Machine Learning (PEC3)

Diciembre 2020

Secuencias promotoras en E. Coli

Los promotores son secuencias de ADN que afectan la frecuencia y ubicación del inicio de la transcripción a través de la interacción con la ARN polimerasa.

Este estudio se basa en los ficheros obtenidos de:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Para más información, se puede recurrir a la siguiente referencia acerca del estudio de promotores en E. Coli: Harley, C. and Reynolds, R. 1987. "Analysis of E. Coli Promoter Sequences." Nucleic Acids Research, 15:2343-2361

Los atributos del fichero de datos son:

1. Un símbolo de {+/-}, indicando la clase ("+" = promotor).
2. El nombre de la secuencia promotora. Las instancias que corresponden a no promotores se denominan por la posición genómica.
3. Las restantes 57 posiciones corresponden a la secuencia.

La manera elegida para representar los datos es un paso crucial en los algoritmos de clasificación. En el caso que nos ocupa, análisis basados en secuencias, se usará la transformación denominada **one-hot encoding**.

El one-hot encoding representa cada nucleótido por un vector de 4 componentes, con 3 de ellas a 0 y una a 1 indicando el nucleótido. Pongamos por ejemplo, el nucleótido T se representa por (1,0,0,0), el nucleótido C por (0,1,0,0), el nucleótido G por (0,0,1,0) y el nucleótido A por (0,0,0,1).

Por tanto, para una secuencia de 57 nucleótidos, como en nuestro caso, se obtendrá un vector de 4*57=228 componentes, resultado de concatenar los vectores para cada uno de los 57 nucleótidos.

Una vez realizada la transformación, one-hot encoding el objetivo se trata de implementar distintos algoritmos vistos en el curso para predecir si la secuencia es un promotor o no, y comparar sus rendimientos.

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** si una secuencia de ADN es promotor o no.

Puntos importantes:

1. Implementar una función para realizar una transformación one-hot encoding de las secuencias del fichero de datos `promoters.txt`. En caso de no lograr la implementaación de dicha transformación, se puede utilizar el fichero `promoters_onehot.txt` con las secuencias codificados segun un one-hot para completar la actividad.

2. En cada algoritmo hay que realizar las siguientes etapas: 1) Entrenar el modelo 2) Predicción y Evaluación del algoritmo. Será necesario "tunear" diferentes valores de los hiperparámetros del algoritmo para posteriormente evaluar su rendimiento.

3. Se debe aplicar la misma selección de datos training y test en todos los algoritmos. Utilizando la semilla aleatoria `123`, para separar los datos en dos partes, una parte para training (67%) y otra parte para test (33%). Opcionalmente, se puede escoger otro tipo de partición del conjunto de training para hacer la validación como por ejemplo k-fold crossvalidation, boostrap, random splitting, etc. Lo que es importante es mantener la misma selección para todos los algoritmos.

4. En todos los casos se evalua la calidad del algoritmo con la información obtenida de la función `confusionMatrix()` del paquete `caret`.

5. 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.

6. 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. Título: igual que el de la PEC, autor, fecha de creación e indice de apartados de la PEC.

2. Sección de lectura y de transformación de los datos. 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. (*Puntuación*: 10%)

3. Sección de aplicación de cada algoritmo para la clasificación. Está formado por subsecciones que corresponden a cada algoritmo y en este orden: k-Nearest Neighbour (se explorarán los valores para el número de vecinos $k = 1, 3, 5, 7$), Naive Bayes (se explorará la opción de activar o no 'laplace'), Artificial Neural Network (se explorarán el número de nodos de la capa oculta $n = 4, 5$), Support Vector Machine (se explorarán la funciones kernel lineal y rbf), Árbol de Clasificación (se explorará la opción de activar o no 'boosting') y Random Forest (se explorarán la opción de número de árboles $n = 50, 100$). (*Puntuación*: 60%)

   En cada algoritmo hay que realizar las 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.