Para los que estéis aún preprocesando los datos y en toma de contacto, animaros a ir acabándolo y, sobre todo, tened en cuenta que la curva de aprendizaje al principio es lenta, pero una vez se va cogiendo el tema, se acaba haciendo exponencial, ya lo veréis. Lo bueno (y lo malo) de R es que tiene miles de librerías y ofrece un sinfin de posibilidades en el trabajo con datos, y los paquetes están bastante testados, pero digo lo malo porque hoy por hoy la Ciencia de los Datos está creciendo a pasos agigantados, y cada dos por tres sale un nuevo algoritmo......

Se ha adjuntado en el apartado documentos el resto del material docente, correspondiente al siguiente trabajo que se hace con R, la modelización. Las funciones que hayáis hecho personalizando la presentación de vuestros resultados en gráficos y tablas, podéis aprovecharlas aquí. Tened en cuenta que la presentación en esta ciencia es muy importante. De hecho, a los profesionales dedicados a esto se les llama Data Artist (Artista de los Datos) y son cada vez más demandados.

La modelización está orientada, al igual que en años anteriores (pero con Matlab), a introducirnos en los métodos de Monte-Carlo (llamados así por el casino más importante de la época), la Regresión no paramétrica y la Estadística bayesiana (como derivado importante de los algoritmos de Metrópolis-Hastings), donde podemos incorporar información a priori al modelo en base al conocimiento adquirido de situaciones semejantes previas.

La simulación estadística, además de permitirnos trabajar "a la bayesiana", nos permite validar los modelos sin depender de hipótesis, bien a través del bootstrap o de la validación cruzada (quizás la más utilizada actualmente), y es por esto que tenéis un apartado inicial en el documento de machine learning abordando el tema de la validación de los modelos, la curva ROC por su importancia en la clasificación, y la forma en que se comparan los desempeños de distintos algoritmos.

Asimismo, también el bootstrap permite elaborar modelos que ofrecen una gran precisión, como el Random Forest (en mi opinión, quizás el mejor modelo del Machine Learning). Al ser un ensemble que utiliza árboles CART como algoritmo de base, se explican también las características de dicho algoritmo, del Bagging y el Extreme Randomized Tree. La Estadística Bayesiana tiene su importante aportación también aquí, se muestra como ejemplo el algoritmo Naive-Bayes.

Por último, al margen de la gran utilidad e importancia que ha adquirido la Regresión no parámetrica por sí misma en el mundo de la Economía, señalar que combinada con los Modelos Lineales Generalizados, son la base de los denominados Modelos Aditivos Generalizados, el modelo más flexible y potente de la Estadística clásica, y que compite perfectamente con cualquiera de los algoritmos de Machine Learning. Por mi experiencia, en muestras medias y pequeñas mejor GAM, en muestras grandes Random Forest.

Les he elaborado un capítulo donde se repasa el modelo GLM, excluyendo el diseño experimental (comentario para los que estén también en Métodos Estadísticos). Aun así, el modelo GLM es mucho más amplio. En este documento se refleja lo fundamental, y que entiendo (puedo estar errado) que alguien que cursa Máster lo debe de conocer ya, está principalmente para que veáis las funciones a usar. Dicho esto, cualquier duda me dicen. Si alguien aquí, quiere conocer el diseño experimental, como vean...

Espero que puedan sacar utilidad de estos dos nuevos modelos que se han añadido. Los temas se abordan como sigue:

• Primer tema: introducción y GLM.

• Segundo tema: Métodos de Monte Carlo y GLM Bayesiano.

• Tercer tema: Regresión no paramétrica y GAM.

• Cuarto tema: Introducción al Machine Learning

Además de la teoría, donde se muestra el código con ejemplos, les he incluido complementariamente:

• el pdf del paquete MCMCPack, que es el utilizado en inferencia bayesiana.

• un ejemplo de artículo que estima macromagnitudes en áreas pequeñas con regresión bayesiana.

• el pdf del paquete caret, que contiene funciones para configurar la validación, el preprocesado (escalar variables, smote, muestra, selección variables,...), un gran número de algoritmos para regresión y clasificación, así como otras funciones para comparar estadísticamente el desempeños de los diversos modelos que podamos estimar sobre una misma base de datos.

• el libro de Max Kuhn, el autor de la librería caret, donde tenemos ejemplos y la información relativa a los algoritmos utilizados y las librerías que utiliza. Lo que hace caret es homogeneizar los resultados de las librerías que hacen los modelos. Si alguno tiene experiencia programando, también podemos incorporar a caret nuestros propios modelos

Les dejo aquí los ejercicios propuestos.

Reciban atento saludo