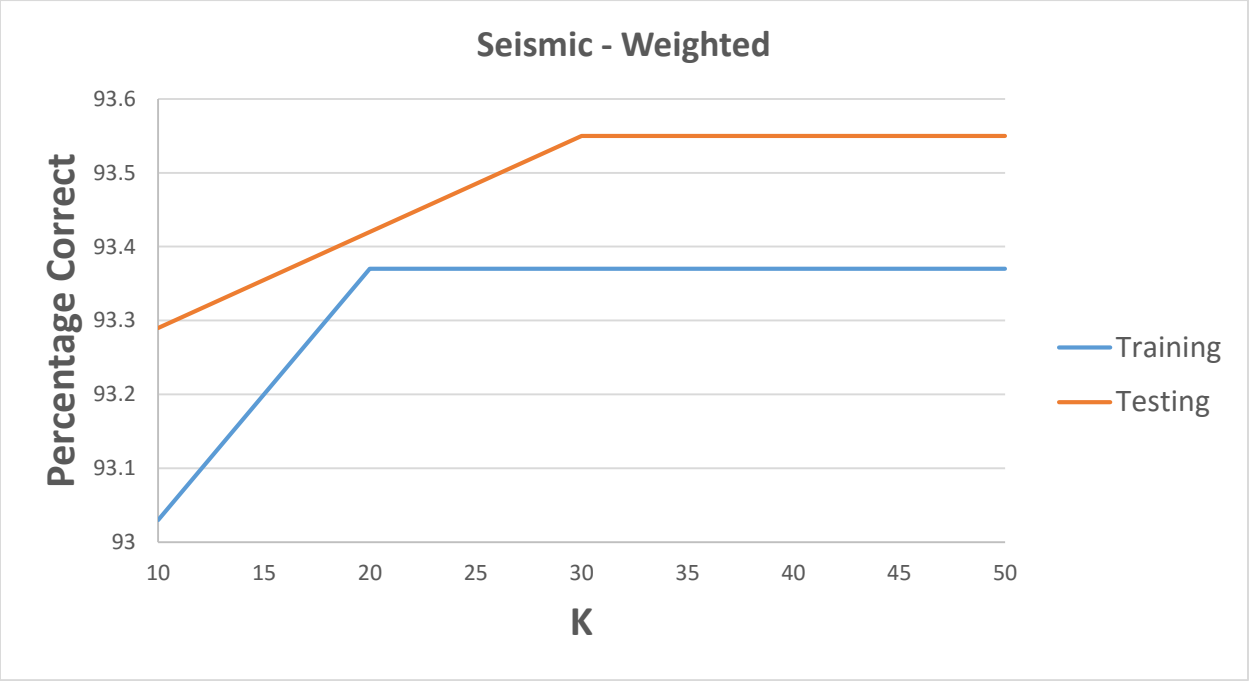


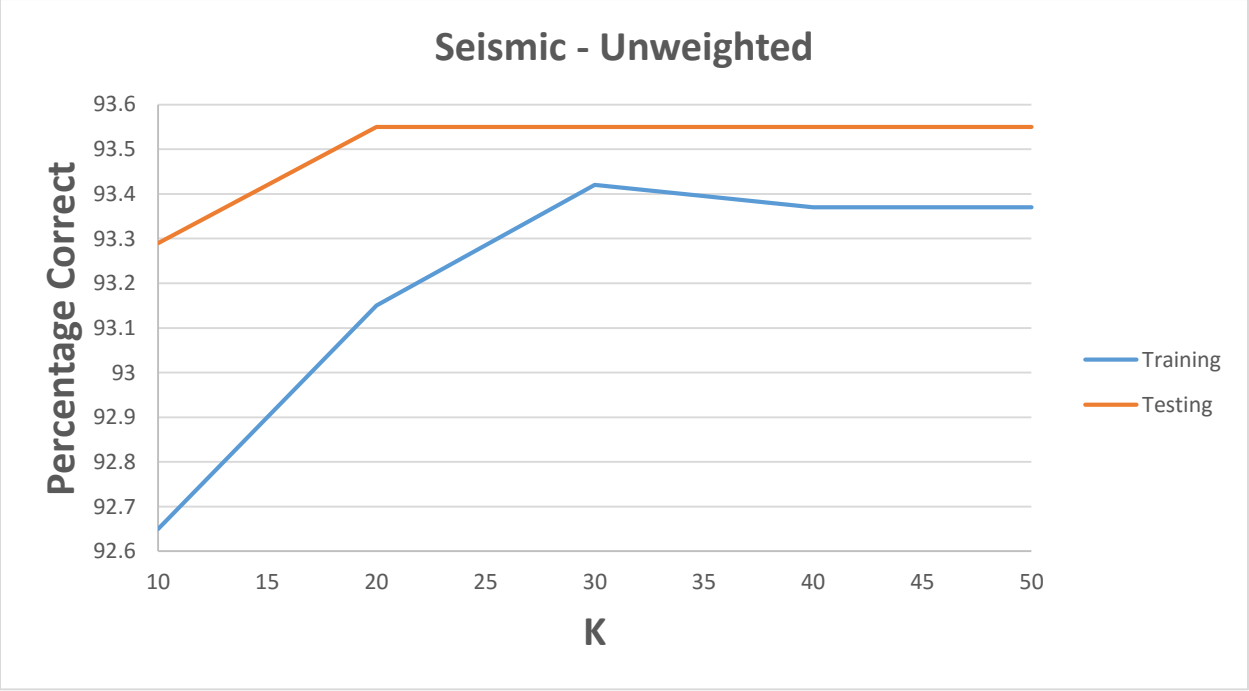
The results for the neural networks for the seismic dataset were interesting and dynamic. Training times were different depending on the dataset and variations in testing were based on learning rate and momentum values. The shortest time recorded among all the datasets to build and complete was roughly 3 minutes. Independent of completing, the shortest build time was 2.34 seconds. Based on the path of the lines, it's easy to notice that both the training and testing lines are decreasing in correct percentage. This is probably due to the fact that it is already good enough at the beginning and is learning more than necessary. Another thing to point out is the red line which represents the testing line at a learning rate of 0.4 and a momentum of 0.1. This line is interesting because of its sudden drop from y = 500 to y = 1500. Here the network tries to extrapolate the function by continuity, without having extracted the influence of the other parameters. Overfitting also starts occurring between testing and testing lines when the learning rates are at 0.4 and 0.3 and the momentums are at 0.2 and 0.1. Also note that it occurs at the second iteration so almost immediately. It occurs slightly as the number of iterations increases as the lines for the training and testing times begin to diverge. On the other extreme, it's interesting to point out that when the learning rate is at 0.2 and momentum is at 0.3, at 4500 epochs the training line and testing line cross paths. For the other times with the exception of the beginning, they do not cross paths. These two lines converged slower than the remaining variations. This set of values would be the results for this network despite correct percentage.



It is important to note that the red and blue lines are being overlapped by the grey line and the yellow and dark blue lines are being overlapped by the green line. For both testing and training the results were the same amongst themselves regardless of variation in learning rate and momentum. The results remained stringently constant especially for the testing data portion. The high level of consistency can be described by the nature of the attributes. The major factors to consider are the time attributes with the restriction of if a solar flare occurred within 24 hours of the previous solar flare. There are also 10 attributes to consider which are all numeric or binary values. Because of these strict conditions, while counter-intuitive, it is easy with lots of room for error added to determine an occurrence of a solar flare. The attributes also are discrete values as opposed to the seismic neural network which are continuous making it easier to learn. Lastly, the size of the solar flares dataset is significantly smaller terms of instances compared to the seismic dataset. There were not enough data points for which the neural network could learn on during the later iterations of the test.

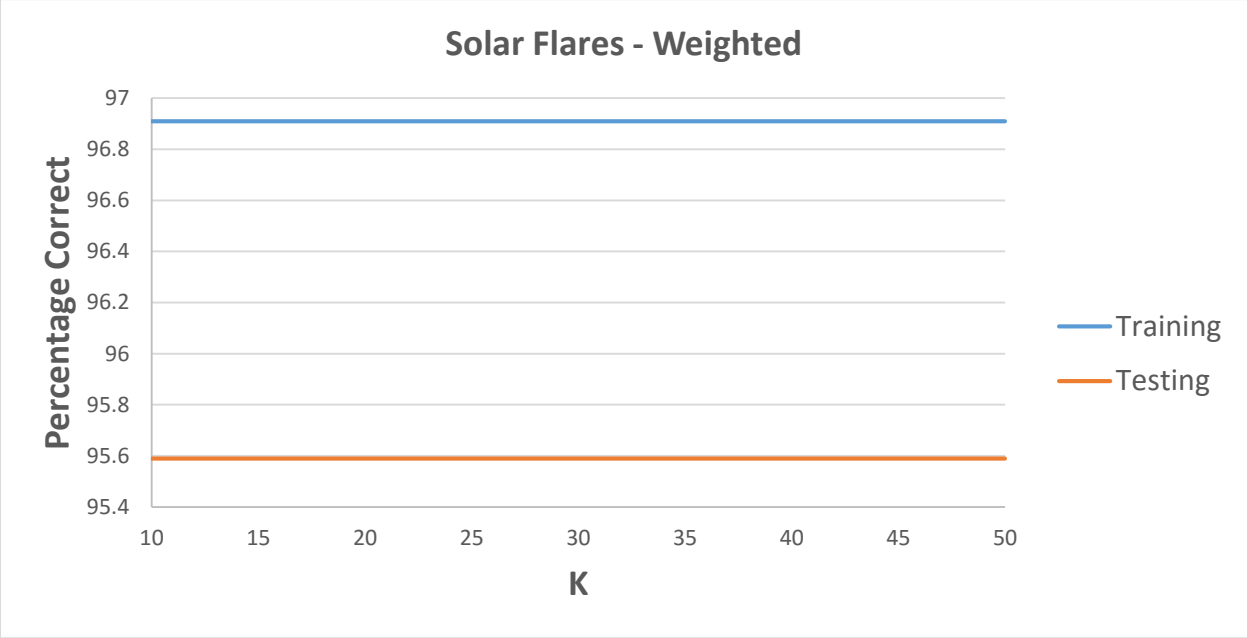
KNN (Ibk):

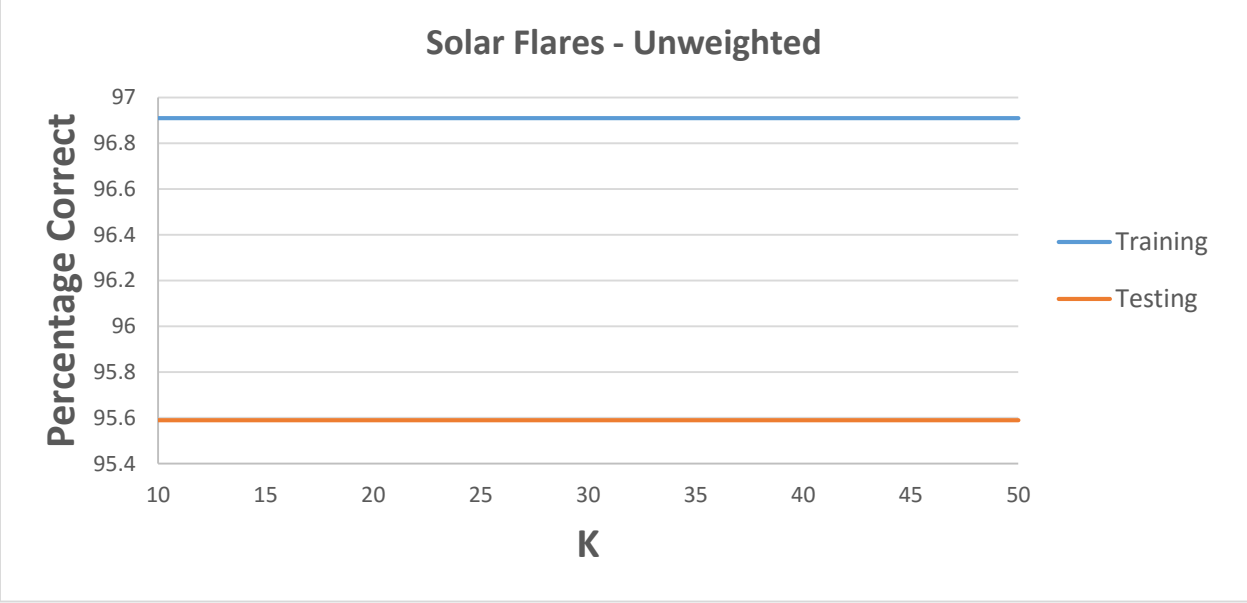




Running the kNN algorithm on the seismic dataset produced interesting results for both the unweighted and weighted runs on the testing data portion. The results were good because there is no indication of overfitting. As the value of K increased, on both the unweighted and weighted graphs, the training and testing data improved until it hit a constant. The one exception is the training data for the unweighted graph which was the only line that showed signs of decreasing in correct percentage. Therefore by the time it reached the highest K value it produces

a slightly worst accurate result than it potentially could have. This may be due to the algorithm clustering multiple classes together as K increased but eventually converged to a constant x value. Among the two, it seems that the weighted kNN algorithm yielded better results albeit being very similar to the unweighted results. From an external viewpoint, it makes sense that the nearest neighbors are more likely to contribute to the end result than an arbitrary data point. Especially in this case when both the training and testing data stay consistent at an early stage of the K value.





Running the solar flares dataset through the kNN algorithm produces identical results for both unweighted and weighted cases. There is absolutely no error in because it is consisted across the board at the same percentage correct locations. This is interesting to me because I expected some fluctuations. The nature of this line can be primarily related to the small number of instances this dataset has. There are not enough influential factors to achieve a non-linear line resulting in constant lines. The only notable difference is the percentage correct for training and testing and in this situation, training beats testing in both cases as opposed to the seismic dataset results.

SVM (Libsvm):

Seismic Activity:

Polykernel	Degree	Training	Testing
	1	67.4406%	48.3871%
	2	66.8325%	50.5416%
	3	58.8170%	48.7742%
	4	29.5191%	15.0968%
	5	23.9912%	6.4516%

RBF	Gamma	Training	Testing
	0.0.1	93.3665%	93.5484%
	0.25	93.3665%	93.5484%
	0.5	93.3665%	93.5484%
	0.75	93.3665%	93.5484%
	1.0	93.3665%	93.5484%

This algorithm produced the most interesting results in reference to the polykernel portion of the seismic activity. First, I would like to note the run time for the polykernel portion was significantly longer the lower the degree value. It was comparable to the time it took to run the neural networks and in some cases took even longer to complete. There are 4 outliers in terms of results which are the values for testing and training at degree 4 and 5. These two tests took significantly less time to compute compared to the first three variations which can be reflected by the percentage classified correct. There is a huge gap between the third variation and the fourth. The results were very inconsistent for the polykernel portion, but was perfectly consistent for the RBF portion where the variations were based on the gamma values. Running the algorithm through the RBF function also took significantly less time compared to polykernel. It was almost instant whereas the polykernel instances took about 20 to 30 minutes for the first two degrees. The decrease in accuracy could be due to failing to split the attributes into more clearly separated hyper-planes. The algorithm did not maximize the gap between different classifications which decreased the value of the results. For polykernel, since the boundary is of some defined but arbitrary order, the seismic dataset was not a good fit. There are many attributes which cause fluctuation and polykernel or more or less supposed to result in a linear model with some slight curve. RBF uses normal curves around the data points and sums these so

that the decision boundary can be defined by a topology condition such as curves where the sum in above a certain value. This complements the seismic dataset because it is the opposite of what is the result of polykernel.

Solar Flares:

Polykernel	Degree	Training	Testing
	1	95.5882%	96.9072%
	2	95.5882%	96.9072%
	3	95.5882%	96.9072%
	4	95.5882%	96.9072%
	5	95.5882%	96.9072%

RBF	Gamma	Training	Testing
	0.0.1	96.9072%	95.5882%
	0.25	96.9072%	95.5882%
	0.5	96.9072%	95.5882%
	0.75	96.9072%	95.5882%
	1.0	96.9072%	95.5882%

Similar to running the kNN algorithm on the solar flares dataset, the results were consistent across the board for both the polykernel and RBF case. This is simply due to the lack of instances and the relationships among the attributes. This was to be expected because the because regardless of the size of the gamma value or degree which both influence the radius of the support vectors, the model is too constrained and cannot capture the complexity or “shape” of the data.