Collaboration Assignment

Xi Chen

November 13, 2017

Kaggle Open Call Projects

Q1.

I created a Kaggle account with the username as uchixic.

Q2.

The competition "Quora Question Pairs" is very interesting to me, because I always read questions and discussions in Quora during my spare time. However, there is a serious issue that undermines the experience for Quora writers, seekers, and readers: a large amount of duplicate questions. As described in the session of "Overview", not only could duplicate questions cause seekers to spend more time finding the answer to their question, but also make writers feel they need to answer the same questions for a couple of times. Therefore, this competition aims to call for an efficient technique to identify question pairs that have the same intent.

In order to make a submission, firstly, I would need to download the data from the session of "Data", and familiarize myself with the dataset construction by conducting explanatory data analysis with the training data. I would also review the session of "Evaluation" to understand the submission requirement. Secondly, since Quora currently uses a Random Forest model, I would apply the same model to get the baseline results. Then, I would use several classification methods, and compare their performance in predicting if the questions are duplicate. The classification methods I plan to use would be linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), K-nearest neighbor (KNN), logistic regression, and support vector machine (SVM). These analyses could be performed in R. After choosing the best method, I would apply that technique to predict the probability that the questions are duplicates, as suggested in the submission requirement.

The linke to this Kaggle competition:https://www.kaggle.com/c/quora-question-pairs

Q3.

There is a dataset named "Iris Species", which is about classifying iris plants into three iris species: setosa, versicolor, and virginica. The dataset has six columns that descirbe several features for each plant: id, length of the sepal, width of the sepal, length of the petal, width of the petal, and species name.

The link to the dataset: https://www.kaggle.com/uciml/iris

```
# Load the dataset
mydata = read.csv("C:/Users/Xi Chen/Desktop/Perspective - Analysis/Assignment/Assignment 5/Iris.csv")
attach(mydata)
# An overiew of the structure of the dataset
names(mydata)
```

```
## [1] "Id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm"
## [5] "PetalWidthCm"  "Species"
```

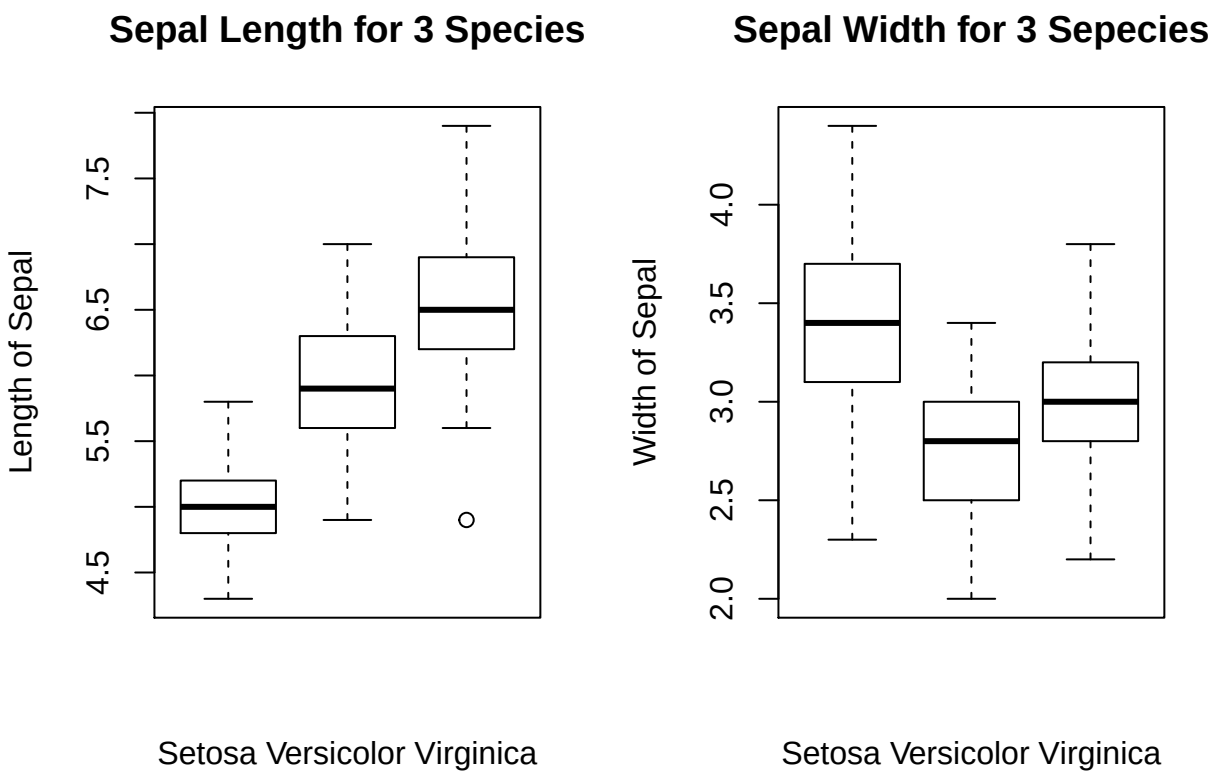
```
head(mydata)
```

```
##   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
## 1  1           5.1          3.5           1.4          0.2 Iris-setosa
## 2  2           4.9          3.0           1.4          0.2 Iris-setosa
## 3  3           4.7          3.2           1.3          0.2 Iris-setosa
## 4  4           4.6          3.1           1.5          0.2 Iris-setosa
## 5  5           5.0          3.6           1.4          0.2 Iris-setosa
## 6  6           5.4          3.9           1.7          0.4 Iris-setosa
```

```
# There are 50 samples for each iris species
with(mydata, table(Species))
```

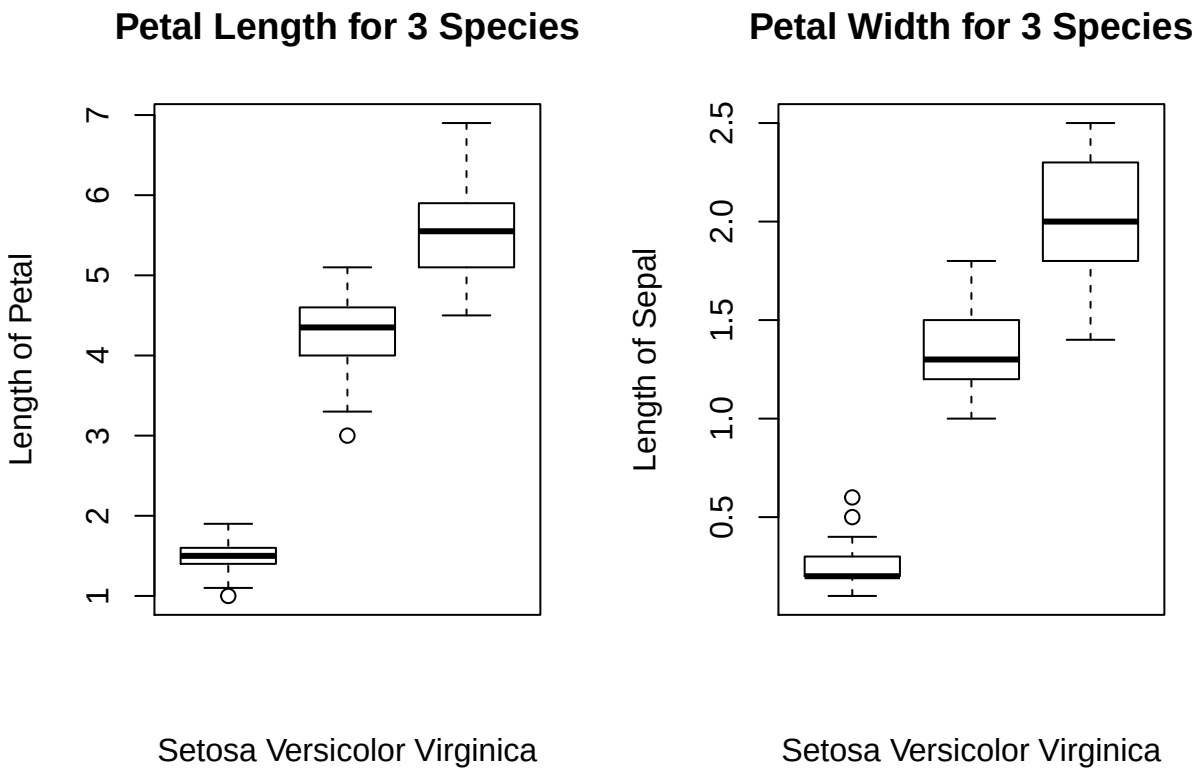
```
## Species
##     Iris-setosa Iris-versicolor  Iris-virginica
##              50              50              50
```

```
# The boxplots show us which features can distinguish the species well
par(mfrow=c(1,2))
boxplot(SepalLengthCm~Species,data=mydata,
        main="Sepal Length for 3 Species",
        xlab="Setosa Versicolor Virginica", ylab="Length of Sepal",
        xaxt="n")
boxplot(SepalWidthCm~Species,data=mydata,
        main="Sepal Width for 3 Sepecies",
        xlab="Setosa Versicolor Virginica", ylab="Width of Sepal",
        xaxt="n")
```




```
boxplot(PetalLengthCm~Species,data=mydata,
        main="Petal Length for 3 Species",
```

```
        xlab="Setosa Versicolor Virginica", ylab="Length of Petal",
        xaxt="n")
boxplot(PetalWidthCm~Species,data=mydata,
        main="Petal Width for 3 Species",
        xlab="Setosa Versicolor Virginica", ylab="Length of Sepal",
        xaxt="n")
```

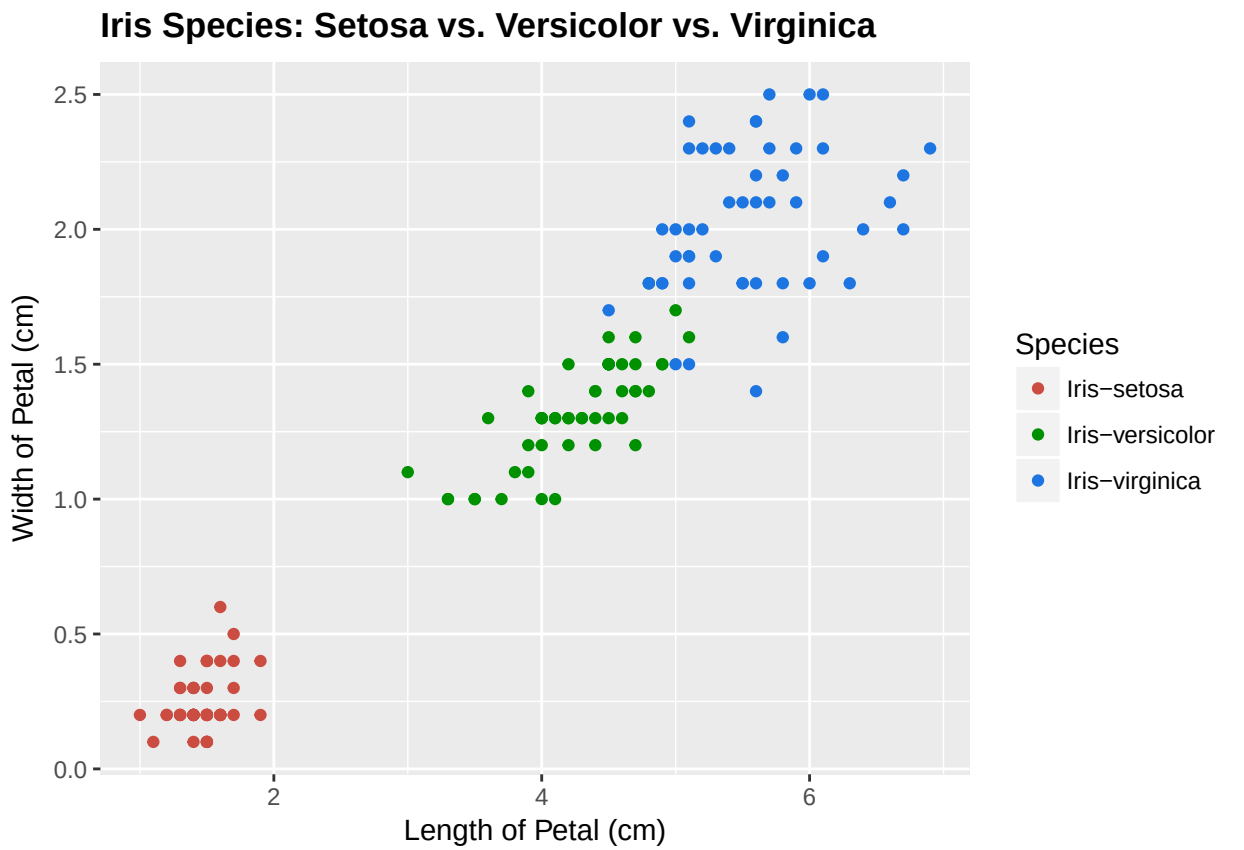


Comments: We can see that "Petal Length" and "Petal Width" are good features to classify three species, so I plot the three iris species by these two features.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
ggplot(mydata, aes(x=PetalLengthCm, y=PetalWidthCm, color=Species)) +
  geom_point() +
  scale_colour_hue(l=50) +
  ggtitle("Iris Species: Setosa vs. Versicolor vs. Virginica") +
  theme(plot.title = element_text(lineheight=.8, face="bold")) +
  xlab("Length of Petal (cm)") +
  ylab("Width of Petal (cm)")
```



Comments: From the above plot, we can see that the iris viginica has the longest and widest petal, the iris versicolor has the second longest and widest pedtal, and the iris setosa has the shortest and narrowest petal. The "Length of Petal" and "Width of Petal" can help us classify the three iris species well.

Improving a Journal Article

Raguhbir and Srivastava (2002) conduct a research about finanical decison-making, and found a systematic difference in people's willingness to pay when using foreign currencies. The results suggest taht, when the nominal value of a foreign currency is a multiple of an equivalent unit of a home currency (e.g., 1,100 Korean Won = 1 American Dollar), people underspend the foreign currency; however, when the nominal value of a foreign currency is a fraction (e.g., 0.89 European Euro = 1 American Dollar), people overspend the foreign currency. They also suggest that "people's willingness to spend and purchase increases or decreases as a function of the relation between the nominal value of the foreign currency and their home currency", according to the exchange rate.

This research is interesting to me. However, during the process of data collection, the researchers recruit U.S. undergraduate business students to participate in the lab study for course credit. For all the six studies, less than one hundred students recruited in each study, such as ninety-seven in study one. My concerns are, firstly, the sample size is relatively small, and secondly, the samples are all from U.S. undergraduate business students, which may not be as representative as expected. Undergraduate students may not have so much experience in financial decision making, especially purchasing with foreign currency. Their familiarity with exchange rate may also bring confounding vairables to the research. Therefore, I would try to reformulate the data collection as a human computation project, which enables the researchers to have access to a large size of more representative samples.

Instead of recruiting students to participant in the lab study, the data would be collected from Amazon

MTurk. Participants would be directed to read a hypothetical scenario, and make a simple purchase decision. This new data collection method would greatly improve this study, because the sample would not only include undergraduate students, but come from a wide population with different ages and backgrouds. With this human-computation way of collecting data, the externality of this research would be highly improved, and we could be more confident to draw conclusions about consumers' decision-making. In addition, researchers can specify the participants' nationality in MTurk. For example, when choosing "UK participants only", researchers can focus on examining UK's consumers' willingness to pay when using foreign currencies. Therefore, this new human-computation method also expand the research's scope and depth. As Salganik (2017) mentions, "Human computation projects are ideally suited for easy-task-big-scale problems". By asking each MTurk participant a simple purchasing scenario question, researches can get a large scale of decision-making data when using foreign currency.

Reference

Raghubir, P., & Srivastava, J. (2002). Effect of face value on product valuation in foreign currencies. Journal of Consumer Research, 29(3), 335-347.

Salganik, Matthew J. (2017). Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press. Open review edition.

Alternative Assignment: InfluenzaNet

Q1.

In the following, I will compare the differences between the traditional reporting systems, InfluenzaNet, and Google Flu from three perspectives.

Design

Traditional influenza tracking systems rely on physicians' incidence records to monitor the influenza trends. Only when the patients go to the hospital to seek medical care, would their cases be reported to the public health and medical systems. However, an internet-based tracking system, InfluenzaNet can potentially monitor a wide range of cases by recruiting volenteers from the general population (Noort et al. 2015). During the influenza season, participants who are afflicted by influenza-like-illnesses would receive newsletter and questionnaires to report their potential symptoms. Another online-based tracking system is Google Flu Trends, which monitors health-seeking behavior in the form of queries to Google search engine (Ginsberg et al. 2009).

Costs

Compared to the internet-based tracking systems, the traditional physician-based reporting system would have highest cost, because it requires doctors, nurses, and other workers affiliated with the health institutions to record and report the medical cases. Google Flue Trends may have the lowest cost, because the data is from the public's queries to Google search engine, instead of the case that doctors manually report the cases to the public health and medical institutions. InfluenzaNet may have relatively middle range of cost, because it doesn't need so much human capital as the traditional systems, but it still needs to manage the online system and database, such as sending out newsletters and questionnaires to collect data.

Likely errors

For the traditional systems, selection bias could be one of the possible errors, because the patients who go to medical centers for help may not be as representative as expected. For example, people from low-income groups and/or lack of comprehensive health insurance would be less likely to visit a doctor when they are sick. For the InfluenzaNet, the data rely on self-reported answers, which may bring validity problems, such as social desirability bias. In addition, everyone has different standards for self-evaluation, which may bring measurement errors. For Google Flu Trends, the potential errors may come from the big data's drawbacks, such as algorithmically confounded.

Q2.

If there is a swine flu outbreak, each system could have similar and/or different errors. For the two internet-based systems, the similar error for them could be "non-representative", because the online population may not represent the general population well. In addition, the Google search engine may have more queries about swine flu at this time. However, it may not correctly indicate the seriousness of the outbreak, but just show that the public pays more attention to this issue. Therefore, the Google Flu Trends may make mistakes in predicting the influenza trend. Besides, if there are more serious outbreaks among the relatively poor areas (which is highly possible according to the history of swine flu), the low-income population are less likely to seek medical care, so the traditional physician-based reporting systems may result in an underestimation of the seriousness of the flu trend. Therefore, no systems are perfect but may have different types of possible errors.

Reference

Tilston, N. L., Eames, K. T., Paolotti, D., Ealden, T., & Edmunds, W. J. (2010). Internet-based surveillance of Influenza-like-illness in the UK during the 2009 H1N1 influenza pandemic. BMC public health, 10(1), 650.

van Noort, S. P., Codeço, C. T., Koppeschaar, C. E., Van Ranst, M., Paolotti, D., & Gomes, M. G. M. (2015). Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour. Epidemics, 13, 28-36.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012-1014.