An Introduction to Statistical Modeling of Extreme Values

Stuart Coles

Springer-Verlag, London, 2001
208 pàgines

This is a really interesting, clear, updated and very well documented book. It summarizes an important part of the new research on the extreme value theory and is directly oriented towards real practical application. Most aspects of extreme modeling techniques are covered, including historical and contemporary techniques. A wide range of worked examples using genuine datasets illustrate the various modeling procedures. My personal opinion is clearly positive, I simply like this book.

First of all, I would like to remark the point of view of the author. The first thinks one finds in the book are genuine datasets. Chapter 1 shows the examples that will be deeply studied in the text. A practical point of view is always considered, complemented by the theoretical framework of extreme value models. All the computations are carried out using S-PLUS, and the corresponding datasets and functions are available on Internet Web page: http://www.stats.bris.ac.uk/ masgc/ismev/summary.html

Chapter 2 gives a nice summary of the most updated likelihood methods for inference, including profile likelihood for quantile estimations. This provides a coherent and global method really useful to consider models with parameters which depend on time and covariants. Maximum likelihood methods adjust automatically the complex models, allowing us to quantify the uncertainty of the model and giving us a way to check the goodness of fit. Moreover, Chapter 9, the last one, deals with a brief introduction to more advanced topics as Bayesian inference and Markov chain Monte Carlo methods.

The core of the book is chapter 3 and chapter 4, in which the classical extreme value theory is developed. They make the theory available for statisticians and non-statisticians alike thanks to its elementary treatment, with heuristic proves often replacing more detailed mathematical proves. The material includes the generalized extreme value distributions and the threshold models with a generalized Pareto distribution.

In general, the book is a good complement of a more classical text, because it considers more current statistical inference techniques for using these models in practice.

In chapters 5 and 6, series of dependent observations are studied. The first one develops methods for stationary series, and the other one for non-stationary series. In the first ca-

se, the same methods are used in independent series work. The same limit distributions arise as natural limits, but the degree of accuracy depends on the degree of dependence. Chapter 5 ends with some applications to Dow Jones financial series. In order to study non-stationary series Chapter 6 introduces covariants, order statistics and all kind of information related to the data. The Chapter includes the study of a nice dataset on the annual sea level in Venice.

Chapter 7 shows an elegant formulation of the extreme value behavior from the theory of point process. This is now the newest point of view of the theory. Chapter 8 focuses on the multivariate extremes. The same methods of single theory can be extended, but new problems arise. The dimensionality raises problems for validation and computation. Special attention is payed to the two-dimensional case.

Finally, I would like to say that the extreme value theory is closely related to the heavy tailed distributions theory and this one is now really popular in mathematical finance. In the last years, many new books on extreme values have appeared. I think this one is an essential reference both for students and researchers in statistics and finance, and the book will also appeal to practitioners looking for practical help to solve real problems.

More information about this book can be found in http://www.springer.de/.

Joan del Castillo
Universitat Autònoma de Barcelona

The Elements of Statistical Learning
Data Mining, Inference and Prediction

Trevor Hastie, Robert Tibshirani i Jerome Friedman

Springer-Verlag, 2002
533 pàgines

El libro esta dedicado al apasionante tema de la modelización estadística, en el sentido más general del término, desde la perspectiva de la predicción. Este es un problema de gran actualidad en el entorno de la minería de datos. Esto es, se utilizan los datos para aprender de ellos, sin suponer la existencia de un modelo teórico a confirmar, sino que se pretende utilizar la información disponible para construir modelos que permitan hacer predicciones sobre la(s) variable(s) de respuesta.

La primera característica del libro que sorprende es la alta calidad de la edición, la impresión en color de los gráficos y títulos es sencillamente espectacular. Pero también el contenido esta a la altura. La gran amplitud de los modelos presentados, muchos de ellos sucintamente, pero el nivel conseguido y la claridad en el enfoque, muestran el conocimiento y la experiencia de los autores, fruto de años de investigación y aplicación de los modelos descritos, siendo los apartados de mayor dificultad teórica señalados mediante una tarjeta amarilla (por suerte no aparecen tarjetas rojas). En este sentido y dado el número de modelos diversos que se presentan, el primer problema es ordenarlos mediante un encadenamiento lógico que relacione todos los modelos. Debemos decir que esto lo consiguen en gran parte, siendo este otro de los atractivos del libro, presentar de forma relacionada y comparativa modelos, en principio disconexos.

Otra característica del libro es que intenta unificar la nomenclatura estadística con la utilizada en la comunidad informática de «machine learning» de cara a utilizar un solo léxico para los mismos problemas y sus soluciones. En este sentido es muy útil la presentación en la mayoría de modelos de su correspondiente algoritmo, que facilita su comprensión y permite su implementación mediante una herramienta informática de alto nivel, como S-plus, R o Matlab y a su vez tiende puentes de entendimiento con la comunidad de «machine learning». Señalemos que la solución informática adoptada en el libro es el S-Plus, de la que el libro hace una exuberante demostración de posibilidades estadísticas y gráficas. Los métodos siempre vienen de la mano de su aplicación a problemas reales y actuales, lo cual no es óbice para utilizar datos simulados cuando interesa dilucidar el distinto comportamiento de los modelos sobre unos datos. Utiliza como «datasets» conjuntos de datos disponibles via web, con problemas actuales que

van desde el reconocimiento de la escritura manuscrita, el filtraje de mensajes spam, o la predicción de diferentes tipos de cáncer a partir de información sobre los genes, etc. Lo cual por otro lado, es muy útil desde el punto de vista docente para prácticas de laboratorio, siendo también útiles en este sentido los problemas de final de capítulo. También unifica los problemas de regresión y clasificación bajo un solo marco, puesto que se trata del mismo problema.

A continuación presentamos los «principales» modelos presentados para dar idea de la amplitud de libro, lista que no cubre todos los modelos presentados, pero si da una idea del contenido del libro.

El libro empieza con una presentación de los modelos de predicción más sencillos, los lineales, para hacer a continuación un salto al aproximador universal más flexible, de los k-vecinos más próximos. Este es uno de los mejores momentos del libro, señala las limitaciones de ambos extremos, la simplicidad de los primeros y el problema de la dimensionalidad en los segundos y justifica la necesidad de encontrar modelos situados entre ambos extremos, a lo cual dedica el resto del libro.

La primera generalización que presenta es permitir más flexibilidad en los modelos lineales mediante expansiones del espacio de características, ya sea por funciones polinómicas, por splines de regresión, alisados (smoothing splines), wavelets. Otra forma de superar la rigidez de los modelos lineales es realizando una regresión local en la vecindad de un punto, mediante regresión kernel y su generalización (y simplificación) al problema de clasificación (Naive bayes) y los «radial basis». A continuación los modelos aditivos generalizados y los árboles de decisión. Es interesante la presentación de los árboles de decisión como un modelo aditivo, el cual permitirá más adelante su generalización para el caso multivariante (MARS, Multivariate Adaptative Regression Splines).

El tema de la complejidad de los modelos se trata en el capítulo 7, a mayor complejidad mejor ajuste pero menor poder de generalización, este problema lo soluciona por regularización, esto es, penalizar el criterio a optimizar por la complejidad del modelo, en este sentido señalar la presentación de la «ridge regressión» como un problema de regularización y su generalización en los «smoothing splines». Ligado a la complejidad del modelo está la selección del modelo, al necesario equilibrio entre sesgo y variancia en las predicciones. El criterio a optimizar es en general la suma de cuadrados residuales penalizados, también la entropía en problemas de clasificación, y también la maximización de la verosimilitud y el método bayesiano. El error de predicción se mide mediante métodos heurísticos, de la muestra test, validación cruzada o bootstrap, pero también se explican los criterios AIC, BIC y la dimensión de Vapnik Chernovenkis para la complejidad del modelo.

También se trata los métodos de consenso, tales como el «bagging» (por promedio de las predicciones) como un método minimizar el error de predicción de los modelos, y

el de consenso mejorado, ponderando las observaciones en función de la malclasificación («Boosting»), como una forma de producir un clasificador fuerte a partir de un clasificador débil inicial, el cual se reduce a un modelo aditivo particular.

Finalmente se presentan los modelos de predicción basados en factores construidos a partir de los datos, Projection Pursuit Regression (si bien aquí estaría bien empezar por la regresión sobre componentes principales (PCR) y también la regresión PLS) y su generalización en redes neuronales. También se presentan las generalizaciones del análisis discriminante «Support Vector Machines», el cual construye hiperplanos no lineales separadores y los «flexible discriminants».

Los últimos capítulos se dedican a una serie de técnicas de minería de datos no directamente relacionadas con el problema de predicción, tales como el aprendizaje no supervisado donde se ha incluido las reglas de asociación (Market Basket Analysis) y por supuesto los métodos de «clustering», en particular el «k-means», mapas de Kohonen, «vector quantization» y métodos aglomerativos. También las Componentes Principales no lineales, «Independent Component Análisis», método relacionado con Factor Analysis y Multidimensional Scaling.

Ciertamente es imposible describir en profundidad todos y cada uno de los modelos, pero a menudo es suficiente para empezar en su praxis, si bien un conocimiento previo facilita en gran medida una mejor comprensión.

En resumen, se trata de una obra de referencia imprescindible en la biblioteca de cualquier investigador o aplicador de las técnicas de mineria de datos, y que a su vez puede utilizarse como guía docente de varios cursos cuatrimestrales de minería de datos.

Información complementaria sobre este libro se puede encontrar en http://www.springer.de/.

Tomàs Aluja
Universitat Politècnica de Catalunya