Localización subcelular de proteínas

Machine Learning. PEC 3. 2019

Introducción

En esta PEC se va a realizar un informe basado en el "Yeast Data Set" de la *UCI Machine Learning Repository* en el enlace https://archive.ics.uci.edu/ml/datasets/Yeast. Los datos estan disponibles en el fichero "yeast.data".

Este conjunto de datos corresponde al análisis de varias características útiles para determinar la localización subcelular de proteínas. En concreto, la descripción de los atributos es:

Sequence Name: Accession number for the SWISS-PROT database

mcg: McGeoch's method for signal sequence recognition.

gvh: von Heijne's method for signal sequence recognition.

alm: Score of the ALOM membrane spanning region prediction program.

mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.

erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.

pox: Peroxisomal targeting signal in the C-terminus.

vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.

nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

Class Distribution. The class is the localization site.

- CYT (cytosolic or cytoskeletal) 463
- NUC (nuclear) 429
- MIT (mitochondrial) 244
- ME3 (membrane protein, no N-terminal signal) 163
- ME2 (membrane protein, uncleaved signal) 51
- ME1 (membrane protein, cleaved signal) 44
- EXC (extracellular) 37
- VAC (vacuolar) 30
- POX (peroxisomal) 20
- ERL (endoplasmic reticulum lumen) 5

Objetivo:

En esta PEC se analizan esto datos mediante la **implementación** de los diferentes **algoritmos estudiados**: *k-Nearest Neighbour, Naive Bayes, Artificial Neural Network, Support Vector Machine, Arbol de Decisión y Random Forest* para **predecir** la localización de proteinas.

Puntos importantes:

1. Solamente se considerarán observaciones cuya localización sea de las clases: CYT, ME1, ME2, ME3, MIT i NUC. Además, las observaciones de las clases ME1, ME2 i ME3 se considerarán pertenecientes a una misma clase, etiquetada como MEM.
2. Realizar una exploración de los datos que incluya una estadística descriptiva básica de las variables mediante tablas y gráficos.
3. En cada algoritmo hay que realizar las siguientes tres etapas: 1) Transformacion de los datos (en caso necesario) 2) Entrenar el modelo 3) Predicción y Evaluación del algoritmo. En la fase 3) "tunear" diferentes valores de los hiperparámetros del algoritmo para posteriormente evaluar su rendimiento.
4. Se debe aplicar la misma selección de datos training y test en todos los algoritmos. Utilizando la semilla aleatoria `12345`, para separar los datos en dos partes, una parte para training (67%) y otra parte para test (33%). Si se prefiere, se puede escoger otro tipo de partición de los datos para hacer la selección de training y test como por ejemplo k-fold crossvalidation, boostrap, random splitting, etc. Lo que es importante es mantener la misma selección para todos los algoritmos.
5. En todos los casos se evalua la calidad del algoritmo con la información obtenida de la función `confusionMatrix()` del paquete `caret`.
6. Para la ejecución especifica de cada algoritmo se puede usar la función de cada algoritmo como se presenta en el libro de referencia o usar el paquete `caret` con los diferentes modelos de los algoritmos. O incluso, hacer una versión mixta.
7. Comentario sobre el informe dinámico. Una opción interesante del knitr es poner `cache=TRUE`. Por ejemplo:

   ```
   knitr::opts_chunk$set(echo = FALSE, comment = NULL, cache = TRUE)
   ```

   Con esta opción al ejecutar el informe dinámico crea unas carpetas donde se guardan los resultados de los procesos. Cuando se vuelve a ejecutar de nuevo el informe dinámico solo ejecuta codigo R donde se ha producido cambios, en el resto lee la información previamente descargada. Es una opción muy adecuada cuando la ejecución es muy costosa computacionalmente.

Informe de la PEC

Las soluciones se presentarán mediante un informe dinámico R markdown con la siguiente estructura:

1. Titulo: igual que el de la PEC, autor, fecha de creación e indice de apartados de la PEC.
2. Sección de lectura, exploración de los datos y obtención de los muestras de train y test. Recordar que un primer paso es, si hace falta, transformar las variables leidas al tipo de objeto R adecuado al tipo de variable. La exploración de los datos se aplica a todas las variables leidas. (*Puntuación*: 10%)
3. Sección de aplicación de cada algoritmo para la clasificación. Está formado por subsecciones que corresponden a cada algoritmo: k-Nearest Neighbour, Naive Bayes, Artificial Neural Network, Support Vector Machine, Arbol de Decisión y Random Forest manteniendo este orden. (*Puntuación*: 60%)

   En cada algoritmo hay que realizar las tres etapas mencionadas anteriormente.
4. Sección de conclusión y discusión sobre el rendimiento, interpretabilidad, ... de los algoritmos para el problema tratado. Proponer que modelo o modelos son los mejores. (*Puntuación*: 20%)

Un característica que se valorará es hasta que punto es el informe "dinámico". En el sentido de adaptarse el informe a cambios en los datos, es decir, si el fichero de datos cambia el informe se adapta a los nuevos resultados. (*Puntuación*: 10%)

Se subiran al registro de entregas un **zip** con los siguientes ficheros:

1. Fichero ejecutable (.Rmd) que incluya un texto explicativo que detalle los pasos implementados en el script y el código de los análisis. No olvidar de incluir todos los ficheros complementarios que hagan falta para la correcta ejecución: *ficheros de datos, fichero de bibliografia, imagenes, ...*

   NOTA: Para facilitar la ejecución, no usar un ruta fija para la lectura del fichero, asociarlo al area de trabajo donde este el fichero .Rmd.

2. Informe (pdf) resultado de la ejecución del fichero Rmd anterior.

Antes de enviar el zip, se recomienda **verificar la reproducibilidad del fichero .Rmd** para obtener el informe en formato pdf sin ninguna dificultad.