

INTRODUCCION AL ANALISIS MULTIVARIADO

Lab. No.5 - KNN - SVM - ENSAMBLE

Se quiere predecir el tipo de cliente (bueno o malo) usando las siguientes variables que se analizan antes de otorgar un crédito:

- edad: edad del cliente en años cumplidos.
- antig_laboral: antigüedad laboral del cliente en años.
- vivienda: tipo de vivienda (propia, padres, alquilada, contrato privado y otros).
- estado_civil: estado de civil del cliente (soltero, casado, separado, divorciado y viudo).
- trabajo: tipo de trabajo del cliente (asalariado, independiente, temporal y otros).
- ingreso: salario mensual del cliente.
- gasto: gasto mensual del cliente.
- deuda: deuda mensual del cliente.
- ahorro: deuda mensual del cliente.
- patrimonio: valor del patrimonio de cliente.
- porc_deuda: $\frac{deuda}{ingreso} * 100$
- porc_ahorro: $\frac{ahorro}{ingreso} * 100$
- porc_gasto: $\frac{gasto}{ingreso} * 100$
- sobreendeudado: 1 si el cliente esta sobreendeudado y 0 si no.
- plazo: plazo del préstamo solicitado.
- monto: monto del préstamo solicitado.
- garantia: valor de la garantía.
- montoGarantia: $\frac{monto}{garantia} * 100$

K VECINOS MAS CERCANOS

1. Cargue los paquetes `caret`, `DT`, `ROCR`, `class`, `knn`, `e1071`, `adabag` y `randomForest`. Se utilizará la `base2` que se encuentra en "Credit.Rdata"
2. Procedimiento manual:
 - a. Use los primeros 50 datos de `base2` para entrenamiento y los siguientes 10 para validación. Forme las dos bases.
 - b. Haga la clasificación manualmente usando 3 vecinos más cercanos de los 10 datos de validación, usando las distancias a los 50 datos de entrenamiento. Use la distancia euclídea. La variable "cliente" indica el grupo y es la variable objetivo.
 - c. Use la función `knn` de la librería `class` de la siguiente forma `knn(train,test,clase,k = 3)`, donde `train` son los datos de entrenamiento (sin la variable objetivo), `test` son los datos de validación y `clase` es la variable objetivo de la base de entrenamiento.
 - d. Verifique que la clasificación manual coincide con la realizada con la función `knn`.

3. Validación entrenamiento/prueba:

- a. Escoja solo las variables numéricas de base2 y estandarícelas. Luego pegue la variable objetivo y llame a esta base3.
- b. Haga una base de entrenamiento con el 80% de los datos y una base de prueba con el resto.
- c. Lleve a cabo la clasificación usando diferentes cantidades de vecinos y almacene el error de clasificación. Decida un número adecuado de vecinos.

MAQUINAS VECTORIALES DE SOPORTE

4. Usando la misma base de entrenamiento anterior se realizará un SVM para predecir el tipo de cliente usando la función de Kernel Radial para la base de validación o prueba.
 - a. Haga un modelo clasificación usando la función `svm` de la librería `e1071`. Indique `kernel = "radial"` y `probability = TRUE`. Use la base de entrenamiento.
 - b. Haga la tabla de confusión con la base de validación
5. Repita los pasos anteriores:
 - a. Prediga el incumplimiento en la base de validación usando el kernel sigmoidal (`kernel = "sigmoid"`) con la base de entrenamiento.
 - b. Repita el ejercicio usando el kernel lineal (`kernel = "linear"`).
6. Vea el summary de cada modelo y observe cuantos vectores de soporte se requirieron para hacer la función discriminante en cada caso.
 - Obtenga la suma de la diagonal de cada matriz de confusión para determinar cuántos casos fueron bien clasificados en cada caso. ¿En cuál de ellos se hizo una mejor clasificación?
7. Compare los resultados con los obtenidos con otros métodos:
 - a. Compárelo con el método de 9 vecinos más cercanos.
 - b. Compárelo con una regresión logística.
 - c. Verifique si hay se puede asumir que las matrices de covarianza de los dos grupos son iguales.
 - d. Obtenga la clasificación de los datos de validación con el análisis discriminante adecuado y compárela con los otros métodos. Se usa "lda" si es lineal y "qda" si es cuadrático.