Aprendizaje Supervisado

Simulación

El objetivo de este ejercicio es simular diferentes escenarios para visualizar las ventajas y desventajas de las diferentes técnicas que estudiamos

Se considerarán diferentes distribuciones y en cada una de ellas se analizará la performance de los siguientes clasificadores en base a 500 réplicas de datos simulados. Los resultado se pueden mostrar mediantes tablas y boxplots que permitan comparar los diversos procedimientos.

- Análisis Discriminante Lineal
- Análisis Discriminante Cuadrático
- Naive Bayes
- CART
- Bagging
- Random Forest
- Boosting
- Regresión Logística, LASSO.
- Regresión Probit
- k Vecinos más Cercanos
- Support Vector Machine Lineal, Kernels: cuadrático y radial.

Distribuciones a considerar

1. Generar muestras de entrenamiento de tamaño y otra de test de tamaños para cada población $n_1 = 100$ y $n_2 = 100$ Datos en dimensión tres con densidades normales donde una población tenga distribución

$$(X_1, X_2, X_3) \sim \mathcal{N}_3 \left((\mu_1, \mu_1, 1), \Sigma \right),$$

donde la matriz de correlación $\Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, con $\rho = 0$ y la otra con distribución

$$(X_1, X_2, X_3) \sim \mathcal{N}_3\left((c\mu_1, c\mu_1, 1), \Sigma\right),$$

con $c = 2$. Se pueden considerar diferentes valores de c y ver como varían los resultados al juntar o separar los centros de las distribuciones. En este caso las dos primeras variables son informativas ($p_1 = 2$), y la última variable variable no informativa $p_2 = 1$.

2. Misma distribución que en el item 1., pero con datos correlacionados, por ejemplo, $\rho = 0.8$. Se pueden estudiar diferentes valores de este parámetro.

3. Misma distribución que en el item 1., pero con diferentes matrices de correlación. Por ejemplo, para el primer grupo $\rho = 0.5$ y para el segundo $\rho = 0.9$

4. Misma distribución que en el item 3., pero aumentando el número de variables no informativas, $p_2 = 10, 50, 100$.

5. Misma distribución que en el item 3., pero desbalanceando los tamaños muestrales, es decir considerar $n_1 = \alpha n_2$ con $\alpha = 0.1, 0.25, 0.5$.

6. Generar $n = 200$ observaciones. Cada observación esta compuesta de p_1 variables informativas y p_2 variables no informativas, estos conjuntos de variables son independientes ente si. Las primeras p_1 variables se generan

$$X_1 \sim N(0, \Sigma_1),$$

donde Σ_1 es la matriz de correlación, la diagonal tiene unos y fuera de la diagonal la correlación ρ_1.

De manera análoga, las variables no informativas son p_2 y siguen también una distribución normal multivariada. Ahora la matriz de correlación, Σ_2, tiene unos en la diagonal y fuera de la diagonal la correlación ρ_2.

A partir de estos datos determinamos dos grupos.

GrupoA $\sum_{i_1}^{p_1} X_{1i}^2 > c_1$,

GrupoB $\sum_{i_1}^{p_1} X_{1i}^2 \leq c_1$,

Es decir las observaciones dentro del círculo de radio $\sqrt{c_1}$ pertenecen al Grupo B y las que están afuera pertenecen al Grupo A. Considerar $\rho_1 = 0$, $\rho_2 = 0$, $p_1 = 2$ y $p_2 = 1$. Para que las observaciones en ambos grupos sean parejas considerar $c_1 = \sqrt{-2\ln(1-\alpha)}$, con $\alpha = 0.5$, α indica la proporción de observaciones en el grupo 2

7. Misma distribución que en el item 6., cambiando la correlación de las variables informativas.

8. Misma distribución que en el item 6., agregando variables no informativas $p_2 = 10, 50, 100$.

9. Misma distribución que en el item 6., desbalanceando los tamaños muestrales ,para ello hay que modificar el parámetro c_1 de acuerdo con $c_1 = \sqrt{-2\ln(1-\alpha)}$, para $\alpha = 0.1, 0.25, 0.5$.