Tarea III:
Temas Selectos de Estadística

Por entregar el 2 de junio;

1. **Este ejercicio es sobre Fisher Discriminant Analysis**
   Para datos de clasificación binaria $\{(x_i, y_i)\}$, considera la siguiente función de costo:
   $$\sum_i (\theta(y_i) - \beta^t x_i - \beta_0)^2. \tag{1}$$

   Definimos n_+, n_- el número de observaciones con $y_i = 1$, resp. $y_i = -1$, c_+, c_- el centroide de las observaciones con $y_i = 1$ resp. $y_i = -1$ y c el centroide de todos los datos.

   Como en la clase, construimos las matrices:
   $$S_B = (c_+ - c_-)(c_+ - c_-)^t,$$
   $$S_W = \sum_{i:y_i=1} (x_i - c_+)(x_i - c_+)^t + \sum_{i:y_i=-1} (x_i - c_-)(x_i - c_-)^t$$

   a) Verifica que
   $$S_W = \sum_{i:y_i=1} x_i x_i^t + \sum_{i:y_i=-1} x_i x_i^t - n_+ c_+ c_+^t - n_- c_- c_-^t.$$

   b) Verifica que el vector $S_B\beta$, es un múltiple del vector $(c_+ - c_-)$.

   c) Si definimos $\theta(1) = n/n_+$ y $\theta(-1) = -n/n_-$, verifica que en el mínimo de (1):
   $$\beta_0 = -\beta^t c,$$
   $$(S_W + \frac{n_+ n_-}{n} S_B)\beta = n(c_+ - c_-) \tag{2}$$

   d) Usando el resultado de inciso b, argumenta que (2) implica que en el mínimo:
   $$\beta \sim S_W^{-1}(c_+ - c_-),$$
   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 será a través de la función `lm()`. Ilustra como 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 puede dar un vector $\{w_i\}$ para minimizar:
   $$\sum_i w_i(\theta(y_i) - \beta^t x_i - \beta_0)^2.$$

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

2. **Este ejercicio es sobre clasificación de textos**

Clasificación de textos surge en diferentes contextos: por ejemplo clasificar según autor, tema o género literario. Es muy común en estos problemas representar un texto a través de un vector donde cada entrada es asociada con una palabra particular y refleja cuantas veces esta palabra aparece en el texto. Depende del problema como elegir estas palabras *de referencia*.

Lee el siguiente artículo de divulgación para entender un problema muy específico del área de stilometría:
Who wrote the 15th Book of Oz? an application of multivariate analysis to authorship attribution que se puede accesar desde
`http://www.cs.toronto.edu/~gh/Courses/2528/Readings/Binongo.pdf`

a) Trata de reproducir/aproximar los resultados que ellos reportan uando el método de componentes principales. ¿ Qué nos da FDA?

Puedes accesar los textos desde `www.gutenberg.org` y el libro 15 (ver texto) desde
`http://www.welcometooz.net/etext-15.txt`.

Si quieres, puedes usar el software *Rainbow* para sacar los conteos de las palabras. Corre bajo linux y se puede bajarlo desde
`http://www.cs.cmu.edu/~mccallum/bow/`
(Nota: en mi máquina no había necesidad de hacer paso 2 del INSTALL)

Comandos útiles son:

```
./rainbow -d ~/model --no-stoplist --index ~/oz ~/oz
./rainbow -d ~/model --print-word-probabilities=oz
```

b) Busca tu ejemplo favorito para probar diferentes algoritmos de clasificación para clasificar textos. Quizás el siguiente artículo puede dar algunas ideas:
http://www.cs.toronto.edu/~gh/Courses/2528/Readings/Koppel-etal.pdf

Otro fuente puede ser el conjunto de artículos de diferentes newsgroups: http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes/20_newsgroups.tar.gz

Algunos comentarios:

• Usa algoritmos de clasificación basada en FDA, Kernel FDA, PCA, Kernel PCA (y usar después tu clasificador favorito). Una

variante es proyectar los datos usando (Kernel) FDA y usar un método como vecino más cercano después.

• En la aplicación de libro de Oz, se tomaron como palabras *de referencia* las más frecuentes. Para clasificación de textos por tópico, es más común trabajar con aquellas palabras que más distinguen las categorías. Por ejemplo, sean (X_p, Y) la frecuencia (relativa) de una palabra en un documento y la categoría del documento respectivamente; podemos elegir como palabras *de referencia* aquellas p que muestran la mayor dependencia entre X_p y Y. Es muy popular medir la dependencia usando la información mutua (ver notas curso primer semestre).

• Limítate a clasificación binaria.

• Puedes usar la library `kernlab` para (Kernel) PCA. Kernel FDA tendrás que implementar (más información sigue).