Density estimation. GMM. DBSCAN (Assignment)

Pedro Delicado

The file `BikeDay.Rdata` contains information on the bike-sharing rental service in Washington D.C., USA, corresponding to years 2011 and 2012. This file contains only one data frame, `day`, with 731 rows (one for each day of years 2011 and 2012, that was a leap year) and 16 columns:

- `instant` row index, going from 1 to 731
- `dteday` date
- `season` (1:springer, 2:summer, 3:fall, 4:winter)
- `yr` year (0: 2011, 1:2012)
- `mnth` 1 for January, until 12 for December
- `holiday` weather day is holiday or not
- `weekday` day of the week (0 Sunday to 6 Saturday)
- `workingday` if day is neither weekend nor holiday is 1, otherwise is 0
- `weathersit` :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- `temp` Normalized temperature in Celsius. The values are divided to 41 (max)
- `atemp` Normalized feeling temperature in Celsius The values are divided to 50 (max)
- `hum` Normalized humidity. The values are divided to 100 (max)
- `windspeed` Normalized wind speed. The values are divided to 67 (max)
- `casual` count of rental bikes by causal users (not registered)
- `registered` count of rental bikes by registered users.
- `cnt` count of total rental bikes (casual + registered)

In particular we are interested in the joint distribution of `temp` and `casual` for year 2012:

```
load("BikeDay.Rdata")
X <- as.matrix(day[day$yr==1,c(10,14)])
#pairs(X)
```

Questions

1. Use library `mclust` for the following task. Do a model based clustering of these data assuming a Gaussian Mixture Model, allowing varying volume, shape, and orientation for different components in the mixture. Choose k_{BIC}, the best number of clusters $k \in \{2, \dots, 6\}$ according to BIC.
   Plot the resulting object from `Mclust` (do 4 different graphics: `BIC`, `classification`, `uncertainty` and `density`).

2. Compare the previous `density` plot with the non-parametric density estimation of (`temp`, `casual`) obtained when using the kernel estimator implemented in `sm::sm.density`, with the bandwidths proportional to the standard deviations in both dimensions:
   $h = a \cdot (StdDev(\texttt{temp}), StdDev(\texttt{casual}))$. *Use* $a = 0.25$.

3. For each one of the k_{BIC} clusters obtained above, do the following tasks *(A unique plot should be done, at which the k densities are represented simultaneously)*:

- Consider the bivariate data set of the points in this cluster.
- Estimate non-parametrically the joint density of `temp` and `casual`, conditional to this cluster, using the kernel estimator implemented in `sm::sm.density` with the bandwidths proportional to the standard deviations in both dimensions: $h = a \cdot (StdDev(\mathtt{temp}), StdDev(\mathtt{casual}))$. *Use $a = 0.4$ and compute the standard deviations at each cluster.*
- Represent the estimated bivariate density using the level curve that covers the 75% of the points in this cluster.

4. Use library `fpc` to check if it is possible to merge some of the components in the Gaussian Mixture Model previously estimated. Let k^* be the final number of clusters after the merging process. Do the scatterplot of (`temp` , `casual`) with colors according to the new k^* clusters.
   Indication: Use the function `mergenormals` with the option `method="bhat"` .

5. For each one of the k^* clusters obtained above, do the following tasks *(A unique plot should be done, at which the k densities are represented simultaneously)*:

- Consider the bivariate data set of the points in this cluster.
- Estimate non-parametrically the joint density of (`temp` , `casual`), conditional to this cluster, using the kernel estimator implemented in `sm::sm.density` with the bandwidths proportional to the standard deviations in both dimensions: $h = a \cdot (StdDev(\mathtt{temp}), StdDev(\mathtt{casual}))$. *Use $a = 0.4$ and compute the standard deviations at each cluster.*
- Represent the estimated bivariate density using the level curve that covers the 75% of the points in this cluster.

6. Use DBSCAN to find clusters (and outliers) in the data set (`temp` , `casual`), after **centering and scaling** both variables (do `Xs <- scale(X)`). Try $\varepsilon \in \{0.25, 0.5\}$ and $\mathbf{minPts} \in \{10, 15, 20\}$.
   Which combination of the tuning parameters do you consider the *best one*?
   Compare the DBSCAN clustering corresponding to your favorite combination of tuning parameters with the results of `mergenormals` (print their cross-table).

7. Give an interpretation (or explanation, or description) of the clusters your have found before.
   *(**Indication:** Other variables in the data set can help to describe the clusters).*