Ricardo Aler Mur

PREDICCIÓN/REGRESIÓN EN R

Puntuación: 4 puntos

PREDICCIÓN DE PRODUCCIÓN DE ENERGÍA EÓLICA

1. OBJETIVO

Se trata de resolver problemas de predicción de energía eólica a partir de predicciones meteorológicas y series temporales de predicción proporcionada en la competición KAGGLE 2012. Para entender el problema, aparte de la explicación dada en clase, es conveniente mirar:

- La propia descripción del problema en la competición:
  - https://www.kaggle.com/c/GEF2012-wind-forecasting
- Las transparencias de introducción a la práctica

La práctica tiene 4 partes y cada una de ellas vale un punto

1) **Primera Parte** – exploración visual de los datos y determinación de atributos relevantes con modelos lineales

   a) Separar 2/3 de los bloques de predicción de 48h para **entrenamiento** y el otro 1/3 para **validación**, tal y como se hace en el ejercicio 2 de las transparencias. Usad esta descomposición hasta el fin de la práctica, salvo que se diga lo contrario.
   b) Ahora, coged los datos de entrenamiento y seleccionad aquellos relativos al horizonte temporal de una hora (hors==1) y ved mediante plots las relaciones entre *u*, *v*, *ws* y *wd* y la producción eléctrica *wp*. ¿Son las relaciones entre las entradas y la salida *wp* lineales o no lineales? ¿Hay mucho ruido en la salida?
   c) Responder de manera gráfica: ¿todas las velocidades de viento se producen con la misma frecuencia? ¿todas las direcciones de viento se producen con la misma frecuencia?

d) Con los datos de entrenamiento de *c)*, construid cuatro modelos lineales con lm: *lm(wp ~ u)*, *lm(wp ~ v), lm(wp ~ ws), lm(wp ~ wd)* y calculad el error cuadrático medio con los datos de validación. ¿Cuál de los atributos parece ser más relevante?

e) ¿Hay alguna mejora (con un modelo lineal) si usamos *ws + wd*? ¿Y si usamos *ws + wd* mas un tercer atributo? ¿Y si usamos los cuatro atributos? Desde el punto de vista de los modelos lineales, ¿cuántos atributos merecería la pena utilizar?

2) <u>Segunda parte</u> - <u>Determinación del mejor tipo de modelo</u>

a) Con los datos de 1.c) entrenad *gbm y svm* y validadlos con los datos de validación con sus parámetros por omisión. ¿Se consigue mejorar a los modelos lineales?

b) Probad también *random forest* con parámetros por omisión y determinad las variables más relevantes que saca *random forest* con la función *importance*. ¿Coincide con lo que salía con los modelos lineales construidos con los atributos individuales?. **Nota importante**: random forest es un algoritmo lento, con lo que es conveniente ejecutarlo en otro ordendor mientras se sigue con la práctica en el ordenador principal.

c) Ajustad los parámetros *n.trees*, *interaction.depth* y *shrinkage* de *gbm*. Validad siempre con los datos de validación. ¿Es posible superar lo que salía con los modelos lineales?

d) Probad a hacer lo mismo con la mejor combinación de atributos encontrada en la primera parte (se trataría de construir el modelo con 1, 2, o 3 atributos, lo que mejor funcionara en la primera parte. ¿Es mejor que lo que salía en los resultados correspondientes de la primera parte?

3) <u>Tercera parte</u> – <u>Añadir atributos que representen los 1, 2, o 3 instantes previos de la serie temporal</u>

a) Construid un modelo lineal y un modelo con *gbm* que use los datos que se venían usando hasta ahora (los de 1c) con *hors*==1) pero añadiendo uno, dos o más instantes de la serie temporal previa a la hora de predicción de *wp*. ¿Añadir estos valores mejora los resultados obtenidos hasta ahora? ¿Cuántos valores habría que añadir?

b) ¿Ocurre lo mismo si hacemos la predicción a 24 horas (hors==24) o a 48 horas (hors==48)?

4) <u>Cuarta parte</u> – <u>Apartado libre. Basándose en el análisis anterior, intentad mejorar los resultados o probad algo diferente.</u>