CHARACTERISTICS IN FLIGHT DATA - ESTIMATION WITH LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINES

Claus Gwiggner, Ecole Polytechnique, LIX, Palaiseau, France

Gert Lanckriet, University of Berkeley, EECS, Berkeley, California

Abstract

We analyze data from flight sectors. The questions are whether there are differences between weekend and weekdays and among sectors. We compare expected prediction errors of linear logistic regression and of linear and non linear kernel classifiers. Linear decision boundaries impose an average prediction error of around around 26 % for the weekend data and around 15 % for the sector name data. Non linear boundaries do not improve the predictive accuracy by more than 4 %. Thus, there is some characteristic in the data which is identified by both methods.

General Background

Airspace is divided into geographical regions, called sectors. For safety reasons, no more than a certain number of aircraft is allowed to enter certain sectors during one hour. Such numbers are called sector capacities. Airlines pose a demand to enter sectors before take-off by submitting a flight plan to a control center. A flight plan is essentially a time stamped list of way-points. When demand is higher than capacity either take-off is delayed or aircraft are rerouted. We speak of *initial demand* and *regulated demand* of a sector.

Although pilots have to follow their flight plans, there are differences between the number of aircraft planned to enter sectors and the number that really entered them (the *real demand*). By consequence, safety is not always guaranteed and available capacity is not always optimally used.

We call these differences *planning differences*. They are consequences of uncertain events like weather conditions, delays, en-air reroutings or more. Such events are not taken into account by the current traffic planning. If there are regularities in planning differences, they can be used to improve current traffic planning.

Data Description

We focus on four sectors in the upper Berlin airspace where planning differences are reported to occur. The sectors are roughly equal in size. The average traversal time of a sector is ten minutes. We use regulated demand (number of aircraft planned to enter a sector) and real demand data (number of aircraft that really entered a sector) counted in intervals of 60 minutes for a total of 141 weekdays and 68 weekend days in the period June 2003-April 2004 of the four sectors EDBBUR1-4.

Approach to Uncertainty

We consider the data as a finite number of realizations of random variables [1]. More precisely, we define $REAL^{S}_{t1;t2}$ = 'number of aircraft entering sector **S** between t1 and t2' for the real demand. Similarly, we define REG for regulated demand and $DIFF = REAL - REG$ for the planning differences. A sector is thus represented by a vector of random variables, one variable for each time interval.

Hypothesis of the paper

Planning differences show irregular patterns in every sector: Figures 1 and 3 display eleven days of planning differences for the sectors EDBBUR2 and 3 respectively. However, their empirical probability distributions turn out to have similar shapes invariantly of time and sector [1] (figure 2 shows histograms of the first twelve hours of the day of planning differences for sector EDBBUR2, the shapes of distributions of the other three sectors are similar). Our hypothesis is that planning differences have the same characteristics between sectors and on various days. If this is the case, we

[1] for a definition of terms from probability theory and statistics we refer to [4] or any introductory book of the subject

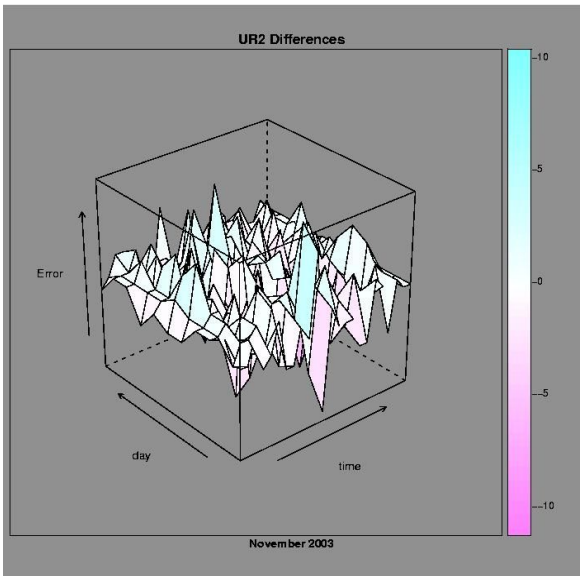

Illustration 1 Planning Differences UR2

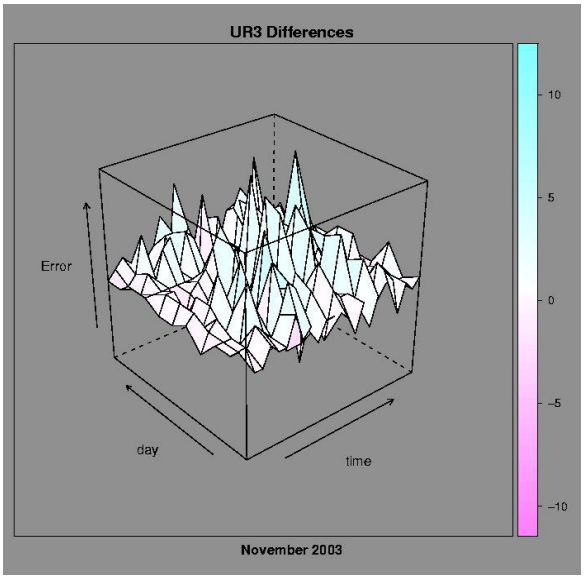

Illustration 3 Planning Differences UR3

can reasonably simplify any future model of planning differences. We formulate the question as a binary classification problem. If we find classes, the hypothesis is unlikely to be true. We are in an exploratory phase of analysis. Formal statistical inference is not intended.

This paper is organized as follows: in the next section we explain briefly the ideas behind logistic regression and support vector machines and why we compare these two classification techniques against each other. We then explain the experimental settings and our results. Finally, we give conclusions and ideas for future work.

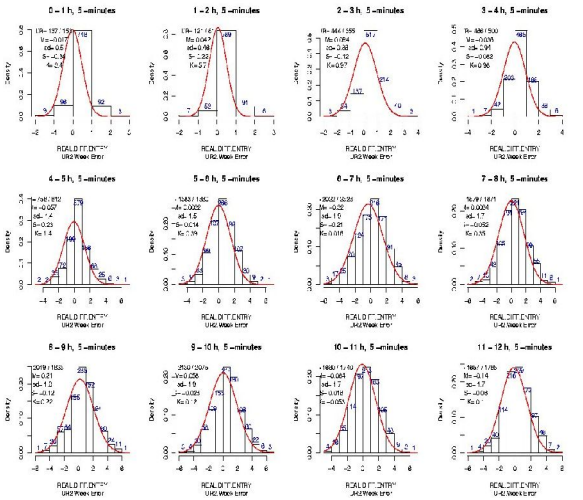

Illustration 2 Similar Shape of Distributions over time

Binary Classification

In a binary classification problem, one has a set of known items belonging to one of two classes 1 and -1 and one likes to predict for a new item, to which class it might belong. We speak of the vector of predictor variables **X** (the 24 hourly variables representing a sector) and the class variable Y (representing the question whether a vector is from a weekday or weekend or from a given sector). There are different approaches to this problem. Some of them refer to a probabilistic and others to a geometric interpretation.

In this section we review the ideas behind one out of each; logistic regression and support vector machines. The different approaches are related to each other. For more detailed information about classification we refer to [2] or [3].

Logistic Regression

Logistic regression explicitly assumes a functional form for class probabilities:

$$P(Y=1|X=x)=\frac{e^{F(X)}}{1+e^{F(X)}}$$

The equation means that a functional relationship F between an item x and the probability, that it belongs to class 1 is assumed. Since there are only two classes, the probability for class -1 follows directly. The fraction on the right side is a transformation that guarantees that the probability

estimates lie in [0,1] for any value of F(x). This transformation is well known in statistics and discussed for example in [5], [6]. Linear logistic regression [7] is a special case, where $F(x)=\beta x$ is chosen to be a linear function. The parameters of the linear model are usually estimated by binomial maximum likelihood. Other forms for F are studied in [4]. Prediction with logistic regression models translates into the question: given a new measurement, what is the probability that it belongs to class 1?

Logistic regression is a popular technique in data analysis and known since the early 19th century [5]. The goal is to understand the role of the predictor variables in explaining the class.

Support Vector Machines (SVMs)

It is also possible to characterize classes by their boundaries that they draw in Euclidean space. For the binary classification problem, only one boundary characterizes the whole solution. Class 1 lies on the one side and class -1 on the other side of the boundary.

The idea behind support vector machines is to directly estimate decision boundaries as a function of the predictor variables. The result is a classifier mapping input values to the one or the other side of the boundary.

SVMs overcome a number of limitations of other related techniques, namely the linearity of the decision boundaries and the problem of overlapping classes [2].

We only characterize informally how SVMs work and refer to literature for detailed information (e.g. [2],[8]). Nonlinear decision boundaries are found by a transformation of the predictor variables in a high, sometimes infinite dimensional space in which a linear boundary is sought. This transformation is calculated efficiently by the use of a kernel function. In the case of overlapping classes, the condition that points of one class have to lie on the one side and those of the other class on the other side of the boundary can be relaxed. In both cases, finding an optimally classifying boundary is formulated as a quadratic optimization problem and solved by standard techniques. Model selection is originally based on the Vapnik-Chernovenkis theory [8]. It is not known, however, whether this technique has an advantage over cross validation [2]. Prediction with SVMs is done by evaluating the obtained classifier on a new point, with class -1 or 1 as a direct result. This classifier generally depends only on few training examples; the support vectors. Since their invention in 1988 [8], Support vector machines have been successfully applied in different domains [9]. Despite their promising technical strength, they are sometimes criticized. Their results cannot be easily interpreted and one does not know the role of the predictor variables [9].

Comparison between Logistic Regression and Support Vector Machines

Logistic regression and SVMs are related: a SVM can be seen as an estimator of the class probabilities [2]. When a linear separation is possible, logistic regression will always find it [2]. Logistic regression implies linear decision boundaries, which can be seen by equaling the two class equations. It is not the scope of this report to elaborate this relationship.

We are mainly interested in the question whether these two approaches - a traditional, linear approach, and a newer, non linear approach - can give us different insight in hidden structures in the sector data.

Experiments

We conduct several classification experiments in order to gain insight into how planning differences behave in different sectors and on different days. For this, we create data in three stages: the randomly permuted sector data, sector data where the number of positive and negative instances is balanced and data, where only a subset of variables is selected. As described above, predictive accuracy of SVMs is critical to free parameters used. We combine thus a large number of SVM parameters systematically. These are the Kernel Functions: linear, Gaussian, polynomial, each in raw and in centered and normalized form, their associated parameters and the loss functions (one norm and two norm). In total, more than 800 SVM models are estimated per experiment. Parameters of the SVMs are estimated by cross validation and of the logistic regression by standard techniques [10]. Expected prediction errors for both are estimated by cross validation. For this, we split our data into 15 parts of equal size, train a model on 14 parts and calculate prediction error for the remaining one. To obtain an estimate of the prediction error, the average for 15 runs for each model is taken. We compare our results with a

Wilcoxon-Mann-Whitney test. We use the best 10 runs for the Logistic regression and for the SVM.

Two categories of classification experiments have been carried out:

Weekend/Weekday Experiment

Data from one sector is classified according to whether it is from a weekday or the weekend. Best classification results are obtained on the raw data, that is, unbalanced. Here, performance of logistic regression and SVM do not differ significantly. EPE are around 26 %. In more detail, SVM perform significantly worse on balanced data and on data with variable selection. Logistic regression performance does not differ between raw data and balanced data but is significantly better on raw data than on data with variable selection. As a baseline, we assigned class attributes arbitrarily with EPE ~ 50 % (tables 1,2,3).

Sector Name Experiment

In this experiment we are interested whether data from a sector S differs from that of the other sectors. Data is attributed to group 1 if belonging to sector S and to group 0 otherwise. The classification experiment is run for each of the sectors. Best SVM performs significantly better than best Logistic Regression but no more than 4 % (on a 10 % level). Average EPE for logistic regression is 18.4 % and 12.8 % for SVM. Kernel Statistics: The top ten performances in the raw data are achieved exclusively by Gauss and Gauss CN Kernels. For the balanced dataset, the situation is similar, but 3 Poly CN appear in the list, as well. No linear kernel appears in the list. Balancing the sample decreases quality significantly for logistic regression and has no impact on SVM. As above, our baseline resulted in EPE ~ 50 % (tables 4,5).

Conclusions and Future Work

We conducted 20 classification experiments on the two datasets 'Weekend/Weekday' and 'Sector Name'. Taking the best results of each method tells us that the expected prediction errors lie around 26 % for the weekend data and around 15 % for the sector name data. Thus, there is some characteristic in the data, which is identified by both methods.

Comparing logistic regression and SVM tells us that there is no significant difference in predictive accuracy on the weekend data and less than 4 % of better accuracy for the SVM than for the logistic regression in the sector name experiment. The best kernels are Gauss and Gauss CN in 98 % and Poly CN in the remaining. No linear kernel appears in the top ten performances of every experiment. We conclude that SVMs do not promise a major improvement on predictive accuracy, even if more parameter tuning experiments should follow.

For future work, we shall identify the reasons for the differences, such as traffic density or sector complexity. The implication of the existence of classifiers for the hypothesis of different underlying probability distributions has to be studied in further detail.

Acknowledgments

The authors like to thank Alexandre d'Aspremont, Laurent el Ghaoui and Devan Sohier for their helpful comments.

Bibliography

[1] Some Spatio Temporal Characteristics of the Planning Error in European ATFM. C. Gwiggner, P. Baptiste, V. Duong. Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems. ITSC 2004.

[2] The Elements of Statistical Learning. Data Mining, Inference, and Prediction, T. Hastie, R. Tibshirani, J. Friedman, Springer Series in Statistics, 2003.

[3] Probabilites, Analyse des Donnees et Statistique. G. Saporta. Editions Technip. Paris. 1990.

[4] Additive logistic regression: a statistical view of boosting. J. Friedman, T. Hastie, and R. Tibshirani. Ann. Statist. 28 (2000), no. 2, 337-407.

[5] The Origins of Logistic Regression. J.S. Cramer. Tinbergen Institute Discussion Papers 02-119/4. Tinbergen Institute. 2002.

[6] Why the logistic function ? A tutorial discussion on probabilities and neural networks. M. I. Jordan. MIT Computational Cognitive Science Report 9503, 1995.

[7] Generalized Linear Models, 2nd Edition. P. McCullagh, J. A. Nelder, Chapman \& Hall, Monographs on Statistics and Applied Probability 37. 1988.

[8] Support Vector and Kernel Methods. N. Christianini, J. S. Shawe-Taylor. In Intelligent Data Analysis. An Introduction, 2nd Edition. M. Berthold, D. J. Hand (Eds). Springer. 2003.

[9] Recent advances in predictive (machine) learning. J. Friedman, Nov. 2003. Technical Report. Stanford University.

[10] R: A language and environment for statistical computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2004.

Annex

The following tables contain the expected prediction errors in columns LReg and SVM and the result of the comparison in the column Comp. Here, the symbol ~ is used for no significant difference.

WE raw	***LReg***	***SVM***	***Comp***
UR1	27.6	22.3	SVM
UR2	25.9	22.3	SVM
UR3	27.1	27.1	~
UR4	25.9	26.4	~

Table 1 Results Weekend raw

WE balanced	***LReg***	***SVM***	***Comp***
UR1	24	25.4	~
UR2	36	20.1	SVM
UR3	27	48.8	LReg
UR4	28	29.7	~

Table 2 Results Weekend balanced

WE 6-19	***LReg***	***SVM***	***Comp***
UR1	32	33.6	LReg
UR2	40	51	LReg
UR3	36	31.1	SVM
UR4	20	31	LReg

Table 3 Results Weekend 6-19 h

Name raw	***LReg***	***SVM***	***Comp***
UR1	13.5	12.1	SVM
UR2	13.95	12.1	SVM
UR3	15.3	14.6	~
UR4	21.4	19.8	~

Table 4 Results Named raw

Name balanced	***LReg***	***SVM***	***Comp***
UR1	18.5	4.1	SVM
UR2	15.9	13.1	~
UR3	19.1	17.1	~
UR4	35.9	17.2	SVM

Table 5 Results Named balanced