

Mini-project assignment
MS1b Statistical Data Mining

Deadline

Mini-project assignments for *MS1b Statistical Data Mining* are due to be handed in at the Examination Schools by **12 noon on Monday 2 May 2011 (week 1 Trinity Term)**.

The details of this mini-project are on pages 2–4 below.

Mini-project MS1b Statistical Data Mining, HT 2011

Traffic Jam Prediction

The aim of the project is some basic form of traffic jam prediction, using historic data of traffic flow on a motorway/highway.

Travel times of passing cars are recorded for different locations of the motorway. Each location has a start and end and the time it takes each car to pass between the start and end point is recorded. The average travel time of cars in each 3-minute bin is reported. The unit is deciseconds. A traveltime of 210 for a given location means thus that it took cars on average 2100 seconds to pass between start and end point of this recording location.

Observations are made at 10 different locations 1,...,10 and at different times of the year 2010. Days are numbered consecutively so that day 1 is 1 January 2010 and day 365 is 31st December 2010. Minutes within a day are numbered consecutively from midnight. A time 10:00 correspond to minute 600 and 23:59 to minute 1439.

There are three files for this project:

- The file `TravelTimes.csv` is a 1560x100 dimensional matrix. Each row corresponds to a specific location on a given day of the year. The column entries in each row show the average travel times on 100 different times of this day, starting at 9am and updating every 3 minutes until 14pm (minutes 541 to 838).
- The file `LocationTimeType.csv` is a 1560x4 dimensional matrix, giving for each row in the matrix in file `TravelTimes.csv` the location (a number in $\{1, \dots, 10\}$) in the first column, the day of the year in the second column, the day of the week in the third column, and finally a class variable Y in the fourth column. The class variable takes values in $\{1, 2\}$ and indicates whether traffic was free-flowing at 2:30pm (in which case $Y=1$) or not (in which case $Y=2$).
- The file `JamLocation1.csv` is a 156-dimensional vector and contains information about the speed of traffic at 14:15 for all rows in matrix `TravelTimes.csv` of location 1 (there are exactly 156 such observations for location 1 on different days of the year). If traffic was free flowing, this class variable takes value 0, for minor delays it takes value 1 and for severe delays value 2.

All three files are comma separated files and can be read in R using the commands

```
> read.table(file= "TravelTimes.csv",      header=TRUE,  sep=",")
> read.table(file= "LocationTimeType.csv",header=TRUE,  sep=",")
> read.table(file= "JamLocation1.csv",     header=FALSE, sep=",")
```

You can download the files from the website

<http://www.stats.ox.ac.uk/~meinshau/MS1/miniproject>

You should produce a report that covers the tasks below.

Take extra care to ensure that your report is clear on how you have performed the analysis, so that an interested reader could reproduce your results. For example, if a classifier needs a choice of tuning parameters, describe how you set these tuning parameters, whether you used cross-validation (and which form of cross-validation, if so), and so on. The report *should not* contain computer code. You may (and are encouraged to) include figures.

Task 1: Multi-dimensional scaling Using the dataset `TravelTimes.csv`, produce a 2-dimensional embedding of the data, using metric multi-dimensional scaling. Would you scale the data here for the multi-dimensional scaling?

Once you have computed the embedding, plot it twice. In the first plot color observations according to the day of the week they are made (using information from the file `LocationTimeType.csv`). In the second plot, color observations according to at which location (out of 1,...,10) they were made. Does the multi-dimensional scaling plot show a grouping and which grouping does it discover to first approximation? **20 marks**

Task 2: Principal Component Analysis Compute the PCA of the unscaled data in file `TravelTimes.csv`. How many components do you need to retain to explain 99% and 99.9% of the variance in the data respectively? Plot the first 5 principal components as a function of the minute of the day. What is the characteristic feature of the first principal component? What is the proportion of all observations that show the pattern of the first principal component? Make a comment on the usefulness of standard PCA for this dataset. **15 marks**

Task 3: LDA analysis Use Linear Discriminant Analysis (LDA) for traffic prediction at location 1. The class variable in the file `JamLocation1.csv` takes one of three values and gives an indication of traffic flow 15 minutes after the end of the observational period, that is at 2:15pm (0: free flowing traffic, 1:minor delays, 2:severe delays).

Use the 156 observations that were made at location 1 as predictor variables in a LDA analysis.

Find and plot the two linear discriminant vectors. Do they look smooth? Project the data into this two-dimensional space, indicating the class of each observation by colour or otherwise. Does LDA seem to be able to find a separation between days with free flowing traffic and minor and severe delays?

What is a major assumption in LDA that is not made in Quadratic Discriminant Analysis (QDA)? Make plots to inspect visually whether this assumption is approximately fulfilled for this dataset. Does LDA or QDA seem more appropriate for this dataset? What would be a challenge when trying to apply QDA to this dataset?

20 marks

Task 4: Classifier comparison For this task, expand your dataset by using all 1560 observations, ignoring the fact that they came from different locations. The class variable

is now the fourth column in the file `LocationTimeType.csv`, giving a binary indicator of whether traffic was free flowing with only minor delays (class 1) or whether there were severe delays (class 2).

Fit LDA and Random Forest to 1000 randomly chosen observations and produce a ROC curve for the classifier output for the remaining 560 observations.

Which of the two estimators would you prefer for predictive accuracy ? **20 marks**

Task 5: Extend the model Can you extend the prediction model from task 4 with the data you currently have in any useful way? You could for example implement quadratic or regularised discriminant analysis or use the location information for better prediction with Random Forests. You can also derive new “features”/variables from the data (e.g. the trend over the past observations before 2pm, the mean travel time over the day, the behaviour of travel times at other locations for the same day, make better use of the time-series aspect of the data etc.). Does your out-of-sample predictive accuracy on the test dataset increase with any of these ideas? It is not required that it does; you should just describe what you have tried. You could also check whether you can reduce the number of features and make the model simpler without reducing predictive accuracy unduly. You are not expected to try all of these ideas and you are free to follow your own ideas.

25 marks