Yichen Du 915170850

PA2 Report

Part1:

1. The MNIST dataset
2. http://yann.lecun.com/exdb/mnist/
3. the features of this dataset are all in the 28x28 pixel images,
   and the target which should be predicted is the number represented by the 28x28 image which's range is 0-9.
4. 60000 examples for training and 10000 for testing
5. all the features are represented by the 28x28 pixel of an image which is one example.
6. (1) all images are unique, and the range of the number represented by the images is 0-9
   (2)70000 examples
   (3)28x28 = 784 features

Part2:

1. for titanic_train.csv
   for feature First_class , min error rate is 0.3249 when value > 0.0
   for feature Sex , min error rate is 0.2199 when value > 0.0
   for feature Age , min error rate is 0.4062 when value > 1.0
   for feature SibSp , min error rate is 0.4062 when value > 1.0
   for feature ParCh , min error rate is 0.3852 when value <= 0.0
   for feature Embarked , min error rate is 0.3838 when value <= 0.0

   for breast_cancer.csv
   for feature age , min error rate is 0.2972 when value <= 5.0
   for feature menopause , min error rate is 0.2972 when value <= 2.0
   for feature tumor-size , min error rate is 0.2972 when value <= 10.0
   for feature inv-nodes , min error rate is 0.2797 when value <= 0.0
   for feature node-caps , min error rate is 0.2832 when value <= 0.0
   for feature deg-malig , min error rate is 0.2797 when value > 0.0
   for feature breast , min error rate is 0.2972 when value <= 1.0
   for feature breast-quad , min error rate is 0.2937 when value <= 4.0
   for feature irradiat , min error rate is 0.2972 when value <= 1.0

2. for titanic_train.csv
   for full decision tree, train error rate is 0.2031

   for breast_cancer.csv
   for full decision tree, train error rate is 0.0210

3. (1) Yes, we can directly use the kNN or Perceptron model to train a classifier on these two datasets.

(2) All the features can be numerical, and each example can be vectorized into a N-dim vector, and the vector can be inputted to the model.

(3) each no-numerical feature can be converted into a one-hot vector, like 5 will convert to [0, 0, 0, 0, 1], and the vectors can compute the inner-products and compute the distances.