Final Report for EE236A
Name: Yifan Shu
Uid: **********

Introduction to the Project:

In this project we have implemented a prototype selection for nearest neighbor classification. The code we implemented consists of four Python files: classifier.py, test_iris_plants.py, test_breast_cancer.py, and test_digits.py. The first Python file contains the classifier class using nearest neighbor with prototype selection. The other three Python files runs the test using different datasets to validate the performance of the classifier.

<u>classifier.py:</u>

This class defines the classifier using nearest neighbor with prototype selection, it contains several methods (further details can be shown on the code).

<u>test_iris_plants.py, test_breast_cancer.py, and test_digits.py:</u>

Using the corresponding dataset to validate the model using the method specified in the project guideline.

Result:

<u>Iris Plants Dataset:</u>

For this dataset, we use a 4-folder validation cross validation method to validate the model. For this dataset, we choose the $\epsilon = 0.5$ and $\lambda = 1/N$, where N is the number of input training dataset.

After running the test, we get the following result: the average accuracy = 0.939, the average prototypes generated = 48.75, average objective value = 5.68, and average cover error = 0.00.

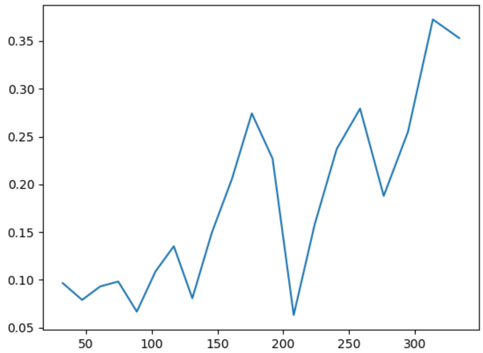
The result shows that the nearest neighbor with prototype selection perform well on Iris data set.

<u>Breast Cancer Dataset:</u>

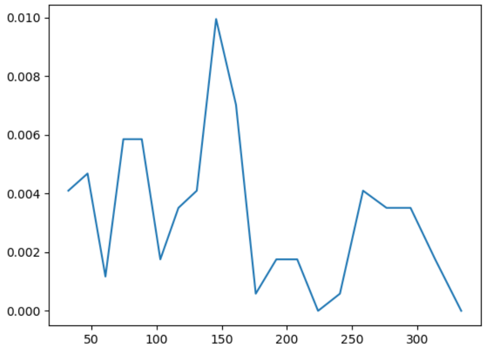
For this dataset, we use a 4-folder validation cross validation method to validate the model. For this dataset, $\lambda = 1/N$, where N is the number of input training dataset. For ϵ, we choose 20 different ϵ ranging between 2 percentiles and 40 percentiles of the inter-point distances.

After running the script, the result can be shown in following graphs:

- Test Error, Cover Error vs. ϵ

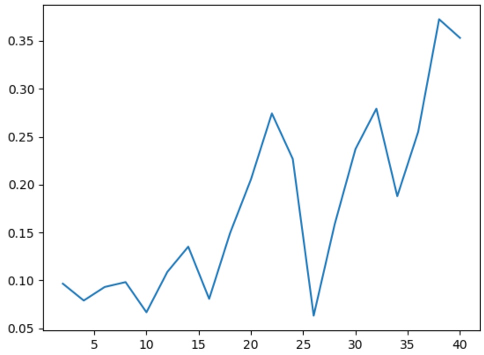


Test Error vs. ϵ

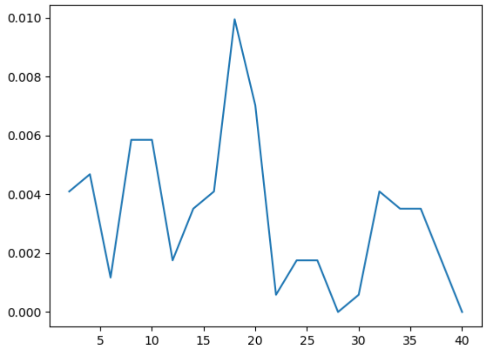


Cover vs. ϵ

- Test Error, Cover Error vs. Percentiles of the inter-point distances

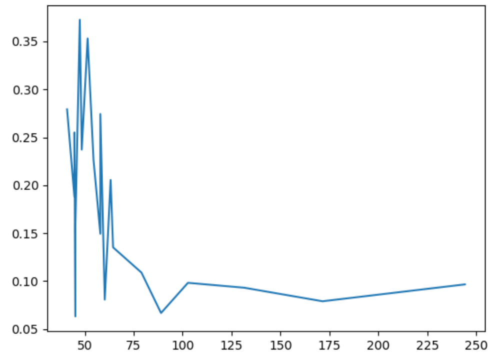


Test Error vs. Percentiles

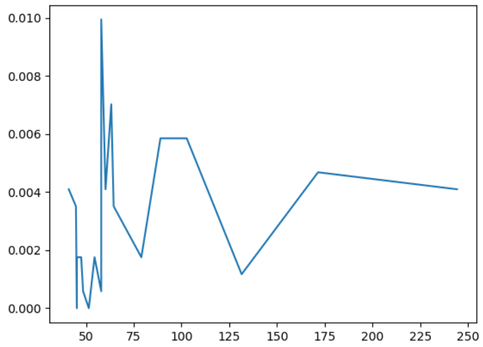


Cover Error vs. Percentiles

- Test Error, Cover Error vs. Average Number of Prototypes



Test Error vs. Average Number of Prototypes

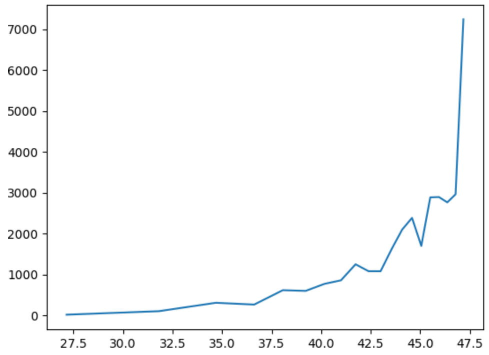


Cover vs. Average Number of Prototypes

Digits Dataset:

For this dataset, we use a 4-folder validation cross validation method to validate the model. For this dataset, $\lambda = 1/N$, where N is the number of input training dataset. For ϵ, we choose 20 different ϵ ranging between 2 percentiles and 40 percentiles of the inter-point distances.

After running the script, the relationship between objective value and ϵ can be shown in the following graphs:



Objective Value vs. ϵ

Modification:

In the original method, we are using L2 norm, which indicates that the ϵ region is a ball. However, we can change the distance metric to form different shape. For example, we can use L1 or L∞ norm to indicate ϵ region is a cube. And we can set the n in Ln norm between 1 and 2 or between 2 and ∞ to make the cube smoother. We will focus on L1 and L∞ norm and will repeat the validation using Iris Plants dataset and Breast Cancer dataset using the same method as before.

Iris Plants Dataset:

	L2 Norm	L1 Norm	L∞ Norm
Accuracy	0.939	0.912	0.886
Cover Error	0	0	0
Number of Prototypes	40	79	36

Breast Cancer Dataset:

	L2 Norm	L1 Norm	L∞ Norm
Best Accuracy	0.070	0.058	0.065
Corresponding Cover Error	0.004	0.004	0.004
Corresponding ϵ	60.87	105.05	63.10
Corresponding Number of Prototypes	139	138	111

Observation:

- The larger the ϵ is, the larger test error tends to be and the smaller cover error tends to be.
- The more the number of prototypes generated, the smaller test error tends to be. But it does not affect cover error too much.
- The larger the ϵ is, the larger objective value is.
- The ϵ shape (by changing using different norm) can have little influence on the performance of the final result