Homework 2 Solutions

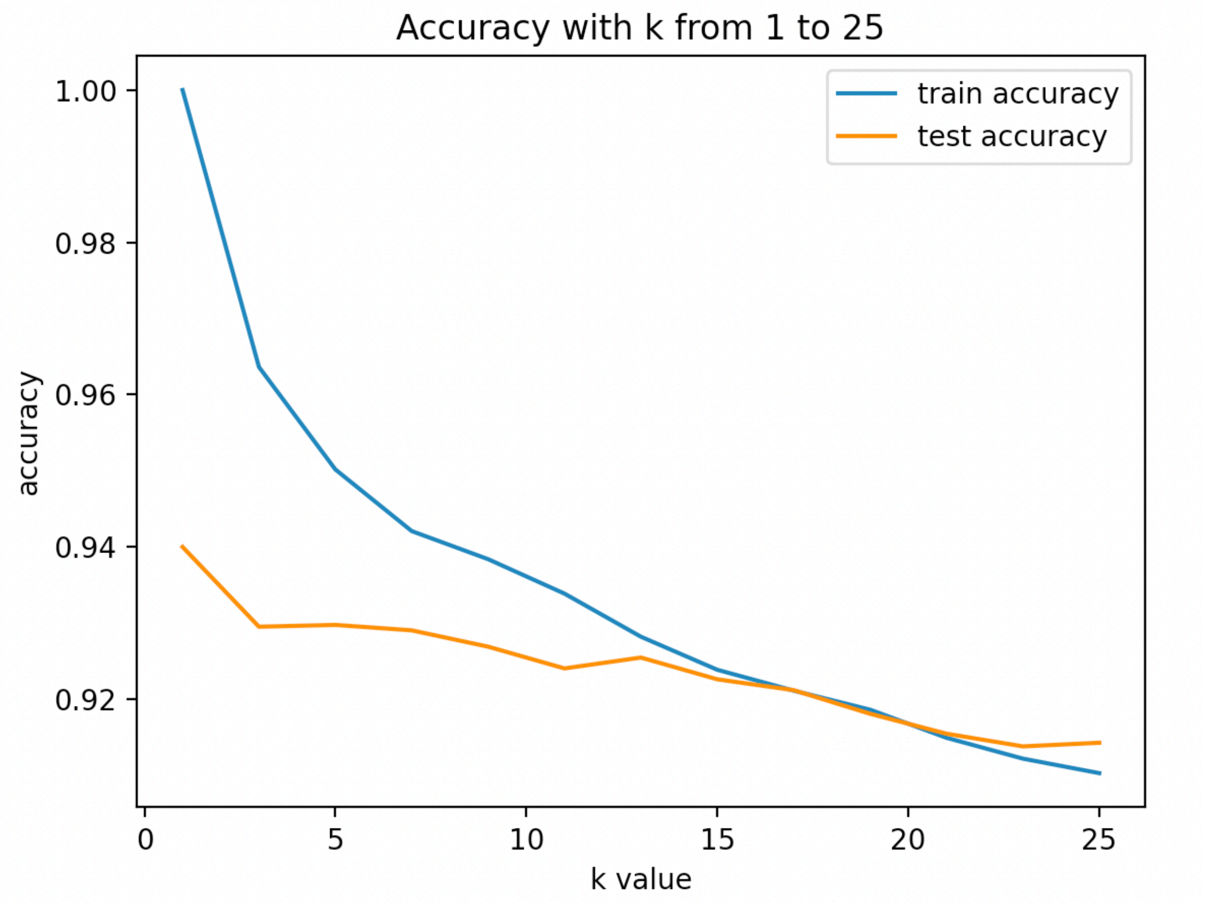
Problem 1:

b) **Logistic Regression Accuracy:**

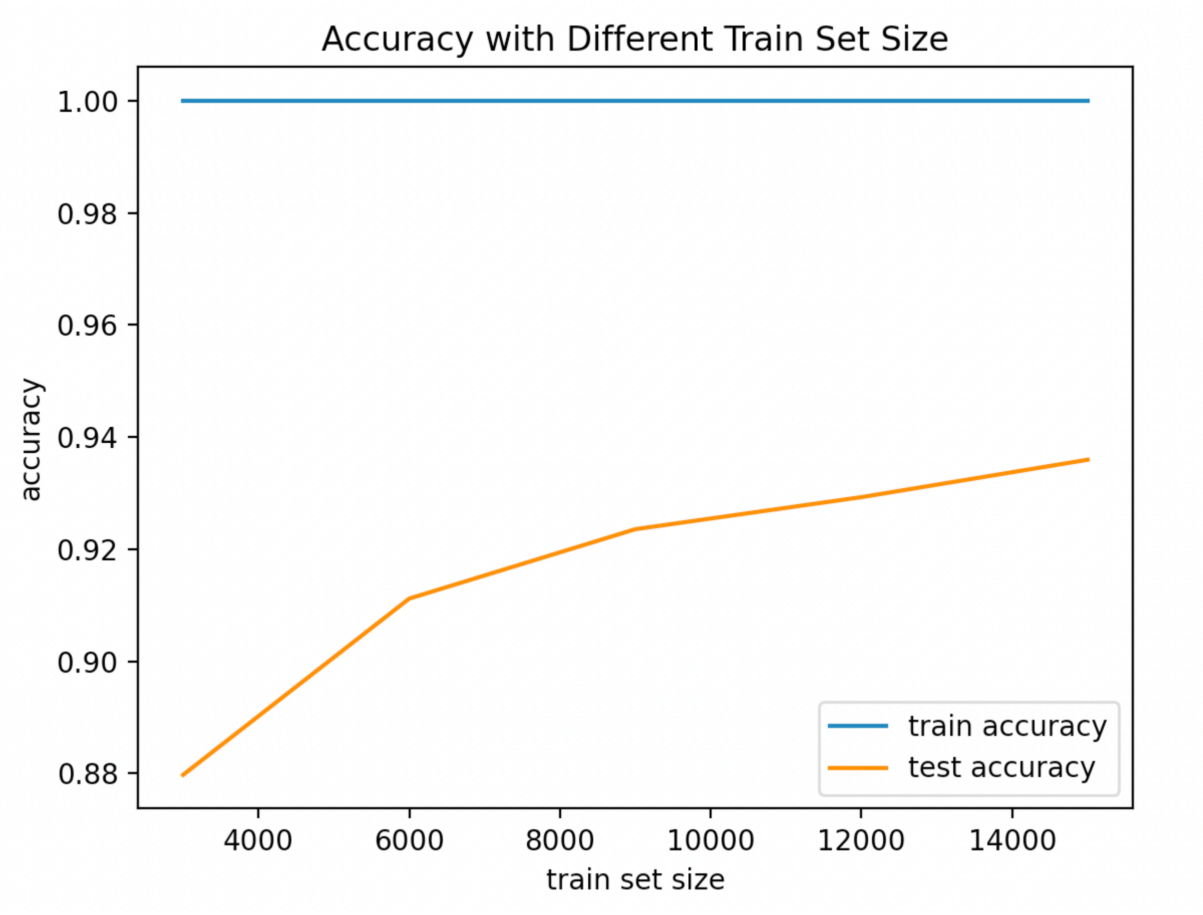
Train Accuracy: 0.99

Test Accuracy: 0.91

c) **kNN train and test accuracy using k from 1 to 25 with a step size of 2:**



e) **kNN train and test accuracy using training set size from 3000 to the full dataset with a step size of 3000**



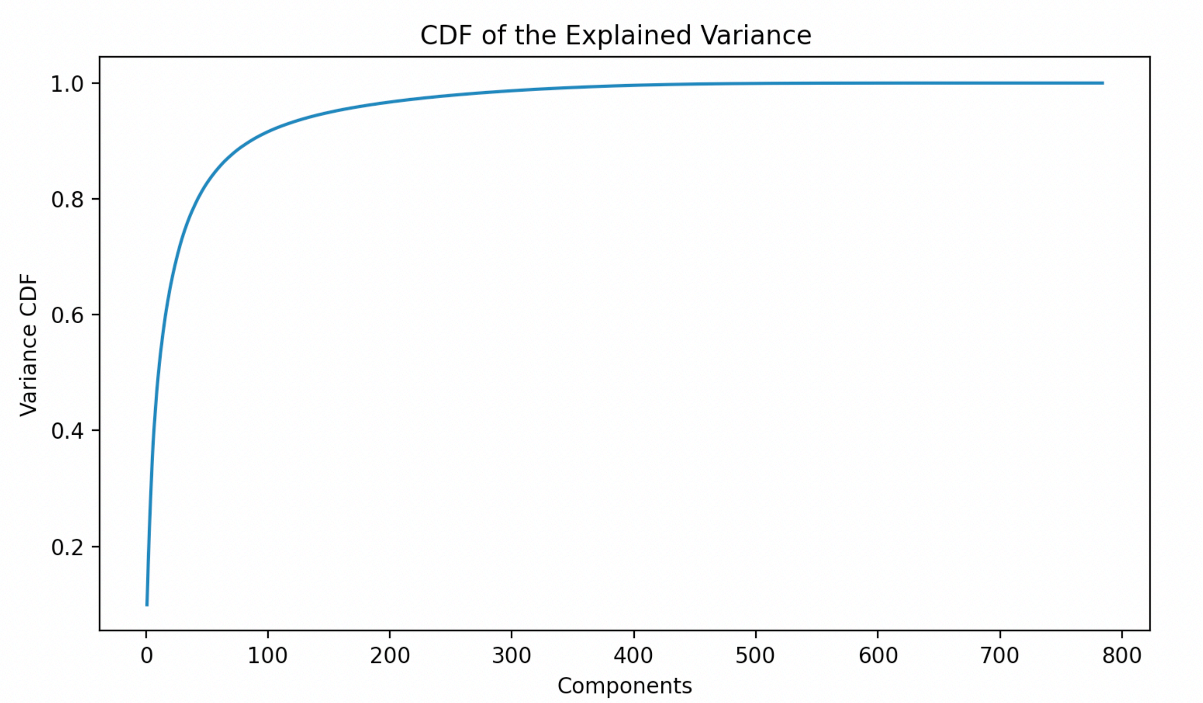
f) **Pros and Cons & Comparison of Logistic Regression and kNN**

- Logistic Regression:
  - Pros:
    - Easy, fast and simple classification method.
    - θ parameters explains the direction and intensity of significance of independent variables over the dependent variable.
    - Can also be used for multiclass classifications.
    - Loss function is always convex.
  - Cons:
    - Cannot be applied on non-linear classification problems.
    - Proper selection of features is required.
    - Good signal to noise ratio is expected.
    - Colinearity and outliers tampers the accuracy.
- kNN:
  - Pros:
    - Easy and simple machine learning model.
    - Few hyperparameters to tune.
  - Cons:
    - k should be wisely selected.
    - Large computation cost during runtime if sample size is large.
    - Proper scaling should be provided for fair treatment among features.
- Logistic Regression (LR) vs kNN :
    - kNN is a non-parametric model, where LR is a parametric model.

  - kNN is comparatively slower than Logistic Regression.
  - KNN supports non-linear solutions where LR supports only linear solutions.
  - LR can derive confidence level (about its prediction), whereas KNN can only output the labels.
- When and why to use Logistic Regression over kNN:
  When we are solving a linear classification problem, especially with a very large dataset, it's better to use logistic regression over kNN to save the training time.

Problem 2:

b) **CDF of the Explained Variance:**



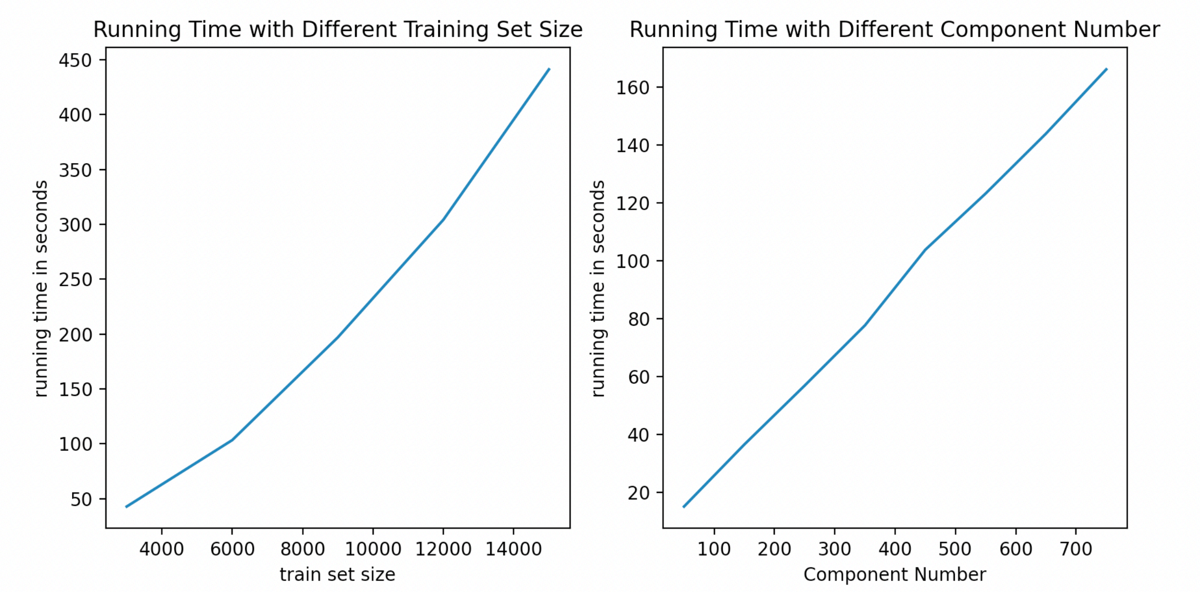
C) **kNN Accuracy with PCA:**

Use the CDF and GridSearchCV to decide on the number of components to be5 0 to get the best performance when k equals to 1.

Train Accuracy: 1.00

Test Accuracy: 0.96

d) **kNN Running Time Comparison with Different Training Size and Components**



e) **Choose k to be 1, component number to be 50, training set size to be 9000:**

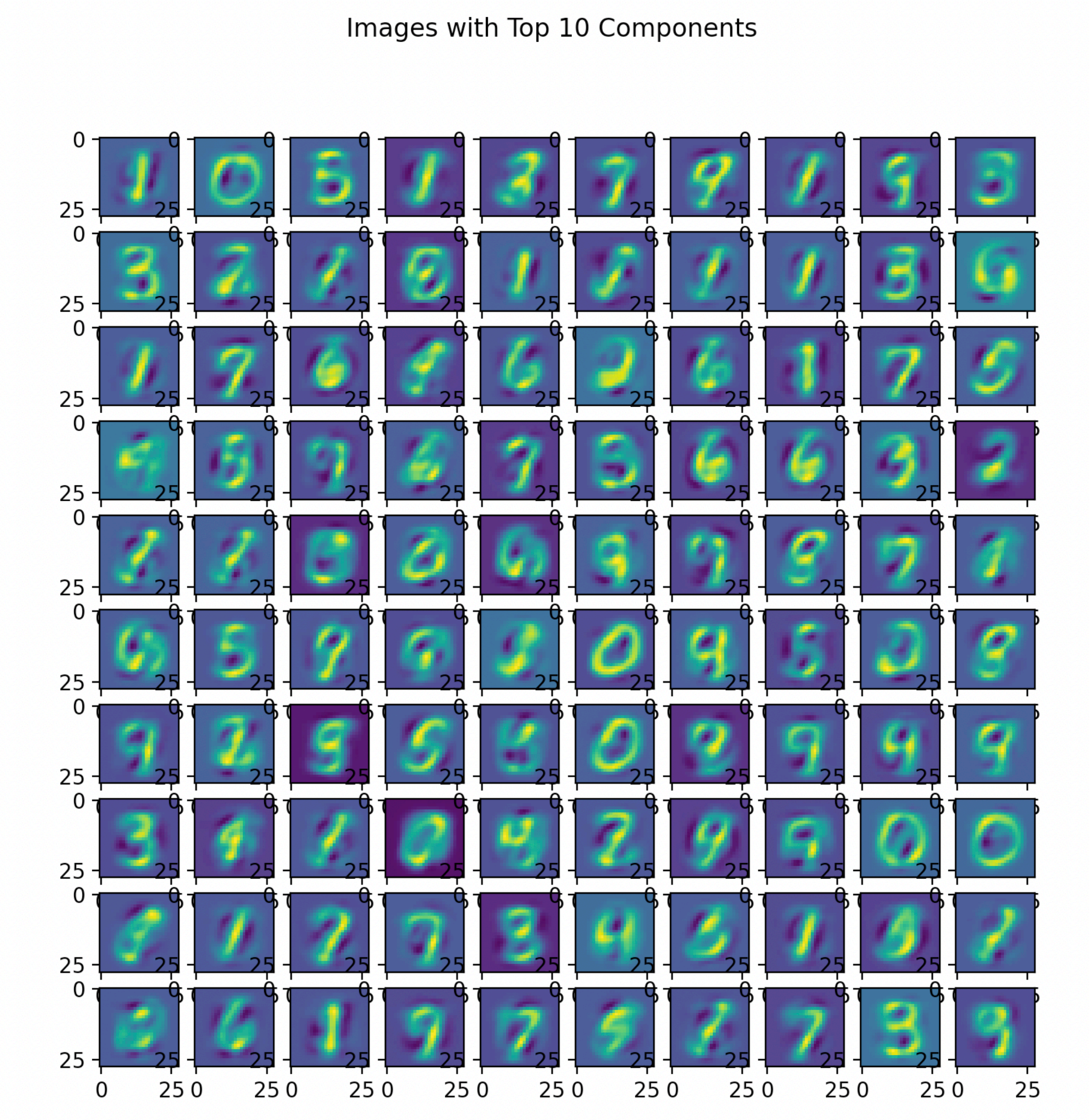
Train Accuracy: 1.0

Test Accuracy: 0.94

Running Time: 7.3

With proper choices of the values of k, component number, the accuracy rate on test data is almost the same as the best performance when using kNN training on the whole data set. However, the running time has dropped dramatically from 450+ seconds to 7.3 seconds.

f) **100 Sample Images of the First 10 Components**



Images with Top 10 Components