Facultad de Matemática, Astronomía y Física y Computación
Universidad Nacional de Córdoba

Ciencia de Datos

Práctico N°6: Regresión Lineal y Logística

Problema 1: La regresión como clasificador.
Considerar un problema con dos clases: `(0,1)` no correlacionadas. Los datos de la primer clase provienen de una distribución gaussiana ($\mu = 0{,}5, \sigma = 0{,}5$), mientras que los de la segunda clase de una distribución gaussiana ($\mu = 2{,}5, \sigma = 0{,}5$).
a) Generar 50 datos sintéticos para cada clase y graficarlos usando la clase como ordenada.
Ayuda: Para los siguientes dos ítems, estudiar la entrada de `scikit-learn` sobre la función logística.
b) Ajustar los datos con una recta utilizando el modelo `LinearRegression` de `scikit-learn`. Tener en cuenta que los modelos requieren los datos como vectores columnas. Para trasponer un arreglo puede usarse: `x = X.reshape((-1, 1))`, donde `X` son los todos los datos concatenados. Superponer la recta de ajuste en el gráfico con los datos:

```
linear = LinearRegression().fit(x, y)
y_lin = linear.coef_* X + linear.intercept_
```

c) Ajustar los datos con la función logística utilizando el modelo `LogisticRegression` de `scikit-learn`. Superponer la función ajustada en el gráfico anterior:

```
X_test = np.linspace(0, 3, 300)
y_log = expit(X_test * logistic.coef_+ logistic.intercept_).ravel()
```

donde la función `expit` es la función logística o sigmoide, definida por $\text{expit}(x) = 1/(1 + exp(-x))$, y la provee `scipy`: `from scipy.special import expit`.
d) Discutir cómo usar las regresiones para clasificar los datos. ¿Cómo puede asignarse probabilidad a cada clase en la clasificación usando las regresiones?
e) Reconstruir el gráfico anterior usando ahora los mismos datos generados para la clse 0, mientras que los de la segunda clase sintetizarlos usando $\mu = 1{,}5, \sigma = 0{,}5$.
f) Observar cómo se modificaron las regresiones. Leer *completo* el siguente hilo de #`estadisticaXtuiter` del Prof. Walter Sosa Escudero.

Problema 2: Implementar `Perceptron` para clasificar el Breast cancer Wisconsin dataset provisto en `scikit-learn`.
a) Evaluar el modelo imprimiendo un `classification_report` y la matriz de confusión.
b) Comparar con los resultados del modelo de Naïve Bayes de la guía anterior.

Problema 3: Implementar LogisticRegressionCV para clasificar el Breast cancer dataset.
a) ¿Por qué motivo este modelo utiliza cross validation? Interpretar los siguientes valores de los parámetros: `cv=5, penalty='l2', solver='liblinear', tol=1e-6, max_iter=int(1e6)`.
b) Evaluar el modelo imprimiendo un `classification_report` y la matriz de confusión. Comparar con los resultados de los modelos implementados anteriormente.