

Aprendizaje Estadístico - FIUBA

Guía 3 - Clasificación

1. En el archivo `productos.txt` tenemos un registro de productos que fueron exitosos y otros que fracasaron, junto con información de su precio y de su presupuesto en marketing.

1. Levantar los datos del archivo y graficarlos, en negro los productos ?exitosos? y en rojo los ?fracasados?.
2. Para determinar un valor óptimo k de vecinos más cercanos se propone:
 - a) Estandarizar el total de los datos con alguno de los dos esquemas propuestos.
 - b) Particionar el conjunto en un training y un testing set.
 - c) Para un valor de k y un mecanismo de estandarización clasificar los elementos del testing set en función del training set, aplicar como función de costo la cantidad de datos incorrectamente clasificados.
 - d) Repetir con diversos valores de k y elegir como modelo aquel que minimice la función de costo.
3. Clasificar un producto con precio \$70 y un presupuesto de marketing de \$100000 utilizando el modelo óptimo encontrado antes.

2. Considere el archivo `abalone.txt`. Construya un modelo de regresión logística para poder predecir si un espécimen es adulto (M o F) o bien Infante (I), basándose en:

1. Longitud solamente.
2. Peso total solamente.
3. Cantidad de anillos solamente.
4. Los tres parámetros anteriores en su conjunto.

Para la selección del mejor modelo, puede separar los datos en un training y un testing set y evaluar la performance de cada modelo en virtud a la función de costo. Alternativamente, para evaluar performance de la clasificación, considere usar las métricas de True Positives (TP), True Negatives (TN), False Positives (FP) y False Negatives (FN), y a través de ellos calcule métricas como precision, recall, accuracy.

3. Tomemos el data set `abalone.txt`, consideraremos para el mismo el modelo de clasificación logística basado en la longitud, el peso total y la cantidad de anillos. En vez de considerar el umbral de clasificación en 0.5, probaremos distintos valores en una grilla entre 0 y 1. Una vez entrenado el clasificador logístico usando un 70 % de los datos, consideraremos diversos valores de corte para el umbral para hacer el análisis sobre el 30 % restante de los datos de evaluación. Para cada uno de esos umbrales θ , calcularemos:

$$TPR(\theta) = \frac{TP(\theta)}{TP(\theta) + FN(\theta)} \quad FPR(\theta) = \frac{FP(\theta)}{TN(\theta) + FP(\theta)}$$

Graficamos entonces una curva de FPR (eje x) vs TPR (eje y), llamada la curva ROC. El punto (0,1) del espacio ROC corresponde a un clasificador óptimo, cualquier punto en la diagonal será un clasificador perfectamente aleatorio, mientras que en (1,0) tendremos un clasificador perfectamente malo (que también es óptimo si uno invierte sus interpretaciones). Para cada umbral, calcular la distancia del punto (fpr, tpr) con respecto al vértice (0,1) o (1,0) más cercano. Graficar una curva de umbral vs distancia y determinar así el mejor umbral. Arme la matriz de confusión correspondiente a dicho umbral con el 30 % de los datos separados.

4. Implemente una función **univariateLDA** que reciba los siguientes parámetros:

- X vector de valores.
- Y vector de clases para cada valor X.
- k cantidad de clases

y devuelva otra función que sea una predictora. Es decir, debe devolver otra función que permita, dado un valor x , obtener su clase.

5. Simulación.

Genere 1000 datos $Z_i \sim \mathcal{N}(1, 2)$, y asigne la clase 0 si el z_i correspondiente es menor que cero y la clase 1 en caso contrario. Defina $x_i = 0.5 + 5z_i + \epsilon_i$ donde ϵ_i sigue una distribución normal estándar, será el valor observado. Obtenga una clasificación LDA usando la función anterior de un conjunto de entrenamiento elegido de 700 elementos y en base a eso clasifique los 300 elementos restantes, obteniendo métricas de error. Comparar con algún esquema de regresión logística.

6. Considere el archivo `iris.data`, que tiene información sobre la longitud del sépalo, su ancho, la longitud del pétalo y su ancho, todo en centímetros. Finalmente, la última columna tiene información sobre la clase a la que pertenece. Implemente un mecanismo de clasificación por LDA para estos datos utilizando todas las variables.

7. Efectúe una clasificación por LDA de los datos `abalone.txt`, discriminando entre adulto e infante, utilizando aquellas variables que considere más pertinentes. Elija el mejor modelo, usando criterio, sentido común y alguno de los métodos propuestos vistos en clase para determinarlo. Se recomienda separar inicialmente un 20 %, 30 % de los datos para poder hacer la evaluación final a través de una matriz de confusión.