Inteligencia Artificial

Seleccion inteligente de variables para la creacion de un modelo predictivo

Sergio Anguita Lorenzo

Aitor Brazaola Vicario

Proyecto

En esta practica se ha realizado la implementacion de lo algoritmos **naive bayes** y **knn**. Ambos se han realizado en R y se adjuntan en este entregable.

Los algorimos desarrollados son los dos algoritmos suspervisados que parten de un dataset de entrenamiento para generar el modelo probabilistico.

En cuanto al desarrollo de la practica en si, se explica su funcionamiento.

1. Se lee el fichero de los datos de entrenamiento. (vehicle-tra.csv)
2. Se lee el fichero de los datos de prueba. (vehicle-tst.csv)
3. Cada fichero, se separa en dos dataset. Uno contiene todos los atributos especificados y el otro dataset contiene las clases a las que pertenece dicho vehiculo. De esta forma cuando se entrene el clasificador con los atributos, estará asociado a una salida o clase.

Como la practica pide que los datos necesitan procesarse y normalizarse, se ha creado la funcion *MinMaxNormalizer()* que realiza dicho proceso dado un dataset. La normalizacion que se utiliza es la min-max, la cual deja todos los valores del dataset original en el rango [0, 1]. Esa es la ventaja que tiene. La desventaja es que no es capaz de eliminar outliers. Como knn funciona mejor con los datos normalizados, espero notar alguna mejora en la precision del modelo.

4. Una vez teniendo ambos datasets (normalizado y no normalizado), se ejecutan los algoritmos naive bayes y knn sobre ambos datasets.

Cada algoritmo, se configura diferente en cada iteracion quedando de la siguiente forma.

- Algoritmo knn: k iteraciones con $k = \{1, 3, 10, sqrt(items\_in\_dataset)\}$
- Algoritmo naive bayes: n iteraciones, siendo n el nivel de smooth en laplace con $n = \{0, 1, 2, 3, 4, 5\}$

Lo que se intenta hacer con laplace es eliminar aquellos elementos que tengan probabilidad $P(x) = 0$

Conclusión

Es normal que cuando se ejecuta el algoritmo KNN con los datos normalizados con min-max se obtenga una precision en el modelo predictivo del 70%, ya que el hecho de que esten normalizados ayuda a ello. Aun asi, con un exito del 70%, en una situacio real, se considera que el modelo probabilistico tiene una tasa de acierto baja ya que lo estandar en la industria es que tengan una precision mayor para apliaciones comerciales. Pero como esto es un mero, ejercicio de clase, lo considero un resultado coherente y factible.