Aprendizaje Estadístico - FIUBA
Guía 2 - Clasificación

---

1. Considere el archivo `abalone.txt`. Construya un modelo de regresión logística para poder predecir si un especimen es adulto (M o F) o Infante, 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 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 derive métricas como precision, recall, accuracy o bien F score.

---

2. 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. Ensaye los siguientes modelos de regresión logística, separando los datos en un training y un testing set
   a) Basado solamente en el precio.
   b) Basado solamente en el presupuesto de marketing.
   c) Basado en ambos factores.
3. Evaluar la performance de los tres modelos y elegir aquel con mejor performance.
4. Con el mejor modelo obtenido, clasifique un producto de precio $70 y un presupuesto de marketing de $100000.

---

3. Consideremos para el data set `abalone.txt` el modelo de clasificación logística basado en la longitud, el peso total y la cantidad de anillos (realizado en el ejercicio 1, ítem d). 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, consideremos 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, calcular:

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

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 sería 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. Genere 1000 datos $Z_i \sim \mathcal{U}(-1, 2)$, 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 training 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. Proceda como en el primer ejercicio, pero para el caso multivariado, llamándolo **multivariateLDA**.

---

7. 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.

---

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

---

9. Compare, para el ejercicio anterior, el "mejor" esquema de clasificación que usted considere entre regresioón logística y análisis discriminante. Comente sobre la métrica o las métricas de comparación que se utilizaron.

---

10. Deducir la función discriminate para el caso del análisis de discriminante cuadrático (es decir cuando las matrices de covarianza no se asumen iguales entre todas las clases).

---

11. Sea $K = 2$ el número de clases. Probar que la regla de decisión en el caso de LDA, luego que los datos son transformados para tener matriz de covarianza la identidad, depende sólo de la distancia de la proyección de cada x_i al subespacio afín que contiene a los centros $\hat{\mu}_1$ y $\hat{\mu}_2$ transformados y del logaritmo de las probabilidades a priori. Reescribir dicha regla en términos de estas distancias. Supongamos que los datos $\{x_i\}_{1 \leq i \leq n} \in \mathbb{R}^p$, $\quad p >> 2$, ¿qué implica el resultado anterior?