Práctica: Regresión Logística

1. Siguiendo la idea de regresión lineal por descenso por el gradiente, adapte el esquema para tener una versión aplicable a regresión logística. Recordando alguna notación:

$$h_\beta(x) = P(Y = 1|x, \beta) = \frac{1}{1 + e^{-\beta^T x}},$$
$$Cost(h_\beta(x), y) = -y \log(h_\beta(x)) - (1 - y) \log(1 - h_\beta(x))$$

de donde tenemos que

$$J(\beta) = -\sum_{i=1}^{n} y^{(i)} \log(h_\beta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\beta(x^{(i)}))$$

Verifique que la regla local de actualización de un parámetro, usando descenso por el gradiente, es:

$$\beta_j := \beta_j - \alpha \sum_{i=1}^{n} (h_\beta(x^{(i)}) - y^{(i)}) x_j^{(i)}.$$

Recuerde de implementar un esquema de **batch update** para la actualización.

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

   (a) Longitud solamente.

   (b) Peso total solamente.

   (c) Cantidad de anillos únicamente.

   (d) Los tres parámetros anteriores en su conjunto.

   Recuerde separar los datos en un **training** y un **testing** set, o bien ensaye un esquema de K-folds, evalúe la performance de cada modelo en virtud a la función de costo definida en el ejercicio anterior, considerando la probabilidad, o bien la cantidad de elementos mal clasificados (es otra alternativa). 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**(`https://en.wikipedia.org/wiki/Precision_and_recall`) o bien **F score** (`https://en.wikipedia.org/wiki/F1_score`). No olvide tampoco considerar la estandarización de las covariables ya que se aplicará un método de descenso por el gradiente (a su elección...).

3. 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.

   (a) Levantar los datos del archivo y graficarlos, en negro los productos "exitosos" y en rojo los "fracasados".

   (b) Ensaye los siguientes modelos de regresión logística, separando los datos en un **training** y un **testing set** o bien, a su criterio, con un esquema de K-folds:

      i. Basado únicamente y hasta linealmente en el precio.

      ii. Basado únicamente y hasta linealmente en el presupuesto de marketing.

      iii. Basado en ambos factores.

   (c) Evaluar la performance de los tres modelos y elegir aquel con mejor performance.

   (d) Con el mejor modelo obtenido, clasifique un producto de precio $70 y un presupuesto de marketing de $100000.

4. (Simulación)

   (a) Genere 1000 datos z_i provenientes de una uniforme $U(-1, 1)$ (**runif**).

(b) Si z_i es menor que 0, entonces asignelo al grupo 0 ($y_i = 0$). Caso contrario, asignarlo al grupo 1 ($y_i = 1$).

(c) Defina $x_i = 2 * z_i + \epsilon_i$ donde ϵ_i sigue una distribución normal estándar. Será el valor "obervado".

(d) Tome los datos (x_i, y_i), con un esquema de K-fold o de training/testing, elija el mejor modelo de regresión logística entre algún esquema basado en un polinomio de x_i. Clasifique posteriormente la totalidad de los datos con el modelo ganador y determine con alguna medida de calidad el ajuste obtenido.

(e) Repita los items anteriores pero ahora considerando $x_i = 0.5 + 2 * z_i + 0.2 * z_i^2 + \epsilon$.

5. (a) Similar al ejercicio anterior, genere z_i provenientes de una uniforme $U(-1, 1)$ y w_i de la misma distribución.

(b) Considere $y_i = 1$ si ambos z_i y w_i tienen el mismo signo, y 0 si no.

(c) Se consideran "observados" y_i, z_i y w_i.

(d) Grafique los datos (z_i, w_i) en rojo si son del grupo cero y en azul si no.

(e) Trate de construir un clasificador logístico de estos datos a través de un esquema lineal en z_i y w_i.

(f) Cómo resultan las predicciones? Por qué podría estar pasando esto?