Subject: ISYE6501

Assignment: HM2

Students (name, GT id):

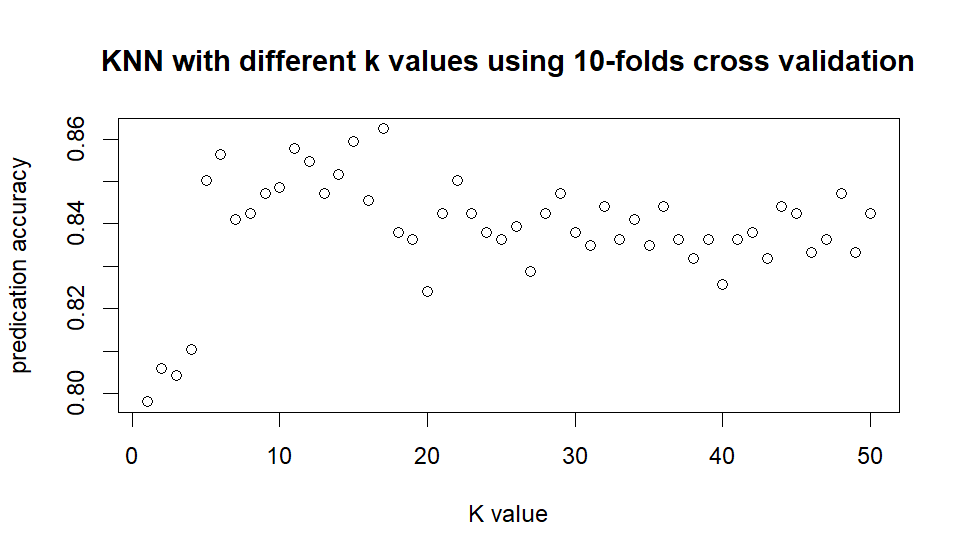
×          Yuanting Fan, 904047984

×          Wenjia Hu, 904057780

×          Sen Yang, 904025995

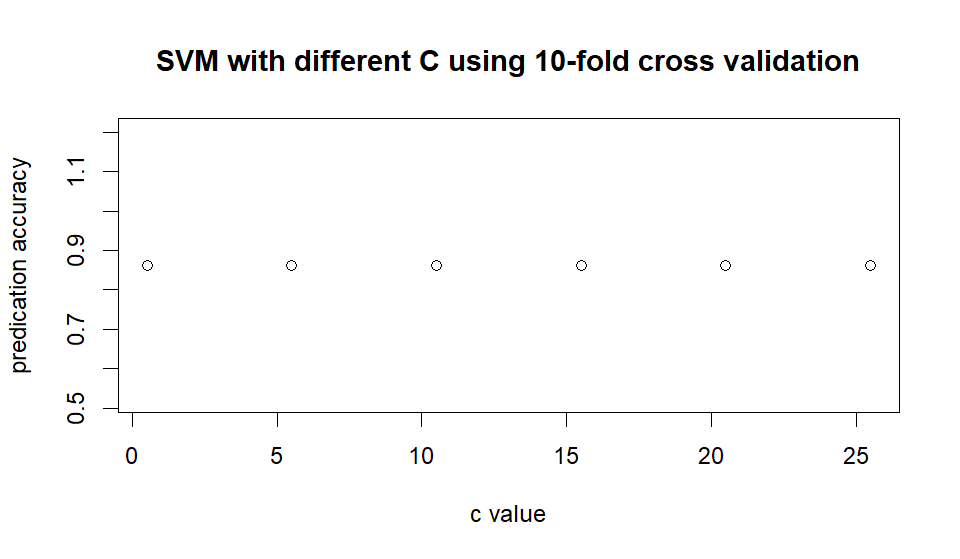
***Question 3.1***

1. Firstly, we use the cv.kknn function to perform 10-fold cross-validation on the entire dataset and determine the best k value based on the average prediction accuracy. We found that when the k value equals 17, the prediction accuracy peaks at 0.8623853. Here’s the relationship between k values and prediction accuracy in cross-validation:

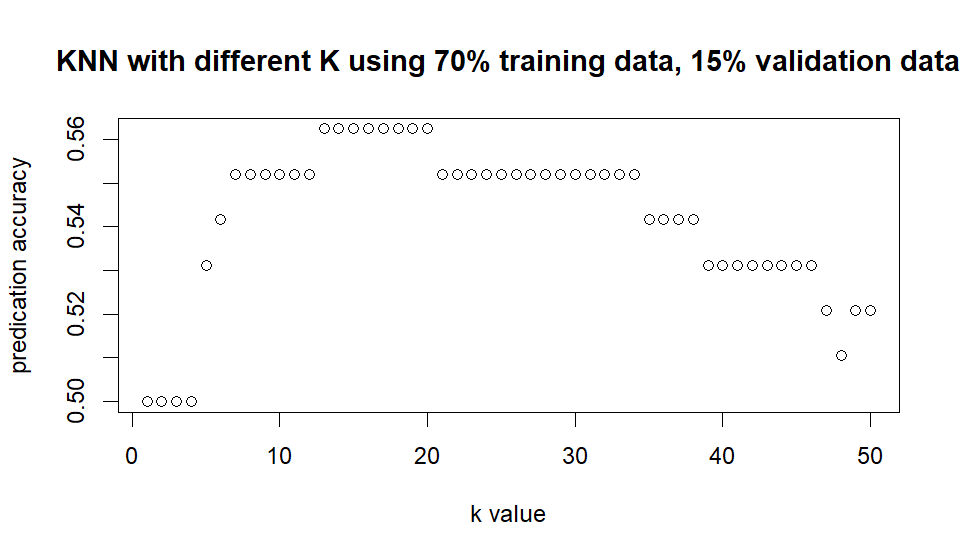


Secondly, we also use the ksvm function (a simple linear SVM model) to perform 10-fold cross-validation on the entire dataset. We found that when C ranges from 0.5 to 30, the prediction accuracy does not change and remains at 0.8629489. The main reason for this is that, as stated in Homework 1, the response is largely determined by one or two predictors (if V5​ equals 1, the response is highly likely to be 1). This means the data is easy to classify correctly, so the SVM does not need to rely on adjusting the margin to improve accuracy.

Again, we obtained another relationship between C values and prediction accuracy in cross-validation (constant):

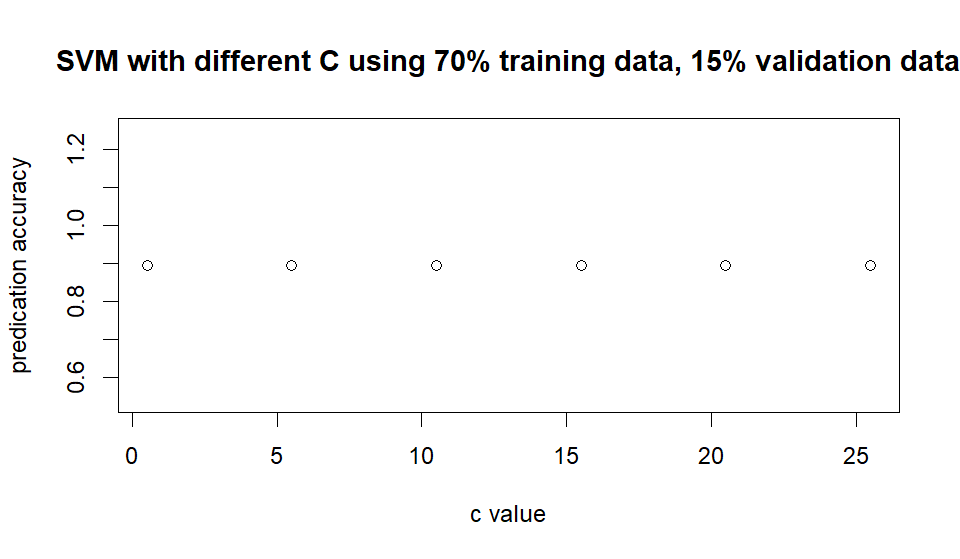


Compared to Homework 1, after cross-validation, the SVM model does not seem to significantly outperform the KNN model for this dataset (but still better). It makes sense because the size of validation data set becomes smaller thus increasing random effects.

1. Firstly, we split the data into training (70%), validation (15%), and test (15%) sets, and used KNN to find a good classifier. We found that when the k value equals 13 (like Homework 1), the validation prediction accuracy peaks at 0.5625. We can see that compared to question (a) the accuracy ratio worsens as the size of the training dataset decreases (from whole dataset to 70%). Here’s the relationship between k values and prediction accuracy:

Select the KNN model with K=13, the test data prediction accuracy reaches 0.4673913, which is lower than the validation procedure.

Secondly, we used the same three datasets to run SVM models. Similarly, we found that when C ranges from 0.5 to 30, the prediction accuracy does not change and remains at 0.8958333, which is higher than the cross-validation accuracy with the whole dataset. Using the SVM model with C=13, the test data

gives us an accuracy ratio of 0.8478261.

Since this dataset is highly linearly separatable, we noticed that KNN model is highly sensitive to the size of training dataset while SVM is more stable.

# Extra:

We also run another two algorithms by splitting data to 80% for cross-validation (K=10) and 20% for test by KNN model (K=5,12,40) and SVM (C=50,250,400). We obtain a table as following:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Linear SVM（C=50）** | **Linear SVM（C=250）** | **Linear SVM（C=400）** | **KNN**  **(K=5)** | **KNN**  **(K=12)** | **KNN**  **(K=40)** |
| **Cross-validation accuracy** | 0.8612404 | 0.8612404 | 0.8612404 | 0.4641746 | 0.4827596 | 0.4675445 |
| **Test data accuracy** | 0.8677686 | 0.8677686 | 0.8677686 | 0.4132231 | 0.4214876 | 0.3884298 |

# It provides similar insights about the features of SVM and KNN model. Noticeably, KNN has become worse.

***Question 4.1***

**Situation**

A clustering model can play an important role in the pre-construction market research of real estate companies. Before formulating a housing development plan, property developers usually need to categorize potential clients into several groups based on the clients’ attributes. By doing so, they can develop more customized strategies in terms of floor plans, pricing, marketing, and financial requirements for clients, to better attract target audience in a competitive market.

Below are five predictors that real estate companies can consider when clustering their clients:

1. **Household family size:** Family size is crucial for determining the home size, the number of rooms and their usage. Companies usually offer floor plans such as studios, 2-bedroom, 3-bedroom or single-family houses to meet clients’ diverse needs.
2. **With or Without Children:** Families with or without children usually have different preferences regarding room size, nearby amenities, and neighborhoods.
3. **Children’s age:** Families with school-age children may require a separate study room. Families with children of different ages tend to have varying housing preferences.
4. **Household income:** It is important to know how much residents can afford before setting reasonable prices for housing products.
5. **Outstanding Loans:** Local financial institutions may have varying requirements for individuals with or without existing loans. When designing housing products, it's essential to consider the financial resources available to potential buyers.

A clustering model can be highly effective in grouping potential buyers. Real estate companies may identify groups that prioritize home size, pricing, amenities, and other factors when choosing their ideal homes, allowing them to tailor their design and marketing strategies accordingly.

***Question 4.2:***

There are several steps to get the final answer:

Find the best k for clustering:

We plot the relationship between total within-cluster sum of squares (WSS) and k and choose k=3 based on the result.A graph of a number of clusters

Description automatically generated

Fix models (with k = 3) by using different combinations of predictors and then comparing their wss. Of all the 15 models (using combinations including 1, 2, 3, 4, 12, 13, 14, 23, 34, 123, 124, 134, 234, 1234), the lowest wss is 7.86 when the combination of predictors is only the third predictor (Petal.lenth).

Test the result by using the third predictor and k = 3 to run the model. Calculate the correct rate between the model result and the real result, which is 94.7%

**A number grid with black text

Description automatically generated with medium confidence**

In a word, the best combination of predictors is only the third predictor (Petal.lenth), and k= 3, leading to clustering correct rate= 94.7%

**Extra:**

In this case, during the process of finding the best combination of predictors, we can exhaustively

enumerate all possibilities because the number of predictors is small. However, such a condition is not

common in the real world. Therefore, we consider two other methods to find the best combination (codes included in R file):

* Use Correlation to Remove Predictors: Remove a predictor when it has a high correlation with other predictors. In this question, we remove the 4th predictor because it has a correlation of over 0.9 with the third predictor.
* Use Principal Component Analysis (PCA) to Reduce Dimensionality: This commonly used method automatically reduces dimensionality when the number of predictors is large. In this question, we can use PCA to find two principal components and then continue the modeling process.