Gene expression patterns of phenotypes subclasses using gene expression profiling and artificial neural networks

Introducción

En esta PEC vamos a realizar un informe que analiza un experimento relacionado con la clasificación de 4 tipos de fenotipos:

1: FNT1 2: FNT2 3: FNT2 4: FNT4

usando información del perfil de expresión génica obtenida mediante técnicas de microarrays.

El objetivo es implementar una red neuronal artificial y un "support vector machine" (SVM) para predecir los cuatro tipos de fenotipos.

Observar que el número de variables es muy grande así que se ha optado por realizar un análisis de componentes principales (PCA) para reducir la dimensión de las variables iniciales y usar solo las 8 primeras componentes como input de la red neuronal. En el caso de SVM se usarán las variables originales (no las variables PCA).

El análisis de componentes principales (PCA, en ingles) es una técnica básica y común en análisis multivariante para reducir el número de variables. Se basa en crear nuevas variables, denominadas componentes principales, como combinación lineal de las originales buscando máximizar la varianza explicada. Como no sé si sabeis realizar PCA en R, tambíen se dispone del fichero "pcaComponents6.csv" resultado del PCA donde las observaciones son representadas con las componentes principales.

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 características. Además, se presente una tabla de sus fortaleza y debilidades para cada algoritmo.
2. Desarrollar un código en R que implemente un clasificador de red neuronal artificial. El código en R debe:
   (a) Leer los valores de las componentes principales `pcaComponents6.csv` y la clase de fenotipos `clase6.csv` donde los valores 1, 2, 3 y 4 representan "FNT1","FNT2", "FNT3" y "FNT4", respectivamente. Solo usar las 8 primeras componentes como variables explicativas del modelo. El que sepa realizar un PCA, lo puede hacer a partir de los datos originales `data6.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) Antes de ejecutar cada uno de los modelos de clasificacion que se piden a continuación, poner como semilla generadora el valor `1234567`.
   (e) Crear dos modelos de red neuronal artificial de una sola capa oculta: un nodo y tres nodos. Aplicar los datos de training para crear los modelos y posteriormente, predecir los cuatro tipos de fenotipos en los datos del test.

(f) 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.

(g) Usar el paquete caret modelo nnet para implementar el modelo de tres nodos en la capa oculta, usando 3-fold crossvalidation. Comentar los resultados.

3. Desarrollar un código en R que implemente un clasificador de SVM. El código en R debe:

   (a) Leer los valores de expresión génica de `data6.csv` y la clase de fenotipos `clase6.csv` donde los valores 1, 2, 3 y 4 representan "FNT1","FNT2", "FNT3" y "FNT4", 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) Antes de ejecutar cada uno de los modelos de clasificación que se piden a continuación, poner como semilla generadora el valor `1234567`.

   (d) Utilizar la función lineal y la RBF para crear el modelo de SVM basado en el training para predecir los cuatro tipos de fenotipos 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 funciones usadas.

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

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

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.

Puntuacions de los apartados

Apartado 1 (5%), Apartado 2 (40%), Apartado 3 (40%), Apartado 4 (5%), Calidad del informe dinámico (10%).