Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks

Introducción

En esta PEC vamos a realizar un informe que analiza un caso basado en los datos del artículo:

Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Khan et al. Nature Medicine, 2001, 6, 673-679

Los datos se pueden obtener directamente de la revista Nature Medicine.

En dicho artículo se investiga la predición del diagnóstico de un tipo de cancer, "small, round blue cell tumors (SRBCTs)" en la infancia usando información del perfil de expresión génica obtenica mediante técnicas de microarrays.

Se estudian 4 tipos de canceres:

The small, round blue cell tumors (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non- Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), are so named because of their similar appearance on routine histology1. However, accurate diagnosis of SRBCTs is essential because the treatment options, responses to therapy and prognoses vary widely depending on the diagnosis. As their name implies, these cancers are difficult to distinguish by light microscopy, and currently no single test can precisely distinguish these cancers.

El proceso de clasificación (ver Fig. 1) es mucho más elaborado que los presentados en las unidades. No reproduciremos este esquema, aunque es importante que se entienda. Se basa en una red neuronal artificial usando 3-fold crossvalidación y repitiendo el proceso 1250 veces mediante particiones aleatorias (proceso similar al bootstrap) para poder estudiar la robustez del modelo. Observar que el número de variables es muy grande (2308) así que se ha optado por realizar un análisis de componentes principales para reducir la dimensión de las variables iniciales y usar solo las 10 primeras en el algoritmo.

El análisis de componentes principales (PCA, en ingles) es una técnica básica y muy utilizada en análisis multivariante para reducir el número de variables creando nuevas variables como combinación lineal de las originales buscando máximizar la varianza explicada. Como no sé si sabeis realizar en R un PCA he optado por crear un fichero con el resultado del PCA llamado "pcaComponents.csv".

En esta PEC se usará los datos del artículo para implementar el algoritmo de red neuronal artificial y "support vector machine" (SVM) para predecir los cuatro tipos de canceres.

Enunciado

1. Escribir en el informe dos secciones con los titulos: "Algoritmo Red Neuronal Artificial" y "Algoritmo Support Vector Machine" en el que se haga una breve explicación de su funcionamiento y sus caracteristicas y se presente una tabla de sus fortaleza y debilidades para cada algoritmo.
2. Lectura del artículo (especialmente las secciones de Introducción y Métodos)
3. Desarrollar un script en R que implemente un clasificador de red neuronal artificial. El script debe:
   (a) Leer los valores de las componentes principales `pcaComponents.csv` y la clase de tumor `clase.csv` donde los valores 1, 2, 3 y 4 representan "EWS","BL", "NB" y "RMS" respectivamente. Solo usar las 10 primeras componentes como variables explicativas del modelo. El que sepa realizar un PCA, lo puede hacer a partir de los datos originales `data.csv` centrados y escalados (media 0 y desviación tipica 1)

(b) Normalizar las variables

(c) Utilizando la semilla aleatoria `12345`, separar los datos en dos partes, una parte para training (67%) y una parte para test (33%).

(d) Utilizar un y tres nodos para crear el modelo de red neuronal artificial basado en el training para predecir los cuatro tipos de canceres en los datos del test.

(e) Comentar los resultados de la clasificación en función de los valores generales de la clasificación como "accuracy" y otros. Comparar los resultados de clasificación obtenidos para los diferentes valores de nodos usados.

(f) Usar el paquete caret modelo nnet para realizar el modelo de tres nodos con 3-fold crossvalidation. Comentar los resultados

4. Desarrollar un script en R que implemente un clasificador de SVM. El script debe:

   (a) Leer los valores de expresión genica de `data.csv` y la clase de tumor `clase.csv` donde los valores 1, 2, 3 y 4 representan "EWS","BL", "NB" y "RMS" respectivamente.

   (b) Utilizando la semilla aleatoria `12345`, separar los datos en dos partes, una parte para training (67%) y una parte para test (33%).

   (c) Utilizar la función lineal y la RBF para crear el modelo de SVM basado en el training para predecir los cuatro tipos de canceres en los datos del test.

   (d) Comentar los resultados de la clasificación en función de los valores generales de la clasificación como "accuracy" y otros. Comparar los resultados de clasificación obtenidos para los diferentes funciones usadas.

   (e) Usar el paquete caret modelo svmLinear para realizar el modelo de SVM con 3-fold crossvalidation. Comentar los resultados

5. Comentar todos los resultados obtenidos y escoger que modelo puede ser el mejor.

Informe de la PEC

Las soluciones se presentarán mediante un informe dinámico R markdown con la estructura habitual de los ejercicios no evaluables realizados hasta ahora. En primer lugar, el informe tendra un titulo (igual que el de la PEC), el autor, la fecha de creación y el indice de apartados de la PEC. En segundo lugar, se crea una sección con el título “Algoritmo Red Neuronal Artificial” donde se haga una breve explicación de su funcionamiento y sus caracterisiticas. Además, se presenta la tabla de sus fortaleza y debilidades. En tercer lugar, se crea la sección “Algoritmo Support Vector Machine” similar a la anterior sección. En cuarto lugar se realizan los diferentes apartados de la PEC pero con la estructura de Step1 hasta Step5 para cada tipo de algoritmo. Al final se crea una sección “Discusión final” para comentar todos los resultados obtenidos y escoger el mejor modelo.

Un caracteristica que se valorará es hasta que punto es el informe “dinámico”. En el sentido de adaptarse el informe a cambios en los datos.

Se entregaran dos ficheros:

1. Fichero ejecutable (.Rmd) que incluya un texto explicativo que detalle los pasos implementados en el script y el código de los análisis.
2. Informe (pdf) resultado de la ejecución del fichero Rmd anterior.

Puntuacions de los apartados

Apartado 1 (5%), Apartado 3 (40%), Apartado 4 (40%), Apartado 5 (5%), Calidad del informe dinámico (10%).