

Ciencia de Datos

Tarea 3

Para entregar el 3 de mayo de 2018

1. Este ejercicio es sobre la evaluación de diferentes algoritmos de clasificación binaria.

Tenemos un conjunto de prueba y un conjunto de entrenamiento y dos algoritmos de clasificación C_1 y C_2 obtenidos con el mismo conjunto de entrenamiento. Define X_i como la variable que indica si el algoritmo C_i clasifica un dato del conjunto de prueba como 0 o 1 y define Y_i como la variable que indica si el algoritmo C_i lo clasifica correctamente o no. Lo anterior nos da dos tablas de contingencia (una basada en (X_1, X_2) y otra en (Y_1, Y_2)).

- a) Una manera para cuantificar si los clasificadores se comportan de manera similar es verificar si las distribuciones marginales de Y_i son iguales. Muestra que lo anterior es equivalente a verificar si $p_{0,1} = p_{1,0}$ con p las probabilidades subyacentes a la tabla de contingencia de (Y_1, Y_2) . Deriva el estimador de Máxima Verosimilitud para esta hipótesis. Usa un estadístico de prueba basado en razones de verosimilitud para calcular p -valor bajo esta hipótesis para los siguientes datos:

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	8	7
$Y_1 = 1$	11	21

- b) ¿Qué complicaciones habría al utilizar las X 's en lugar de las Y 's?

2. Para datos de clasificación binaria $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, considera la siguiente función de costo:

$$\mathcal{L} = \sum_i (\theta(y_i) - \beta' \mathbf{x}_i - \beta_0)^2 \quad (1)$$

Definimos n_+, n_- el número de observaciones con $y_i = 1$ y $y_i = -1$, respectivamente $\mathbf{c}_+, \mathbf{c}_-$ el centroide de las observaciones con $y_i = 1$, y $y_i = -1$ y \mathbf{c} el centroide de todos los datos.

Como en clase, construimos las matrices:

$$\begin{aligned} \mathbf{S}_B &= (\mathbf{c}_+ - \mathbf{c}_-)(\mathbf{c}_+ - \mathbf{c}_-)' \\ \mathbf{S}_W &= \sum_{i:y_i=1} (\mathbf{x}_i - \mathbf{c}_+)(\mathbf{x}_i - \mathbf{c}_+)' + \sum_{i:y_i=-1} (\mathbf{x}_i - \mathbf{c}_-)(\mathbf{x}_i - \mathbf{c}_-)' \end{aligned}$$

- a) Verifica que

$$\mathbf{S}_W = \sum_{i:y_i=1} \mathbf{x}_i \mathbf{x}_i' + \sum_{i:y_i=-1} \mathbf{x}_i \mathbf{x}_i' - n_+ \mathbf{c}_+ \mathbf{c}_+' - n_- \mathbf{c}_- \mathbf{c}_-'$$

- b) Verifica que el vector $\mathbf{S}_B\boldsymbol{\beta}$, es un múltiplo del vector $(\mathbf{c}_+ - \mathbf{c}_-)$.
- c) Si definimos $\theta(1) = n/n_+$ y $\theta(-1) = -n/n_-$, verifica que en el mínimo de (1):

$$\beta_0 = -\boldsymbol{\beta}'\mathbf{c},$$

$$(\mathbf{S}_W + \frac{n_+n_-}{n}\mathbf{S}_B)\boldsymbol{\beta} = n(\mathbf{c}_+ - \mathbf{c}_-) \quad (2)$$

- d) Usando el resultado de inciso b, argumenta que (2) implica que en el mínimo:

$$\boldsymbol{\beta} \sim S_W^{-1}(\mathbf{c}_+ - \mathbf{c}_-),$$

es decir la solución coincide con la del Fisher Discriminant Analysis (FDA).

- e) Lo anterior permite implementar FDA usando algún algoritmo de mínimos cuadrados. En R lo haremos con la función `lm()`. Ilustra cómo funciona el método con algunos conjuntos de datos en 2D bien elegidos.
- f) Observamos que (1) muestra que FDA **no** es muy robusto a datos atípicos. Una posibilidad para hacerlo más robusto es usar mínimos cuadrados ponderados. Por ejemplo `lm()` tiene un argumento opcional `weights` donde se pueden proporcionar pesos $w_i, i = 1, \dots, n$ para minimizar:

$$\sum_i w_i (y_i - \boldsymbol{\beta}^t \mathbf{x}_i - \beta_0)^2.$$

¿Cómo elegirías estos pesos? Verifica tu propuesta con algunos ejemplos en 2D.

3. Este ejercicio es sobre el método de clasificación binaria perceptron.
 - a) Implementa el modelo clásico de perceptrón (versión que trabaja en línea). Aplícalo primero a un conjunto de datos artificiales en dos dimensiones y con dos categorías. Discute tus resultados. Incluye gráficas del ajuste.
 - b) Aplica el clasificador al conjunto de datos `pima` que vimos en clase. Usa `pima.tr` para ajustar el modelo y `pima.te` para verificar su calidad predictiva. ¿Qué puedes decir sobre su desempeño? Comenta tus hallazgos.
4. Implementa Kernel FDA. Puedes basarte en el artículo de Mika et al: *Fisher Discriminant Analysis with kernels*. Verifica su desempeño en un conjunto de datos artificiales *apropiados* en dos dimensiones y dos clases. Compara su desempeño con FDA estándar en el conjunto de datos `pima` del ejercicio 3b. Comenta tus hallazgos.

Opcional (puntos extra): realiza la implementación como una librería en C o C++ para R. Incluye la interfaz (o wrapper) adecuado en R para usarse.

5. Los archivos contenidos en las carpetas `email_train` e `email_test` corresponden a correos electrónicos en inglés clasificados como Spam y No-Spam.
 - a) Implementa clasificadores de Spam usando regresión logística, LDA, QDA, FDA, Kernel FDA y redes neuronales. Usa los datos `email_train` e `email_test` para ajustar y probar los métodos, respectivamente. Compara su desempeño.
 - b) Las curvas ROC (Receiver Operating Characteristics) es un método muy común para comparar algoritmos de clasificación binarios basado en la tabla de errores (falsos positivos y falsos negativos) que se cometen. Usa los resultados del inciso anterior para comparar los clasificadores usando este criterio. ¿Cuál método elegirías? Usa el criterio del área bajo la curva (AUC).

Notas: Los correos-e están como documentos EML, con la estructura que usan la mayoría de los administradores de correo. Puedes explotar esa estructura para extraer la información adecuada. Por ejemplo, la librería `tm.plugin.mail` puede usarse junto con `tm` (que ya usaste antes) para leer textos con esta estructura, así por ejemplo,

```
corp <- Corpus(DirSource(rutTrain,recursive=TRUE),
               readerControl=list(language="en_US",reader=readMail))
```

te crea un `Corpus` con correos electrónicos donde cada elemento de la lista del `Corpus` tiene (al menos) los objetos `$content` y `$meta`, que contienen el texto y otros elementos del mensaje, respectivamente.

Usa como características (o covariables) información de los correos basado en frecuencia de palabras y otras que tu consideres importantes. Pon especial atención al preproceso que hagas. Documenta todos los pasos y criterios que usaste.

Hay bastante literatura sobre ROC, como referencia, puedes consultar el paper de T. Fawcett, An Introduction to ROC Analysis. En R, la librería `pROC` puede ser útil.