Técnicas de Reducción y Visualización de Datos

Prefinal – 24/09/2024

El examen se entrega vía campus ***hasta el miércoles 25/09 a las 13 hs*** *en un archivo* `.pdf`*, y acompañado del archivo* `.Rmd` *correspondiente. La resolución debe incluir no solo el código con las soluciones y resoluciones, sino también los comentarios y explicaciones que se piden y/o se consideren pertinentes. Estarán habilitados para entregar el examen vía campus* ***únicamente*** *aquellos que hayan iniciado el examen de forma presencial en el laboratorio el día 24/09 a las 12 hs.* ***Importante: cada uno de estos archivos debe contener el apellido del autor en el nombre, por ejemplo:*** `Prefinal-Bianco.pdf`***.***

El conjunto de datos `stocks.txt` contiene de los rendimientos porcentuales de un índice bursátil a lo largo de 1250 días, desde principios de 2001 hasta fines de 2005 y es una versión alterada de un *dataset* famoso.

Para cada fecha, se han registrado los rendimientos porcentuales de cada uno de los cinco días de negociación anteriores, disponibles en `lag1` a `lag5`. También se ha registrado la cantidad de acciones negociadas el día anterior, en miles de millones, (disponible en `volumen`), y una variable cualitativa que indica si el mercado estaba subiendo (`Up`) o bajando (`Down`) en esta fecha (diponible en `direc`).

A lo largo de este trabajo se buscará predecir la dirección del mercado utilizando las otras características disponibles: `lag1` a `lag5` y `volumen`.

1. (10 puntos) Examiná algunos resúmenes numéricos y gráficos de los datos. Explicá por qué usas esos resúmenes y qué observás a partir de ellos.

2. (85 puntos) En los siguientes apartados, deberás probar diferentes modelos de clasificación: regresión logística (RL), discriminante lineal (LDA) y cuadrático (QDA), y Bayes Naive (BN). Para cada modelo que uses, hacé explícitos los supuestos y su estructura.

   a) (5 puntos) Separá los datos en una muestra de entrenamiento (`train`, datos previos a 2005) y en una muestra de testeo (`test`, datos de 2005). Explicá por qué es razonable hacer esta división con estos datos y no una en la que se sorteen los índices de las observaciones a cada grupo.

   b) (48 puntos) Entrená cada uno de los 4 modelos con los datos de `train` y evaluá su precisión (*accuracy*) con los de `test`. Para el caso de la regresión logística, evaluá si es conveniente incorporar algún tipo de regularización. En caso de considerarlo así, usá el comando `cv.glmnet()`. Para la implementación de Bayes Naive podés usar el comando `naiveBayes()` de la librería `e1071`, cuya sintaxis es análoga a la de `lm()` y también interactúa con otras funciones útiles como `predict()`.

   c) (12 puntos) Completá la siguiente tabla con las métricas pedidas y evaluadas en el conjunto de testeo para cada modelo.

	Accuracy	Sensibilidad	Especificidad
RL			
LDA			
QDA			
BN			

¿Qué modelo sugerirías como el más adecuado para lo que se busca? ¿Por qué?

d) (20 puntos) Evaluá la curva ROC para cada modelo, graficala e indicá el AUC de cada uno en el conjunto de testeo como una medida global de la calidad del ajuste. Para ello, explorá los comandos `prediction()` y `performance()` de la librería `ROCR`. ¿Qué modelo sugerirías como el más adecuado para lo que se busca? ¿Por qué? ¿Cambió tu respuesta respecto del apartado anterior?

3. (5 puntos) Indicá el modelo (o modelos, si considerás que hay más de uno) que recomendarías para tratar este problema de predicción y explicá por qué pensás que logra mejorar la predicción por sobre los otros.