Tratamiento de Datos: Examen de Laboratorio

January 24, 2017

1 Descarga de datos

Cada estudiante debe trabajar sobre un fichero de datos diferente, que puede descargarse de la siguiente URL (no olvide reemplazar su número de estudiante al copiar la dirección en su navegador):

```
http://www.tsc.uc3m.es/~jarenas/MIT2017/YourNIA_data.mat
```

Durante esta sesión trabajará con las siguientes variables

```
xtr_reg, str_reg, xval_reg, k_knn, xtr_rl, xval_rl, ytr_rl, yval_rl
```

que pueden obtenerse empleando los siguientes comandos de python:

```
>>> import scipy.io
>>> data = scipy.io.loadmat('YourNIA_data.mat')
>>> xtr_reg = data['xtr_reg']
>>> str_reg = data['str_reg']
>>> ...
```

2 Parte 1: Regresión (50%)

Las variables mencionadas incluyen un conjunto de entrenamiento que consta de 500 datos consistentes en pares de entrada-salida: $\mathcal{D} = \{\mathbf{x}^{(i)}, s^{(i)}\}_{i=1}^{500}$. Los 500 vectores de entrada se proporcionan como las filas de la variable `xtr_reg`, mientras que sus etiquetas correspondientes están disponibles en el vector `str_reg`. Asimismo, la variable `xval_reg` contiene 100 datos de validación, con la misma dimensionalidad que los datos de entrada del conjunto de entrenamiento. Utilice estas variables tal y como se proporcionan, sin emplear ningún procedimiento de normalización.

(a) Obtenga el estimador de la forma $\hat{s} = s_0$, con s_0 una constante, que minimiza el error cuadrático promedio medido sobre el conjunto de entrenamiento. Calcule también el error cuadrático promedio de dicho estimador sobre los datos de validación. Almacene sus resultados en las variables `s0` y `E2val`.

(b) Obtenga las predicciones sobre los datos de validación de un modelo k-nn utilizando el valor de K que le ha sido asignado, `k_knn`. Calcule el promedio de las predicciones de los datos de validación, y el error cuadrático promedio del modelo sobre los datos de validación. Almacene sus resultados en las variables `s_prom` y `E2val_knn`.

(c) Asuma a continuación que los datos se han generado de acuerdo con el siguiente modelo:

$$s^{(i)} = [1\ ,\ x_1^{(i)}, \cdots, x_N^{(i)}\ , \exp(x_1^{(i)}), \cdots, \exp(x_N^{(i)})\ ]\mathbf{w} + \varepsilon^{(i)}, \quad 1 < i < 500$$

donde N es la dimensión del espacio de entrada. Asuma también que el vector de pesos y las muestras de ruido siguen distribuciones a priori $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, 2\mathbf{I})$ y $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, y que las muestras de ruido $\{\varepsilon^{(i)}\}_{i=1}^{500}$ son independientes entre sí e independientes de $\mathbf{w}$. Para $\sigma_\varepsilon^2 = 0.5$ obtenga el vector de medias y la matriz de covarianzas a posteriori de $\mathbf{w}$. Guarde sus resultados en las variables `w_mean` y `w_cov`.

No olvide guardar las variables

```
s0, E2val, s_prom, E2val_knn, w_mean, w_cov
```

3 Parte 2: Clasificación (50%)

Cada una de las matrices de datos `xtr_rl` y `xval_rl` contiene 500 vectores de datos con 10 dimensiones.

Se sabe que las etiquetas binarias `ytr_rl` e `yval_rl` (con valores en $\{0, 1\}$) se han generado de acuerdo a un modelo de regresion logística

$$p\{y = 1|\mathbf{w}_e, \mathbf{x}\} = \frac{1}{1 + \exp(-\mathbf{w}_e^T \mathbf{x}_e)}$$

siendo $\mathbf{x}_e = (1, \mathbf{x}^T)^T$ y $\mathbf{w}_e = (w_0, w_1, \ldots, w_{10})$. Se sabe además que uno de los coeficientes w_i es nulo. Por tanto, una de las variables, x_i es irrelevante para la tarea de clasificación, si bien se ignora cuál es el valor de i. Consecuentemente, se desea ajustar un modelo que incluya únicamente las variables relevantes:

(a) Calcule la media y la varianza de cada componente de la matriz `xtr_rl`. Guarde el resultado en las variables `mx` y `sx` (vectores de dimensión 10).

(b) Normalice los datos de entrenamiento y validación de modo que todas las variables de la matriz de datos de entrenamiento tengan media cero y varianza unidad. Guarde el resultado en las variables `xn_tr` y `xn_val`, respectivamente.

(c) A continuación, con los datos normalizados, aplique 100 iteraciones de un algoritmo de descenso por gradiente para aproximar el estimador MAP de los coeficientes del modelo de regresión logistica, suponiendo un prior gaussiano. Utilice para ello un valor de $C = 100$ y un paso de adaptacion $\rho = 0.001$, e inicialice el vector de pesos a cero. Almacene el vector de pesos resultante en la variable `w10`.

Guarde también la variable `rho` con el valor del paso de adaptación. Si tuviera problemas de convergencia con el valor propuesto propuesto, puede utilizar otro, pero no olvide guardar el valor utilizado.

(d) Con el estimador obtenido, uilice todas las variables, y determine la log-verosimilitud negativa, medida sobre los datos de validación (`xn_vall` e `yval_rl`). Almacene el resultado en la variable `lg10`.

(e) Determine el número de muestras de validación asignadas por el clasificador a la clase 1. Guarde el resutado en la variable `n1`

(f) Entrene 10 modelos de regresión logística diferentes, omitiendo en cada ocasión una variable, de tal modo que el modelo i-ésimo utilizará todas las variables menos x_i. Calcule la tasa de error de clasificación (sobre los datos de validación) para cada uno de los 10 casos, y conserve el mejor resultado. Guarde el menor error de validación en la variable **`emin`**, un entero indicando el número de variables del modelo en **`nvar`**, y el vector de pesos correspondiente (únicamente para el mejor caso) en la variable **`wmin`**.

No olvide guardar las variables:

```
mx, sx, xn_tr, xn_val, w10, rho, lg10, n1, emin, nvar, wmin
```

4 Entrega de resultados

Guarde todas las variables solicitadas en las secciones previas en un fichero con nombre **`results.npz`**. Use para ello el siguiente comando de numpy:

```
>>> np.savez('results.npz', s0=s0, E2val=E2val, s_prom=s_prom, ...
    E2val_knn=E2val_knn, w_mean=w_mean, w_cov=w_cov, ... )
```

Cree un fichero .zip en el que debe incluir el fichero **`results.npz`** además del script de python que haya escrito para resolver el examen. Suba el fichero comprimido a Aula Global en el plazo indicado por el profesor.