# Part1

1.

when k = 1, classification accuracy = 0.9066666666666666. predicted class labels of each instance in the test shows below.

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-setosa

Iris-versicolor

Iris-versicolor

Iris-virginica

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-virginica

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-versicolor

Iris-virginica

Iris-virginica

Iris-versicolor

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-versicolor

Iris-versicolor

Iris-virginica

Iris-virginica

Iris-virginica

Iris-versicolor

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-virginica

Iris-versicolor

2.

When k = 1

classification accuracy = 0.9066666666666666

time = 41300(microseconds)

when k = 3

classification accuracy = 0.96

time = 41987(microseconds)

when the k=3, the classification accuracy increasing approximately 6% the performances are approaching for k=1 (41300 microseconds) and k=3(41987 microseconds). When k = 1, it will be Overfitting. Which means that it may easy to add noisy to model.

3.

Advantage:

1. effective if the training data is large, easy to use No training required.
2. Insensitive to outliers
3. Suitable for classifying rare events
4. Suitable for multi-model

Disadvantage:

1. Computation cost and memory cost is quite large. Because we need find distance to all known samples for each text
2. Poor interpretability, cannot tell which variable is more important, and cannot give rules like decision trees
3. Negative learning method, lazy algorithm
4. Need to determine value of parameter k. When the sample is unbalanced, such as the sample size of one class is large, and the sample size of other classes is very small, it may cause that when a new sample is input, the sample of the large-capacity class among the K neighbors of the sample accounts for most

4.

K-fold Cross Validation is used to splitting dataset into training data and testing data

Steps:

1. chop the data into 5 equal subsets

For each subset:

– Treat it as the test set

– Treat the rest 4 subsets as the training set

– Train classifier using the training set, apply it to the test set

(2) The training/test process is repeated 4 times (the folds), with each of the 4 subsets used exactly once as the test set

(3) The 4 results from the folds can be then averaged (or

otherwise combined) to produce a single estimation

5.

If class labels are not available, use the K Means Clustering method to group the examples

(1) Set 3 initial “means” randomly from the data set.

(2) Create 3 clusters by associating every instance with the nearest mean based on a distance measure.

(3) Replace the old means with the centroid of each of the 3 clusters (as the new means).

(4) Repeat the above two steps until convergence (no change in each

cluster center).

# Part2

1.

Baseline classifier accuracy: 0.8518518518518519

Accuracy = 0.7777777777777778

2 categories  
16 attributes  
Read 100 instances

FEMALE = false

FATIGUE = false

ASCITES = false

BIGLIVER = false

Class = live, prob = 1.0

BIGLIVER = true

ANTIVIRALS = true

BILIRUBIN = false

Class = die, prob = 1.0

BILIRUBIN = true

AGE = false

HISTOLOGY = false

Class = live, prob = 1.0

HISTOLOGY = true

MALAISE = false

SPLEENPALPABLE = true

SPIDERS = false

SGOT = false

Class = die, prob = 1.0

SGOT = true

ANOREXIA = false

ANOREXIA = true

Class = die, prob = 1.0

SPIDERS = true

Class = die, prob = 1.0

MALAISE = true

Class = die, prob = 1.0

AGE = true

Class = die, prob = 1.0

ASCITES = true

SPIDERS = false

SPLEENPALPABLE = false

ANTIVIRALS = true

BILIRUBIN = true

BIGLIVER = false

Class = die, prob = 1.0

BIGLIVER = true

AGE = false

Class = die, prob = 1.0

AGE = true

SPLEENPALPABLE = true

AGE = false

VARICES = false

Class = live, prob = 1.0

VARICES = true

ANOREXIA = false

Class = live, prob = 1.0

ANOREXIA = true

MALAISE = false

STEROID = false

ANTIVIRALS = false

Class = die, prob = 1.0

ANTIVIRALS = true

MALAISE = true

ANTIVIRALS = false

Class = live, prob = 1.0

ANTIVIRALS = true

AGE = true

Class = die, prob = 1.0

SPIDERS = true

VARICES = false

Class = die, prob = 1.0

VARICES = true

SPLEENPALPABLE = false

Class = live, prob = 1.0

SPLEENPALPABLE = true

BIGLIVER = false

Class = live, prob = 1.0

BIGLIVER = true

ANOREXIA = false

Class = live, prob = 1.0

ANOREXIA = true

SGOT = false

HISTOLOGY = false

AGE = false

MALAISE = false

MALAISE = true

Class = live, prob = 1.0

AGE = true

Class = live, prob = 1.0

HISTOLOGY = true

Class = live, prob = 1.0

SGOT = true

Class = live, prob = 1.0

FATIGUE = true

MALAISE = true

ANOREXIA = true

ASCITES = true

SPLEENPALPABLE = false

Class = live, prob = 1.0

SPLEENPALPABLE = true

BIGLIVER = false

Class = live, prob = 1.0

BIGLIVER = true

ANTIVIRALS = false

Class = live, prob = 1.0

ANTIVIRALS = true

SGOT = false

VARICES = false

Class = die, prob = 1.0

VARICES = true

SPIDERS = false

Class = live, prob = 1.0

SPIDERS = true

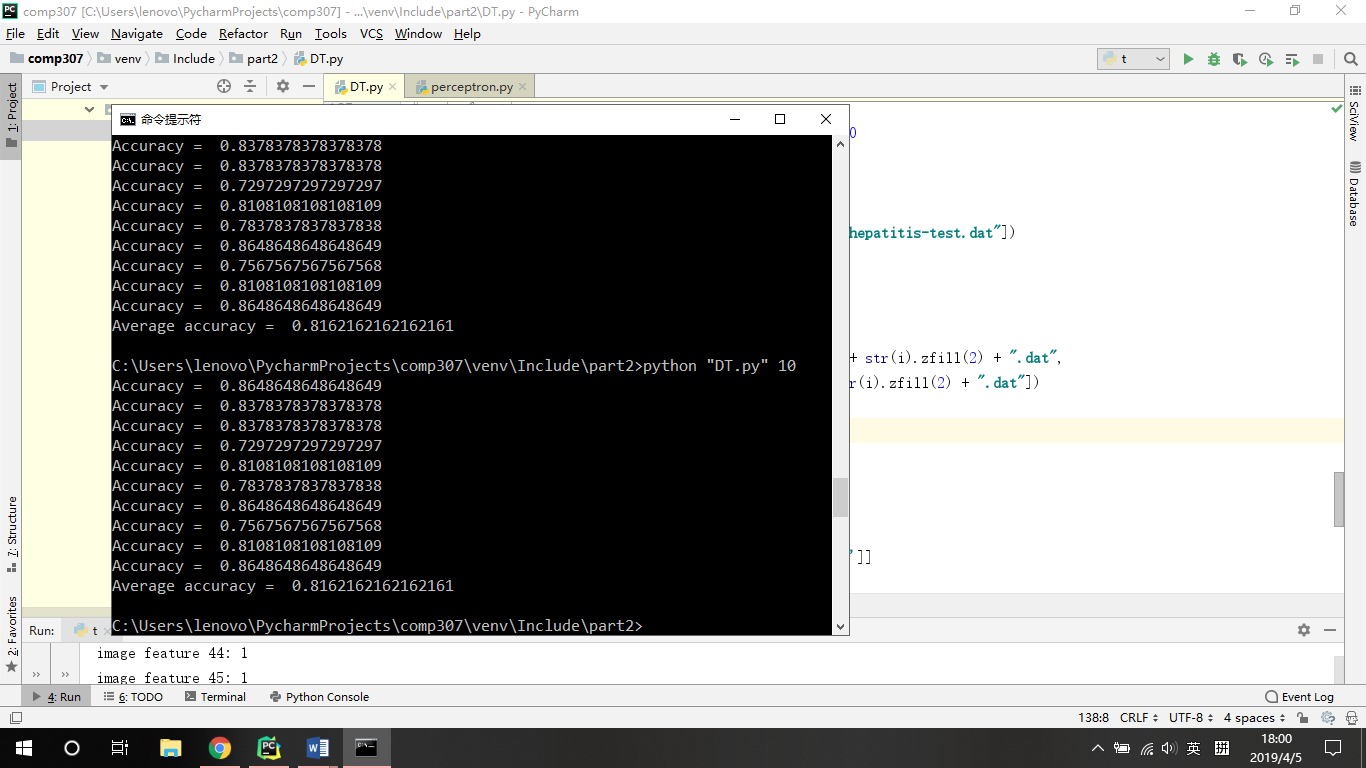
SGOT = true

Class = live, prob = 1.0

FEMALE = true

Class = live, prob = 1.0

2.



3.

(a)

**Reduced Error Pruning**

Step1: Use leaf node to replace subtree,

Step2: Check whether it is beneficial, if the error rate is reduced or unchanged after pruning

Step3: repeat step1 and 2, until the error rate has risen

(b) because we delete some dataset from train dataset when decision tree prune, thus for original train dataset the accuracy of train dataset will decreasing.

(c) Shrink the tree, make it smaller/simpler, to reduce overfitting, thus the accuracy of test dataset may improve.

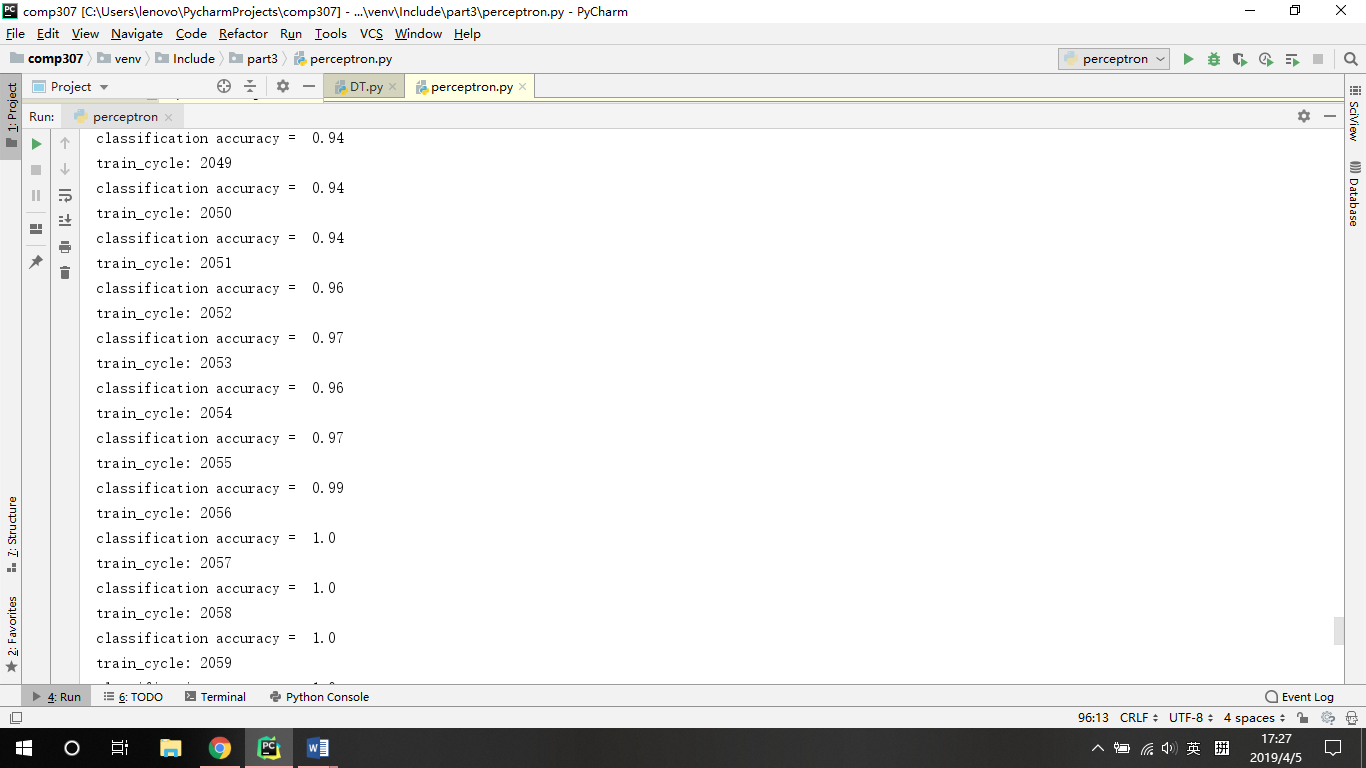
4.

when there are three or more classes, impurity measure may cause false positive if the one of class probability is zero. Then the weighted average impurity will be zero which means it is pure, but the result is incorrect.

# Part3

1.

Accuracy will increase when the train-times big enough. Which means it would find correct set of weights.



2.

the perceptron’s performance on the training data is not a good measure of its effectiveness

because if use train data to train it will be overfitting and match result perfectly.