STA 160 Midterm Project

Xuecheng Zhang
Department of Statistic–STA 160
University of California
Studnt ID: 915942842
hcczhang@ucdavis.edu

Abstract

In this report, I will focus on two data set one is Seed data, another is Automobile data. The main topic of this report is 1) How to compare multiple labels with respect to one single feature? Each label is attached to a 1-dim dataset of feature measurements. 2) How to see intrinsic differences among multiple labels with respect to multiple features? Each label is attached to a K-dim dataset of feature measurements. 3) How to deal with categorical features? 4) How to measure associative relations between a categorical response variable and multiple covariate features.

1 Seed Data set Analysis

In this section I focus on the seed database which contains three kinds of seeds Kama, Rosa and Canadian. It measured seven geometric parameters of wheat kernels: Area, perimeter P, compactness, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove. I will analysis the relationship between each variables and clustering data set.

https://archive.ics.uci.edu/ml/datasets/seeds

1.1 Compare Multiple Labels with Respect to One Single Feature

First of all, I plot a box plot to observe the effect of kernel types on each feature. From the box chart Figure 1, I can see that in general, kernel 2 has the highest value on most variables, and kernel 3 has the lowest value on most variables. However,'coefficient' has a different result: kernel 3 have the highest value, and kernel 1 have the lowest value. 'Compactness' also reveals a different result whereas the value of kernel 1 almost equal to kernel 2 and then followed by kernel 3. In short, the type of kernel affects every parameter in seed data set, Rose has the highest value except coefficient and compactness.

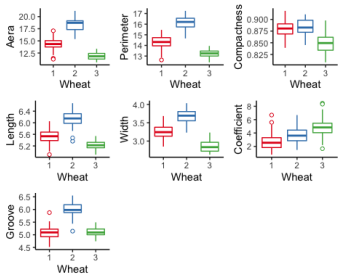

Figure 1: Box Plot of

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

1.2 Intrinsic Differences among Multiple Labels with Respect to Multiple Features

Next, I want to explore whether there is a relationship between each feature, so I spent a correlation matrix plot, from Figure 2, it can be seen that there is a strong linear relationship with the area, perimeter, length, width, this result indicates that this data set has multicollinearity. On another hand, the coefficient seems that have less correlation with other parameters. So I will focus on the area, perimeter, length, and width, analysis is the type of kernel will respect those labels, explore whether the type of kernel has an impact on the distribution in the linear relationship between them.

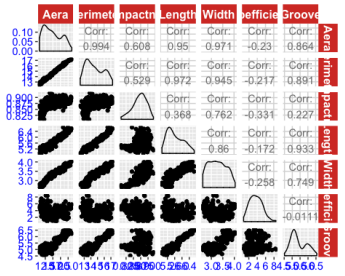

Figure 2: Correlation Matrix of Seed Data

1.3 Three type of Kernels respect to multiple features

First, I concerned about how the kernel's length is related to the kernel's width. From Figure 3, it can be seen clearly that there is the monotonic increasing linear relationship between them.The type of kernel has both effect on the length and width. The kernel 1 is short and narrow, the kernel 3 is long and wide, while the kernel 3 is in the middle.

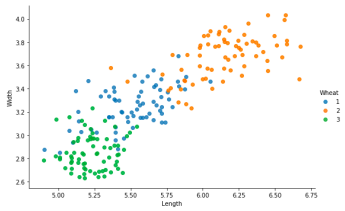

Figure 3: Scatter Plot between Length and Width

However, is the type of kernel will effect both area and perimeter?

Further, I focus on how the kernel's Area is related to the kernel's compactness. From Figure, it can be seen that there is a relationship between area and compactness. The type of kernel affects the area, but compactness is not obvious. In short, kernel 3 has the smallest area and smallest compactness; and the area of kernel 2 is the largest, but the compactness of kernel 2 and kernel 1 are close to each other.

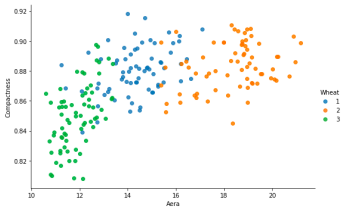

Figure 4: Scatter Plot between Area and Compactness

Therefore, the types of the kernel are obviously distributed in length and width, but not in area and compress, because in the previous box plot, we can see that the types of the kernel have little influence on the compress.

1.4 Heat Map for Hierarchical Clustering Algorithm

The below heat map of the distance matrix shows some clustering. This is no coincidence, since we already know that the seed data set has 3 classes, so they are already nicely sorted.

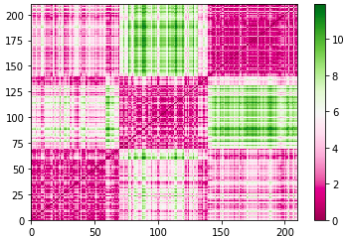

Figure 5: Heat map of Seed Data

we can expect the data belonging to one class to be more similar. However, in standard clustering situations, there is no supervision (no information about classes). So let's consider a random permutation of the data (rows), which is more in line with the unsupervised situation:

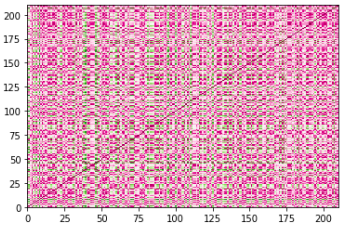

Figure 6: Random Heat map of Seed Data

I use dendrogram to find how many classes we have in our data set, and from the dendrogram we can read there are 3 classes in our data set.

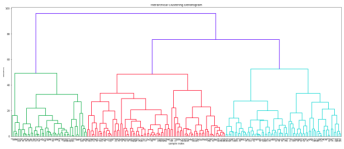

Figure 7: Dendrogram

From heat map of seed data, we can see that there are two obvious green areas in the upper left corner. In the following clustering results, we can see that only single method divides them into the correct label.In addition, in heat map of seed data, the classification on the diagonal is more obvious. After clustering, only Average Method and Ward Method cluster the data on the diagonal.

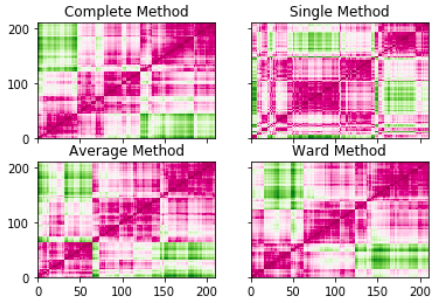

Figure 8: 4 Methods of Hierarchical Clustering

In conclusion, I found that it is not intuitive to use all features for clustering, and the clustering results are not close to the real labels, so I am going to use K-mean for clustering in the next step, and make the image become intuitive and visual.

1.5 K-means Clustering and compare with Hierarchical Clustering

For the K-mean clustering, I need to find the best k values, from the scree plot below, I can see k=3 where the slope of the curve is clearly leveling off (the "elbow) indicates the number of factors that should be generated by the analysis. So I choose k=3.

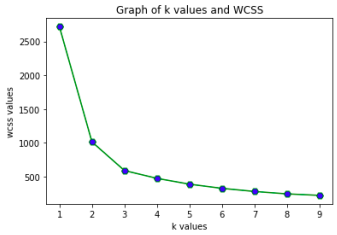

Figure 9: Scree Plot

After choose k value, I did the K-means Clustering and compare with Hierarchical Clustering and Original class. For visualization I will use only two features: Length and Width for the original and predicted data sets. Different classes will have separate color and styles.

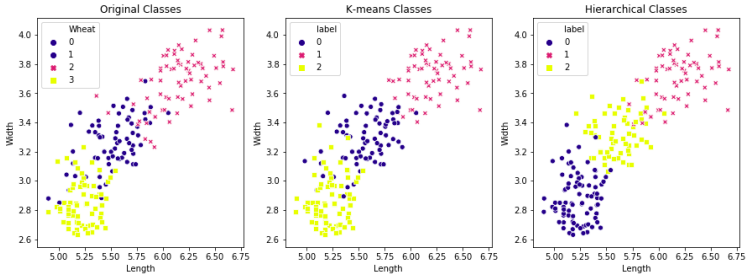

Figure 10: Comparing Original, K-Means and Hierarchical Clustered Classes

From the original figure, we can see that kernel 2 and kernel 1 have overlapped parts, kernel 1 and kernel 3 also have overlapped parts, and kernel 2 and kernel 3 have no overlapped parts. After K-means clustering, it can be seen that it solves the problem of overlapping kernel 1 and kernel 2. After hierarchical clustering, it can be seen that there is a solution, but it is not obvious. Which method is more suitable for length and width clustering? I decided to work out their number and accuracy.

1.6 Conclusion: Clustering

From this Table 1, we can see clearly that when we only consider the length and width of these two features, K-mean clustering is more close to the actual situation than hierarchical clustering. So when only considering the length and width, I suggest K-mean to cluster the data.

Table 1: Predicted Data VS Original Data

Name	Kernel 1	Kernel 2	Kernel 3	Rate
Original Data	70	70	70	100%
K-Means Predicted Data	77	72	61	95.71%
Hierarchical Predicted Data	86	63	61	92.38%

2 Automobile Data set Analysis

In this section I focus on the automobile database which consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. I will explore which futures can influence price changes in numerical features and categorical features.Then I'll build models of prices and find the one that best suits the actual situation.

`http://archive.ics.uci.edu/ml/datasets/Automobile`

2.1 Data Preparation

First of all, I need to solve the problem of missing data, It can clearly see that both numerical features and categorical features have missing data.I choose to fill missing numerical data of normalized-losses, price, horsepower, peak-rpm, bore, stroke with the respective column mean, and fill missing data category Number of doors with the mode of the column.

Table 2: Sum number of missing value

Feature	Missing
Normalized$_{l}osses$	41
Num-of-doors	2
Bore	4
Stroke	4
Horsepower	2
Peak-rpm	2
Price	4

2.2 Numerical Features Analysis

Before analysis numerical data, there are several strongly correlating with each other columns, which could be combined together:$mpg = \frac{city_{mpg}+highway_{mpg}}{2}$ and curb-weight is somehow based on linear sizes: length, width, wheel-base, so I create a now numerical column named mpg and use curb-weight to represent length, width, wheel-base. After that I draw a heat map of cross correlation plot to see the correlation between numerical data. From the plot, I know that mpg, hp, curb-weight, and engine-size have higher correlation with price(mpg have negative correlation).

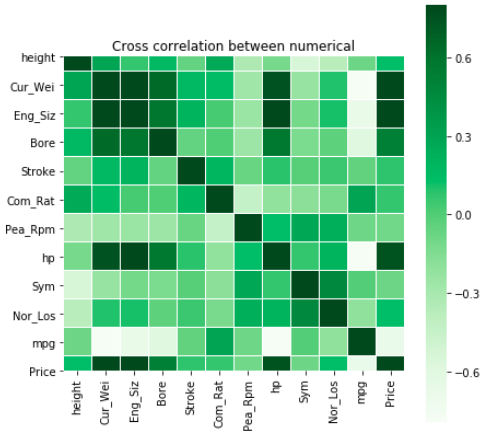

Figure 11: Cross Correlation

I use Pearson Correlation to measures the linear dependence between two variables X and Price, I also use p-value which the probability value that the correlation between these two variables is statistically significant. (see the table below)

Table 3: Correlation P-value

Name	Correlation	P-value
Height	0.1343877957091601	0.05471931679041278
Curb-weight	0.8208247364886393	2.8663274583452043e-51
Engine-Size	0.8617522313557833	9.66974309680653e-62
Bore	0.5323000757708254	2.139812289831734e-16
Stroke	0.08209537049469544	0.2419141372709262
Com-ratio	0.07099045143780537	0.31178188611196345
horsepower	0.7579456217935241	1.591033244659641e-39
Peak-Rpm	-0.10084584127381386	0.14520557459046376
Symboling	-0.08220134587451062	0.24130482805344847
Nor-Losses	0.13399873429239528	0.05543057826382352
MPG	-0.6842007110736521	1.2003440437917137e-29

From the table, we can see that Curb-weight, Engine-size, Horsepower, and MPG, their p-value is lt; 0.001, the correlation between them and price is statistically significant, and the linear relationship is quite strong (>0.6, close to 1).

I choose use box-plots to visualize categorical variables. The below plot shows the relationship between categorical variables and Price. Let's take a look at which variable affects the price. From figure 12, we see that the distributions of price between the different make, aspiration, number of doors, fue-type categories have a significant overlap, and so they would not be a good predictor of price. We still have some categorical variables, so I do the histogram plot next to see their distributions.

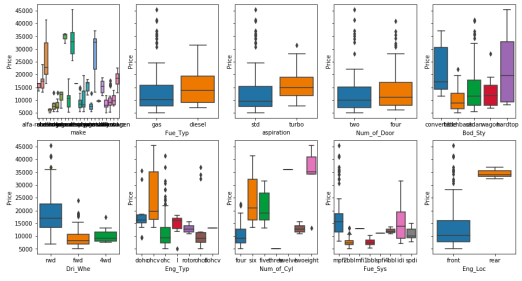

Figure 12: Box Plots with Categorical Variables

From figure 13, the data set is quiet unbalanced in body-style, engine-location, engine-type, num-of-cylinders and fuel-system. We can combine rare values to add more balance to the data. Also there very few examples of vehicles with rear engine, so we won't use this feature in the model.

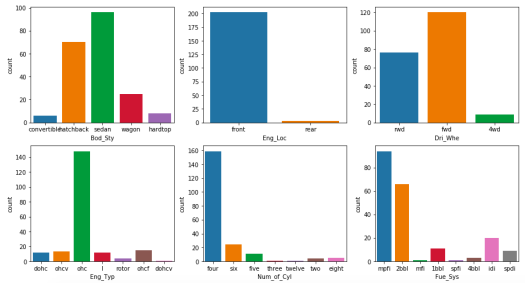

Figure 13: Histogram Plot with Categorical Variables

Finally, I do the ANOVA test for Drive Wheels variable(rwd,fwd,4wd)to get the F-value and p-value. The ANOVA results: F= 67.50368965539572 , P = 3.5392223627637895e-23, with a large F test

score showing a strong correlation between Drive Wheels and Price, and a P value of almost 0 implying almost certain statistical significance.

2.3 Conclusion: Important Variables

After analysis, we can find out which variables are important factors in the prediction of automobile price. I have narrowed it down to the following variablesCurb-weight Engine-size Horsepower, mpg (continuous numerical variables), and Drive-wheels(categorical variables). Next, I'll start building machine learning models to automate our analysis.

2.4 Build Machine Learning Models

First I build linear regression, to avoid multicollinearity, I calculated the Variance inflation factor. From the Figure 14, we know there is not multicollinearity in our variables.

	VIF	Tolerance
hp	4.577590	0.218456
Cur_Wei	5.702634	0.175358
Eng_Siz	5.301557	0.188624
mpg	3.832585	0.260920
Dri_Whe	1.543687	0.647800

Figure 14: Multiple linear regression

So I use Curb-weight Engine-size Horsepower, mpg (continuous numerical variables), and Drive-wheels(categorical variables) to build a Multiple linear regression regression.

From the figure 15, we know that the pearson R is 0.89 and the p-value is 1.2556e-14.

Next, I try Lasso Regression, the pearson R is 0.89 and the p-value is 1.2557e-14.

And ElasticNet Regression,the pearson R is 0.88 and the p-value is 2.3955e-14.

From the three models, we know that fitted values are reasonably close to the actual values, since the two distributions overlap a bit. However, there is definitely some room for improvement.

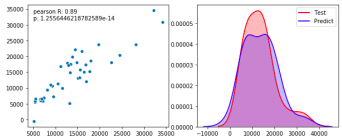

Figure 15: Multiple linear regression

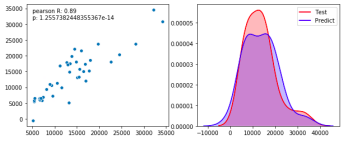

Figure 16: Lasso Regression

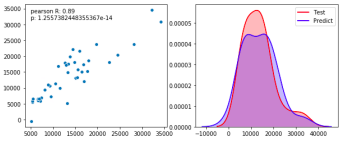

Figure 17: ElasticNet Regression

2.5 Conclusion:Machine Learning Model

Finally I need mean squared error and R2 score to decided which model is better for our data.As is known to all, the model with the higher R-squared value is a better fit for the data and with the smallest MSE value is a better fit for the data. So ElasticNet Regression model is better for our Automobile data set.

Table 4: MSE and R2

Name	Mean squared error	R2 score
Multiple linear regression	11831716.169313844	0.7574993395357078
Lasso Regression	11829931.12465724	0.7575359254799605
ElasticNet Regression	11545386.447921457	0.7633679004067244

3 Conclusion

In seed data, different labels will affect different features, whether single feature or multiple features and k-means clustering is good to cluster data by length and width features. The biggest problem I encountered in this data is how to cluster data. At first, my idea was to choose the hierarchical clustering algorithm, and then analyze four methods: average, ward, single, complete. Which one of them can better cluster, but the result is that they can not cluster well. I think of two ways to solve this problem. One is to change the clustering method, and the other is to replace all features with two or three features, so as to better visualize the data. In the Automobile Data set, I didn't know where to start to analyze the data at the beginning. Later, I separated the digital factor and 2 for analysis, so that the results can be clearly seen. I find out which variables are important factors in the prediction of automobile prices. I have narrowed it down to the following variablesCurb-weight Engine-size Horsepower, mpg(continuous numerical variables), and Drive-wheels(categorical variables. In terms of model building, we can see that my models are all under fit testing data, and the reason may be that I have removed too many features, leading to underfitting, so I think my model has many improvements.