Math 5364

Data Mining 2

Homework 21

Mary Barker

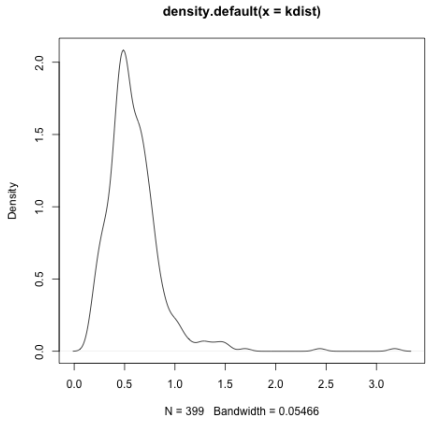
1. Remove all categorical variables from the Auto MPG data set from the UCI Machine learning Repository, and then examine this data set for anomalies, using the following methods for computing outlier scores:

   - Distance for k-nearest neighbor, where k = 5
   - Density
   - Local Outlier Factor method.

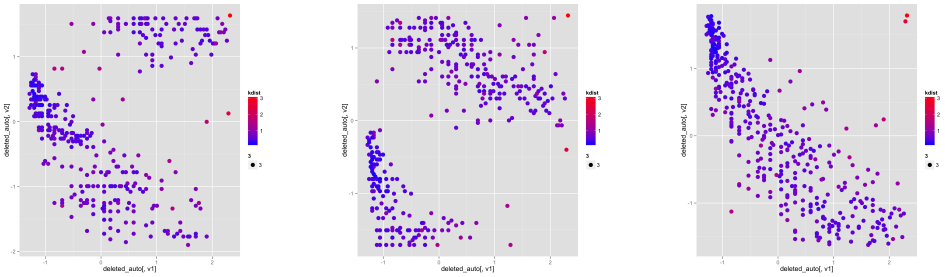
   For each of these methods, plot the density of the outlier score, and produce scatterplots with a heat map of the outlier scores.

   Do you have any comments on this data set after performing the outlier analysis?

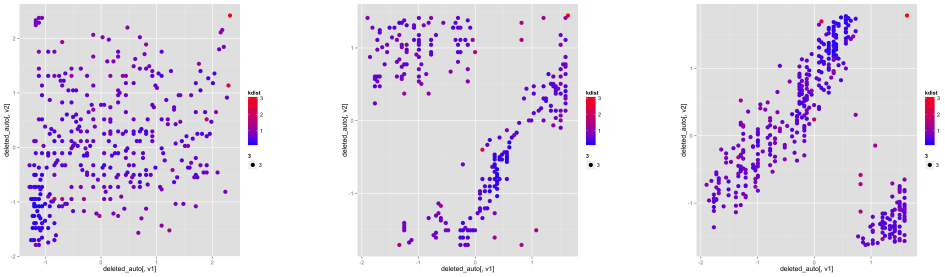
   - Using distance, there were two likely looking places for determining cutoff for clusters as can be seen from the plot below. The first was at 1.5, and the second at 1.25.



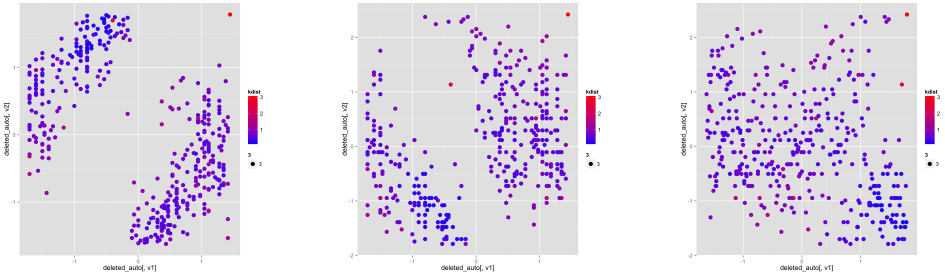
   The plot of values for each point using distance as a metric is shown below.



(i) (ii) (iii)

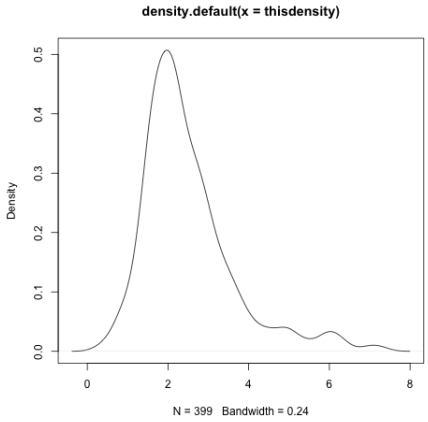


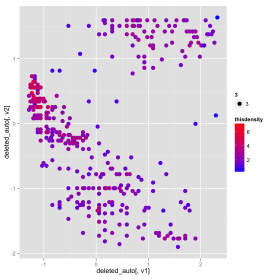
(iv) (v) (vi)



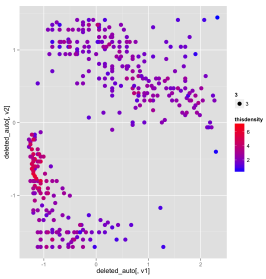
(vii) (viii) (ix)

- Using density as a metric, there are 3 likely cutoff points. The three considered here are values of 1.8, 2.0, and 2.5.

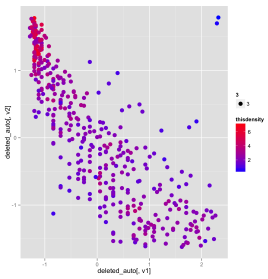




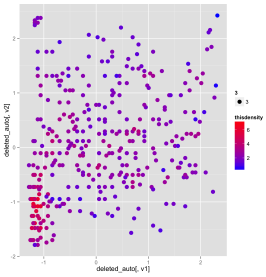
(x)



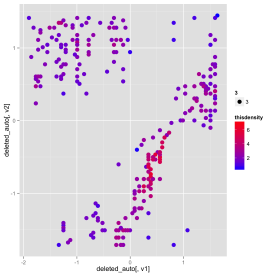
(xi)



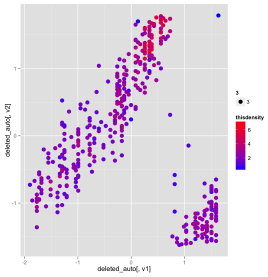
(xii)



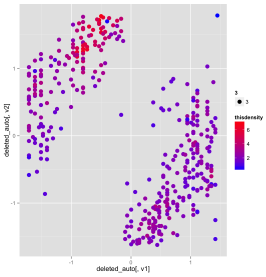
(xiii)



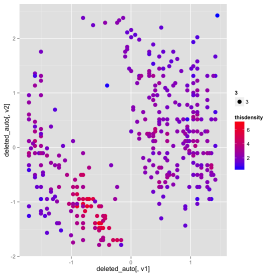
(xiv)



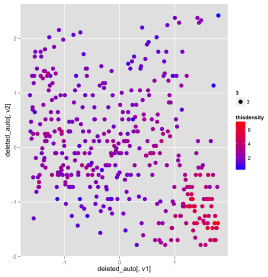
(xv)



(xvi)



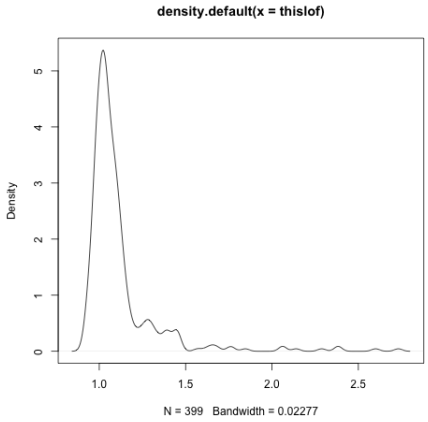
(xvii)

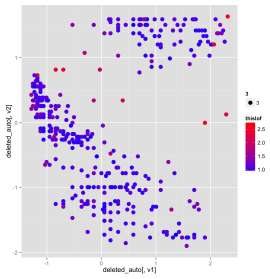


(xviii)

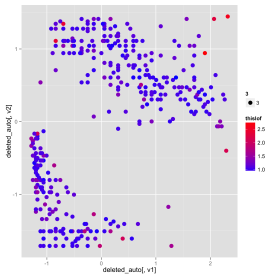
- Using local outlier factor as a metric, there are 3 likely cutoff points. The three

considered here are values of 1.8, 2.0, and 2.5. For the ggplot visualization, the cutoff was set at 2.0.

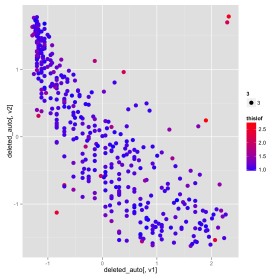




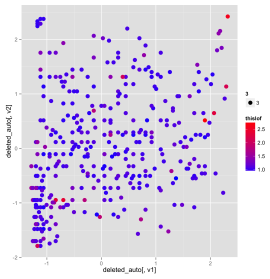
(xix)



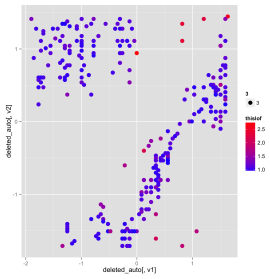
(xx)



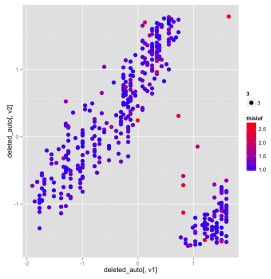
(xxi)



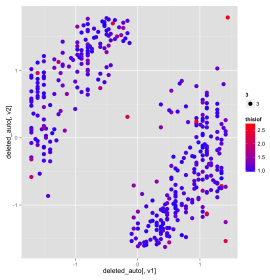
(xxii)



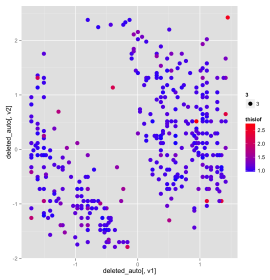
(xxiii)



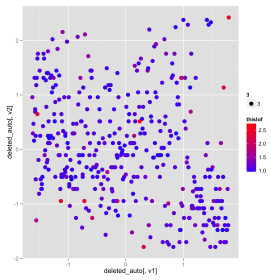
(xxiv)



(xxv)

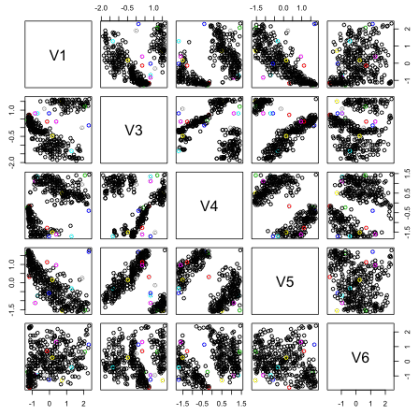


(xxvi)



(xxvii)

The entries selected as outliers are shown below plotted in color, while the remaining entries are colored black.



The values categorized as outliers are shown below for each of the three methods with the different cutoff values discussed.

Metric	cutoff value	outliers
Distance	1.25	1, 30, 73, 113, 211, 245, 329, 335, 336, 359, 366, 389
	1.5	1, 30, 113, 389
Density	5	41, 42, 43, 44, 45, 46, 65, 66, 67, 70, 71, 89, 94, 117, 158, 190, 192, 210
	5.5	41, 43, 44, 45, 46, 65, 66, 67, 70, 89, 117, 190, 192, 210
LOF	1.5	1, 15, 27, 30, 73, 113, 205, 245, 265, 300, 332, 336, 359, 366, 388, 389

The nodes chosen as outliers for Density were entirely different from those chosen by Distance metric or by local outlier factor metric. There were some in common between the latter two, although not all the same either. The variance in which entries were determined to be outliers between the three metrics makes it very difficult to tell whether any one of the values is indeed an outlier.

There is some likelihood that the entries in rows 1, 113, and 389 are outliers as they occur in both of the rows associated with the distance metric as well as that associated with the local outlier factor metric.

```
#Data Mining hw 21

source('~/Dropbox/Tarleton/data_mining/generic_functions/dataset_ops.R')
source('~/Dropbox/Tarleton/data_mining/class_notes/outliers.R')
auto <- read.csv("~/Documents/testout/data_mining/dfiles/auto_data.csv",
                  header=F,  na.strings='?', dec='.', strip.white=T)


#1. Remove all categorical variables from the Auto MPG data set from the UCI Machine
#   learning Repository, and then examine this data set for anomalies, using the
#   following methods for computing outlier scores:

#    * Distance for k-nearest neighbor, where k = 5

#    * Density

#    * Local Outlier Factor method.

#   For each of these methods, plot the density of the outlier score, and produce
#   scatterplots with a heat map of the outlier scores.

#   Do you have any comments on this data set after performing the outlier analysis?
    cols = c(1, 3, 4, 5, 6)
    combinations <- combn(cols,2)
    name <- names(deleted_auto)
    path='~/Documents/testout/data_mining/hw21/'

    for(i in cols){
        auto[,i] <- as.numeric(auto[,i])
    }

    idx = auto[,4] != 'NA'
    deleted_auto <- auto[idx == TRUE,]
    deleted_auto <- standardize(deleted_auto, cols)

    ###################################################################
    # knn:
    ###################################################################
    kdist = my.kdist(deleted_auto[,cols],5)
    plot(density(kdist))
```

```
dev.copy(png,'~/Documents/testout/data_mining/hw21/density_kdist.png')
dev.off()

(1:nrow(deleted_auto))[kdist> 1.5]
(1:nrow(deleted_auto))[kdist>= 1.25]

for(i in 1:ncol(combinations)){
    v1 = combinations[1,i]
    v2 = combinations[2,i]
    plots <- ggplot(data=deleted_auto[,cols],
                    aes(x=deleted_auto[,v1],y=deleted_auto[,v2],col=kdist,size=3))+
                    geom_point()+
                    scale_colour_gradientn(colours=c('blue','red'))
                    ggsave(plots, filename=paste(path,'using_kdist',v1,'_',v2,'.png',sep=''))
}


#####################################################################
# density:
#####################################################################
thisdensity = my.density(deleted_auto[,cols],5)

for(i in 1:ncol(combinations)){
    v1 = combinations[1,i]
    v2 = combinations[2,i]
    plots <- ggplot(data=deleted_auto[,cols],
                    aes(x=deleted_auto[,v1],y=deleted_auto[,v2],col=thisdensity,size=3))+
                    geom_point()+
                    scale_colour_gradientn(colours=c('blue','red'))
                    ggsave(plots, filename=paste(path,'using_density',v1,'_',v2,'.png',sep=''))
}


plot(density(thisdensity))
dev.copy(png,'~/Documents/testout/data_mining/hw21/density_density.png')
dev.off()

(1:nrow(deleted_auto))[thisdensity>=5]
(1:nrow(deleted_auto))[thisdensity>=5.5]
```

```
    ##################################################################
    # local outlier factor:
    ##################################################################
    thislof = lofactor(deleted_auto[,cols],k = 5)
    for(i in 1:ncol(combinations)){
        v1 = combinations[1,i]
        v2 = combinations[2,i]
        plots <- ggplot(data=deleted_auto[,cols],
                        aes(x=deleted_auto[,v1],y=deleted_auto[,v2],col=thislof,size=3))+
                        geom_point()+
                        scale_colour_gradientn(colours=c('blue','red'))
                        ggsave(plots, filename=paste(path,'using_lof',v1,'_',v2,'.png',sep=''))
    }

    plot(density(thislof))
    dev.copy(png,'~/Documents/testout/data_mining/hw21/density_lof.png')
    dev.off()

    (1:nrow(deleted_auto))[thislof >= 1.25]
    (1:nrow(deleted_auto))[thislof >= 1.5]
    ##################################################################


outliers = (thislof>=1.5)
coloring=rep('black',nrow(deleted_auto))
coloring[outliers] <- 1:sum( 1.0 * (outliers))
plot(deleted_auto[,cols], col=coloring)
dev.copy(png, '~/Documents/testout/data_mining/hw21/all_plot.png')
dev.off()
```