Chunyuan Shen

ID:801322013

Homework 4

<https://github.com/cryyin/ECGR5105/tree/main/homework4>

**Problem 1 (50pts):**

Use the cancer dataset to build an SVM classifier to classify the type of cancer (Malignant vs. benign). Use the PCA feature extraction for your training. Perform N number of independent training (N=1, …, K).

1. Identify the optimum number of K, principal components that achieve the highest classification accuracy.
2. Plot your classification accuracy, precision, and recall over a different number of Ks.
3. Explore different kernel tricks to capture non-linearities within your data. Plot the results and compare the accuracies for different kernels.
4. Compare your results against the logistic regression that you have done in homework 3.

Make sure to explain and elaborate your results.

Chart

Description automatically generated

As we can see, when K=2 the model reach the best result, maybe because it only affected by a few main factors, others are interference

Graphical user interface, text, application, email

Description automatically generated

Linear has the best accuracy, poly has the fastest train speed.

 In last homework when k=9 it has the best performance:

Accuracy: 0.9736842105263158

Precision: 0.9558823529411765

Recall: 1.0

It is batter then this times.

**Problem 2 (50pts):**

Develop a SVR regression model that predicts housing price based on the following input variables:

Area, bedrooms, bathrooms, stories, mainroad, guestroom, basement, hotwaterheating, airconditioning, parking, prefarea

1. Plot your regression model for SVR similar to the sample code provided on Canvas.
2. Compare your results against linear regression with regularization loss that you already did in homework1.
3. Use the PCA feature extraction for your training. Perform N number of independent training (N=1, …, K). Identify the optimum number of K, principal components that achieve the highest regression accuracy.
4. Explore different kernel tricks to capture non-linearities within your data. Plot the results and compare the accuracies for different kernels.

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

The lose in homework 1 is 0.04780662856311236, poly and rbf are smaller than it.

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

This is the plot for rbf kernel, we can see when k=9 it got the best result.

Text

Description automatically generated

For different kernels, rbf is the best one.