The **BBVA Foundation** promotes and disseminates knowledge for the benefit of society. It centers its activity on the promotion of scientific research and advanced training, focusing especially on the analysis of emerging issues in five strategic areas: Environment, Biomedicine and Health, Economy and Society, Basic Sciences and Technology, and Arts and Humanities. The BBVA Foundation designs, develops and finances research projects in these areas; facilitates advanced, specialist training through grants, courses, seminars and workshops; organizes award schemes for researchers and professionals whose work has contributed to the advancement of knowledge; and communicates and disseminates new knowledge through publications, debates and lectures. The **BBVA Foundation** is guided in its work by the five values of scientific objectivity, innovation, independence, transparency and a commitment to excellence.

Fundación BBVA

Plaza de San Nicolás, 4
48005 Bilbao
Tel.: 94 487 52 52
Fax: 94 424 46 21

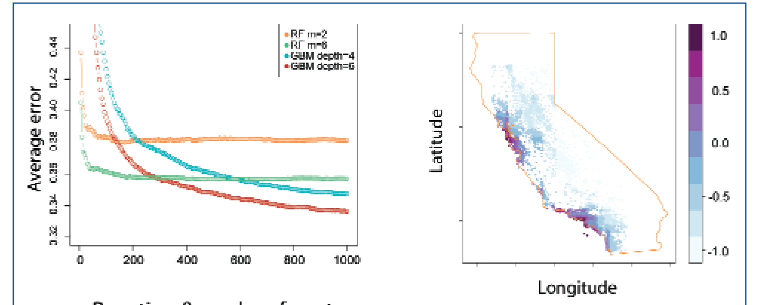
Paseo de Recoletos, 10
28001 Madrid
Tel.: 91 374 54 00
Fax: 91 374 85 22

informacion@fbbva.es
www.fbbva.es

Impreso en papel reciclado

Créditos fotográficos: Ingram Image Library • *Diseño:* LA FACTORÍA DE EDICIONES
D.L.: M-23762-2009

Fundación BBVA



Boosting & random forest performance

Partial dependence plot

2-3 July 2009 • 9:00-18:00
Fundación BBVA
Palacio del Marqués de Salamanca
Paseo de Recoletos, 10
28001 MADRID

Workshop

Statistical Learning

INSTRUCTORS:

Prof. Trevor Hastie
Professor of Statistics and Biostatistics
Stanford University, USA

Prof. Michael Greenacre
Professor of Statistics
Universitat Pompeu Fabra, Barcelona

Statistical Learning

PRESENTATION

The world we live in is swimming in data: data that are generated explicitly, as in microarray experiments, data that are gathered in the course of business, as in insurance claims and supermarket purchases, and data that arise naturally, as in WWW communications. Along with the oceans of data, there is a growing need for data analysts to develop methodologies for learning from these data.

Many modern learning problems can be cast in terms of prediction:

a) Predict what movies a customer might be interested in, given their ratings of past selections.
b) Predict the movement of an investment portfolio, based on market information and ticker data.
c) Predict whether an incoming email is spam, based on the content of the email, and the history of emails received by the client.
d) Predict whether an insurance claim is fraudulent, based on customer data.
e) Predict the prognosis of a cancer patient, based on the gene-expression profile of a sample of tumor tissue.
f) Predict the demographics of a web user, based on the demographics of the web sites visited.

Often the methods we use are classical and simple, because we have far more features than observations. In many modern applications where data are gathered automatically, we have very large samples, and so can afford to fit very rich models.

In this workshop, presented by Trevor Hastie, and assisted by Michael Greenacre in the afternoon practical sessions, we cover a spectrum of approaches, all of which are very current and are in daily use in industry and science. Computers will be provided for use during the practical sessions and participants are urged to bring their own data sets. All our applications are programmed in the R language, an attractive environment for data analysis and presentation.

Workshop participants should have a working knowledge of the R computing environment, which will be used in the course tutorial sessions each afternoon. This should include writing and editing programs, importing data, installing packages, and producing graphical output. There are many introductory texts in R as well as free online tutorials, for example: http://faculty.washington.edu/tlumley/Rcourse/R-fundamentals.pdf.

Much of the material in the course is extracted from a similar course designed and taught by Trevor Hastie and Rob Tibshirani. Course attendees will benefit by reading the book *Elements of Statistical Learning* (Hastie, Tibshirani & Friedman, Springer, second edition, 2009).

```
dim(spa
1] 4601   57
> library(randomForest)
> set.seed(303)
> rf1<-randomForest(x=spam.x[spam.train.inde
> rftest<-rf1$test
> spam.rferror<-rftest$err.rate[,"Test"]
> varImpPlot(rf1)
> library(gbm)
> gbm0<-gbm.fit(x=spam.x[spam.train.index
> gbm.pred0<-predict(gbm0,spam.x[-spam.
> gbm.gbmerror<-apply((spam.y[-spam.tra
m0)
Class   Mode
numeric
```

PROGRAM

July 2

9:00-10:30	Introduction to the learning problem, with an overview of the variety of methods. Contrast the two arenas leading to wide versus tall data
10:30-11:00	Coffee break
11:00-12:30	Linear regression methods, model selection, and a variety of popular approaches to regularization
12:30-13:00	Coffee break
13:00-14:30	Linear models for wide data, ridge regression, nearest shrunken centroids and supervised principal component analysis
14:30-16:00	Lunch time
16:00-18:00	Practical session in R: use 10-fold cross-validation to drive forward stepwise selection

July 3

9:00-10:30	Support-vector machines and kernel methods
10:30-11:00	Coffee break
11:00-12:30	Random forests and boosting
12:30-13:00	Coffee break
13:00-14:30	Continuation. Boosting and ensemble learning
14:30-16:00	Lunch time
16:00-18:00	Practical session in R: compare random forests and boosting on a large classification problem

INSTRUCTORS

Prof. Trevor Hastie

Trevor Hastie was born in South Africa in 1953. He received his university education from Rhodes University, South Africa (BS), University of Cape Town (MS), and Stanford University (Ph.D Statistics 1984). In March 1986 he joined the statistics and data analysis research group at AT&T Bell Laboratories. After nine years at Bell Labs, he returned to Stanford University in 1994 as Professor of Statistics and Biostatistics.

His main research contributions have been in the field of applied nonparametric regression and classification, and he has written two books in this area: *Generalized Additive Models* (with R. Tibshirani, Chapman and Hall, 1991), and *Elements of Statistical Learning* (with R. Tibshirani and J. Friedman, Springer, second edition, 2009). He has also made contributions in statistical computing, co-editing (with J. Chambers) a large software library on modelling tools in the S language ("Statistical Models in S", Wadsworth, 1992), which form the basis for much of the statistical modelling in R and S-plus. His current research focuses on applied problems in biology and genomics, medicine and industry, in particular data mining, prediction and classification problems.

Prof. Michael Greenacre

Michael Greenacre is Professor of Statistics at the Universitat Pompeu Fabra in Barcelona. His research interests are in the analysis of large data sets in the social and environmental sciences, having authored and co-edited six books and numerous journal articles on correspondence analysis and data visualization.

He has given courses on multivariate analysis to social and environmental scientists in 15 countries, and has participated as a statistician in various environmental research projects and the ARCTOS network for research in the Arctic.

PREVIOUS BBVA FOUNDATION WORKSHOPS ON STATISTICS:

Structural Equation Modelling with LISREL,
by Karl Jöreskog and Dag Sörbom,
November 2004

Correspondence Analysis in Practice,
by Michael Greenacre and Antonietta Curci,
March 2005

Multivariate Analysis for Environmental Biologists (first edition),
by Michael Greenacre and Raul Primicerio,
November 2005

Multivariate Analysis for Environmental Biologists (second edition),
by Michael Greenacre and Raul Primicerio,
May 2006

Multivariate Statistical Modelling of Ecological Data (third edition),
by Michael Greenacre and Raul Primicerio,
April 2007

Statistical Learning (first edition),
by Trevor Hastie and Michael Greenacre,
April 2008

Multivariate Statistical Modelling of Ecological Data (fourth edition),
by Michael Greenacre and Raul Primicerio,
September 2008

Course fees: companies, 150 euros / university staff and students, 100 euros
(includes course material, refreshments and lunches)
Registration: June 22 to 26, 2009
To register, please contact: registration@fbbva.es
Simultaneous interpretation will be provided