Homework 5
Introduction to Data Analysis and Mining
Spring 2018
CSCI-B 365

Siyi Xian

March 25, 2018

All the work herein is solely mine.

Directions

Please follow the syllabus guidelines in turning in your homework. I am providing the LaTeX of this document too. This homework is due Sunday, April 8, 2018 10:00p.m. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. All the work should be your own.

K-Nearest Neighbors (KNN) Algorithm in Theory

ALGORITHM `K-nearest neighbors`
INPUT

- training data Δ
- test data Δ'
- distance metric d, i.e., $d : \Delta^2 \to \mathbb{R}_{\geq 0}$
- integer k: nearest neighbors number

OUTPUT

- class label of each $z \in \Delta'$

for $z = (\mathbf{x}', y') \in \Delta'$ **do**
    Compute $d(\mathbf{x}, \mathbf{x}')$, the distance between z and every example $(\mathbf{x}, y) \in \Delta$
    Select $\Delta_z \subseteq \Delta$, the set of closest k training examples to z
    Voting:

- majority voting: $y' = \text{argmax}_v \sum_{(\mathbf{x_i}, y_i) \in \Delta_z} I(v = y_i)$
- distance-weighted voting: $y' = \text{argmax}_v \sum_{(\mathbf{x_i}, y_i) \in \Delta_z} w_i \times I(v = y_i)$ where $w_i = \frac{1}{d(\mathbf{x}', \mathbf{x_i})^2}$

end for

K-Fold Cross Validation for Model Selection

ALGORITHM `k-fold cross validatiaon`
INPUT

- training data $\Delta = (\mathbf{x_1}, y_1), \ldots, (\mathbf{x_m}, y_m)$
- set of parameter values Θ
- learning algorithm $\mathcal{H}$
- integer k

OUTPUT

- $\theta^* = \text{argmin}_\theta[error(\theta)]$
- $h_{\theta^*} = \mathcal{H}(\Delta; \theta^*)$

Randomly partition Δ into $\Delta_1, \ldots, \Delta_k$
*** $\Delta_1 \cup \Delta_2 \ldots \cup \Delta_k = \Delta$ and $\Delta_i \cap \Delta_j = \varnothing$ for $i \neq j \in [1, 2, \ldots, k]$
for $\theta \in \Theta$ **do**
    for $i = 1 \ldots k$ **do**
        *** Train a model for each training set
        $h_{i,\theta} = \mathcal{H}(\Delta \setminus \Delta_i; \theta)$
    end for
    *** Use the trained models over Δ_i (test data sets) to evaluate the models for each parameter
    $\text{error}(\theta) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}_{\Delta_i}(h_{i,\theta})$
end for

In this homework, you are asked to train several classifiers using k-nearest neighbors (KNN) and naive bayes algorithms over car evaluation and credit approval data sets. The links for the data sets are provided below:

- Car Evaluation Data Set
- Credit Approval Data Set

Problem 1: K-Fold Cross Validation [20 points]

Create 5 training and 5 test data sets from each data set using 5-fold cross validation and save these 20 files. You will use these data sets to answer the rest of the questions. You are allowed to use R packages for k-fold cross validation. However, *students who implement it will receive* 15 *extra points from this question.*

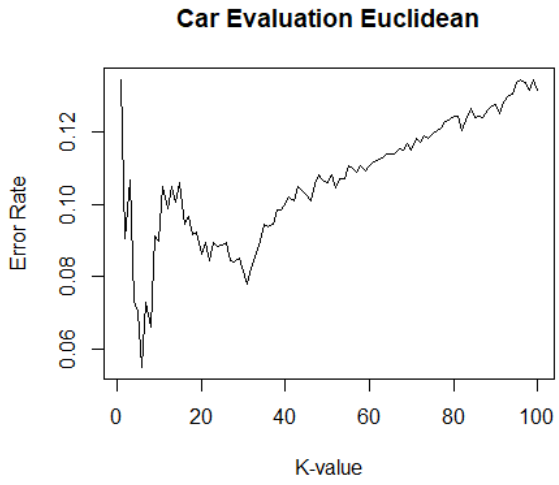
`Please see specific code in crossValidation.R (Starts from line 163).`

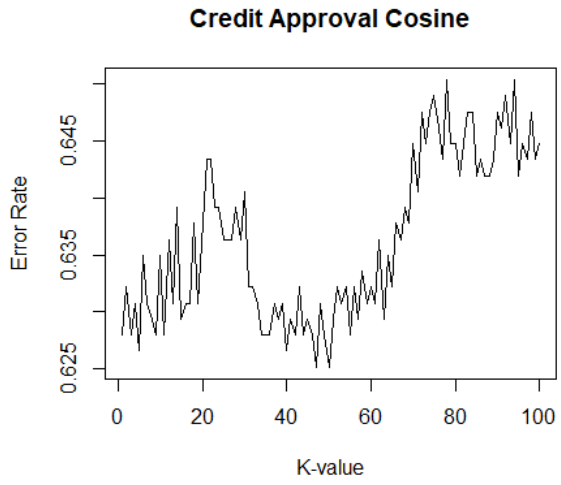
Problem 2: K-Nearest Neighbors (KNN)[40 points]

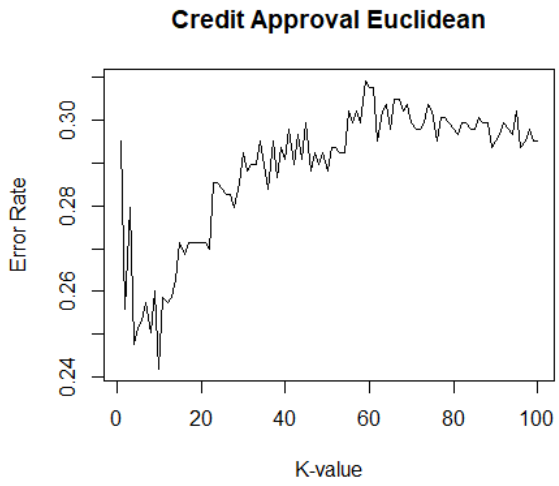
2.1 Implement KNN algorithm with two different distance functions. You can either use existing distance functions, i.e., Euclidean or design your own.

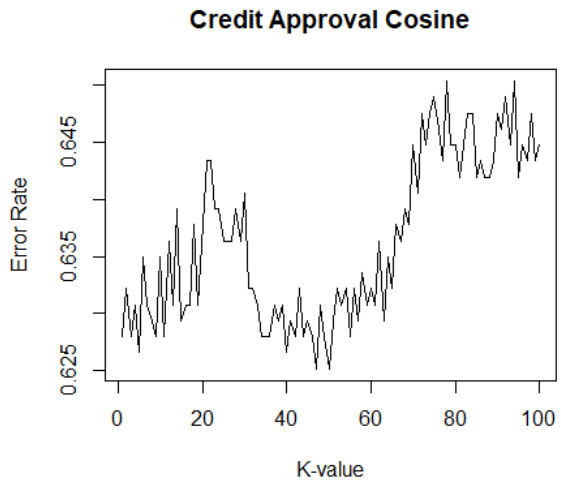
`Please see specific code in knn.R (Starts from line 26).`

2.2 Use the data sets created in problem 1 to determine the optimal k over each data set for KNN algorithm. First, pick 5 different k values and then calculate average error rate of KNN classifiers over test data tests for each k to find the optimal k for each data set and distance function. Report the average error rate for each k, distance function and data set. What are the optimal k and distance function for each data set?





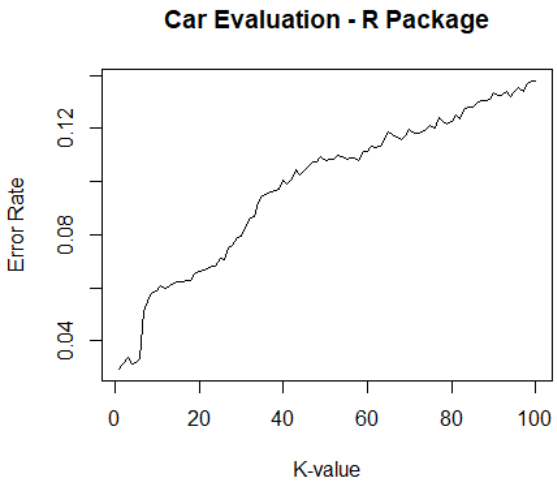


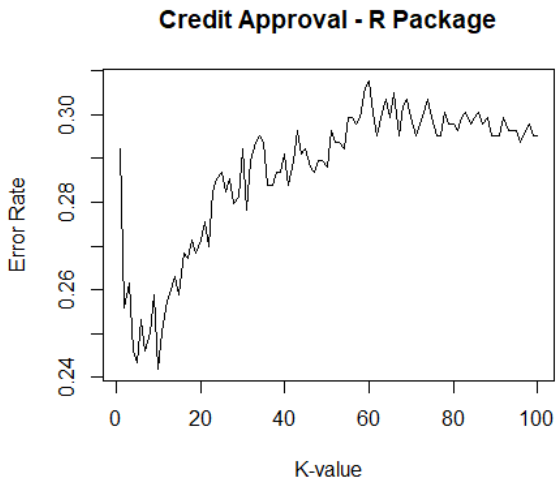



```
Acoording to the graph above, we can decided that:
1. For both data sets, Euclidean will be a better way to calculate distance that Consine.
2. For Car Evaluation Data Set, when K = 7, the error rate will be smallest.
3. For Credit Approval Data Set, when K = 11, the error rate will be the best one to choose.

For specofoc code, please look at the end of crossValidation.R (Starts from line 206).
```

2.3 Use the KNN package in R to validate your results from question 2.2.





When using R pakage, knn will be more efficet due to the run time is much more smaller. However, for the sesualt, it is similar. We will better to choose 2 or 5 for K value in Car Evaluation Data Set. And as for Credit Approval Data Set, we will going to choose K = 11 too. On the other hand, the fluctuation is also similar

Please see specific code in knnValidation.R.

Problem 3: Naive Bayes Classifier vs. K-Nearest Neighbors [20 points]

In this question, you are first asked to train Naive Bayes classifiers to find the optimal Naive Bayes model for car evaluation and credit approval data sets. Second, you will compare your optimal KNN and Naive Bayes models over car evaluation and credit approval data sets. Answer the following questions:

3.1 Train Naive Bayes classifiers over training data sets and test each classifier over corresponding test data. Report the error rates of the classifiers in a figure. Create one figure for car evaluation data set and another one for credit approval data set. You are allowed to use R packages for Naive Bayes algorithm.

```
# Please see detailed code in naiveBayes.R file
> CarEvaluationErrorRates
[1] 0.1987458
> CreditApprovalErrorRates
[1] 0.2106917
```

3.2 Pick the optimal Naive Bayes classifiers (one for car evaluation data set and another one for credit approval data set.) from question 3.1 and compare them with your best KNN models from question 2. Discuss the performances of the optimal classifiers over car evaluation and credit approval data sets, i.e, which one performed better?

For Car Evaluation Data Set, our best error rate is arround 0.04. So for here, KNN can do a better job that Naive Bayes. When we comes to Credit Approval Data Set, the minimum error rate using KNN is 0.24. It is larger than the error rate when doing Naive Bayes. In conclusion, we can find out that the efficence for different algorithm is based on database. Thus, we need to figure out choosing algorithm after studying data set.

Problem 4 [10 points]

7. Consider the data set shown in Table 5.1

(a) Estimate the conditional probabilities for $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$, and $P(C|-)$.

* $P(A=1|+) = \frac{3}{5} = 0.6$
* $P(A=0|+) = \frac{2}{5} = 0.4$
* $P(B=1|+) = \frac{1}{5} = 0.2$
* $P(B=0|+) = \frac{4}{5} = 0.8$
* $P(C=1|+) = \frac{4}{5} = 0.8$
* $P(C=0|+) = \frac{1}{5} = 0.2$
* $P(A=1|-) = \frac{2}{5} = 0.4$
* $P(A=0|-) = \frac{3}{5} = 0.6$
* $P(B=1|-) = \frac{2}{5} = 0.4$
* $P(B=0|-) = \frac{3}{5} = 0.6$
* $P(C=1|-) = \frac{0}{5} = 0$
* $P(C=0|-) = \frac{5}{5} = 1$

(b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample $(A=0, B=1, C=0)$ using the naïve Bayes approach.

Let $P(A=0, B=1, C=0) = K$.

$$
\begin{aligned}
&P(+|A=0, B=1, C=0) \\
&= \frac{P(+|A=0, B=1, C=0) \times P(+)}{P(A=0, B=1, C=0)} \\
&= \frac{P(+|A=0, B=1, C=0) \times P(+)}{K} \\
&= \frac{P(A=0|+)P(B=1|+)P(C=0|+) \times P(+)}{K} \\
&= \frac{0.4 \times 0.2 \times 0.2 \times 0.5}{K} \\
&= \frac{0.008}{K}
\end{aligned}
$$

$$
\begin{aligned}
&P(-|A=0, B=1, C=0) \\
&= \frac{P(-|A=0, B=1, C=0) \times P(-)}{P(A=0, B=1, C=0)} \\
&= \frac{P(-|A=0, B=1, C=0) \times P(-)}{K} \\
&= \frac{P(A=0|-)P(B=1|-)P(C=0|-) \times P(-)}{K} \\
&= \frac{0.6 \times 0.4 \times 0 \times 0.5}{K} \\
&= 0
\end{aligned}
$$

```
We will going to choose +
```

(c). Estimate the conditional probabilities using the m-estimate approach, with $p = 1/2$ and $m = 4$.

* $P(A = 1|+) = \frac{3+2}{5+4} = \frac{5}{9}$
* $P(A = 0|+) = \frac{2+2}{5+4} = \frac{4}{9}$
* $P(B = 1|+) = \frac{1+2}{5+4} = \frac{3}{9}$
* $P(B = 0|+) = \frac{4+2}{5+4} = \frac{6}{9}$
* $P(C = 1|+) = \frac{4+2}{5+4} = \frac{6}{9}$
* $P(C = 0|+) = \frac{1+2}{5+4} = \frac{3}{9}$
* $P(A = 1|-) = \frac{2+2}{5+4} = \frac{4}{9}$
* $P(A = 0|-) = \frac{3+2}{5+4} = \frac{5}{9}$
* $P(B = 1|-) = \frac{2+2}{5+4} = \frac{4}{9}$
* $P(B = 0|-) = \frac{3+2}{5+4} = \frac{5}{9}$
* $P(C = 1|-) = \frac{0+2}{5+4} = \frac{2}{9}$
* $P(C = 0|-) = \frac{5+2}{5+4} = \frac{7}{9}$

(d) Repeat part **(b)** using the conditional probabilities given in part **(c)**. Let $P(A = 0, B = 1, C = 0) = K$.

$$P(+|A = 0, B = 1, C = 0) = \frac{0.0247}{K}$$
$$P(-|A = 0, B = 1, C = 0) = \frac{0.0549}{K}$$

`We will going to choose -`

(e) Compare the two methods for estimating probabilities. Which method is better and why?

`The m-estimate approach is better, becasue it can avoid 0 whihc happend in naive Bayes approach.`

Extra Credit [30 points]

- Problem 1: K-fold cross validation implementation [**15 points**]

  `Please see code in crossValidation.R`

What to Turn-in

Submit a .zip file that includes the files below. Name the .zip file as "usename-section number", i.e., hakurban-B365.

- The *tex and *pdf of the written answers to this document.
- *Rfiles for:
  - Question 1: crossValidation.R, output of cross validation: training and test data sets
  - Question 2.1: knn.R, Question 2.3: knnValidation.R
  - Question 3: naiveBayes.R
- A README file that explains how to run your code and other files in the folder