Lab1 - Machine Learning

Agustí Serra 205271
Alba Torra 194994

En aquest laboratori se'ns demana aplicar alguns dels conceptes vistos a teoria, com ara els models de regressió i la predicció de resultats a partir de l'anàlisi de diversos atributs relacionats entre si, que pertanyen a un mateix camp, repassant també algunes de les peces claus d'estadística, com ara el càlcul de mitjanes.

Exercici 1

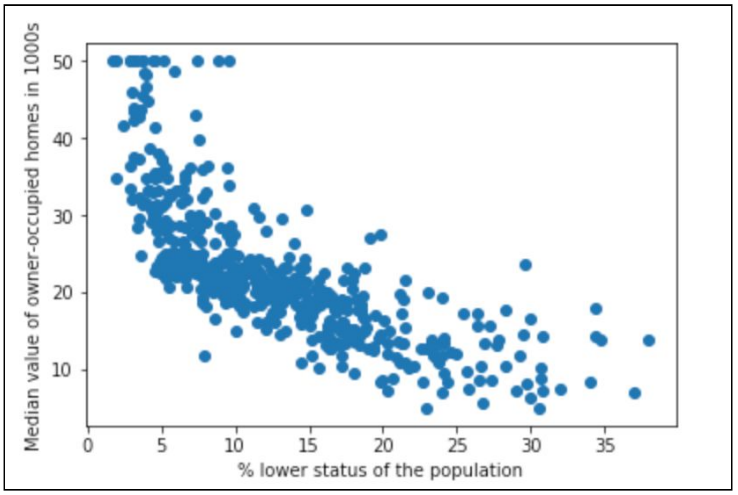
Al primer exercici treballem amb només 3 atributs: RM, LSTAT i PTRATIO. Hem de calcular la mitjana, la mediana, la std dev, màxims i mínims. El que farem serà bàsicament aprofitar les funcions predefinides de les llibreries de python per trobar el que se'ns demana, tal i com veiem al codi, afegint np. davant de cada funció. En el cas de la std dev primer haurem de trobar la variança. Per la correlació, aprofitarem també la funció corr() per calcular-la amb els 3 atributs en funció de medv i, tot seguit, fem un mapa per representar totes les correlacions possibles amb les dades de les que disposem. D'aquesta manera veiem que les correlacions més altes amb MEDV són RM, LSTAT i PTRATIO. Finalment, per fer l'histograma simplement utilitzarem la funció hist(). Realitzant l'histograma apreciem fàcilment que la distribució de les dades es polinomial. Tot això és fàcil de veure amb el codi correctament comentat.

Exercici 2

En el segon exercici ens demanen que treballem amb l'atribut més significatiu que en aquest cas és LSTAT. Tenim que determinar quin tipus de regressió és millor per poder definir aquest atribut, creiem que es el polinomial amb grau 3, ja que quan fem la gràfica d'aquest atribut amb l´atribut de MEDV ,que en aquest cas es tracta de la "y" en la nostra fórmula polinomial, ens surt un plot que al principi vem creure que es tractava d'una regressió exponencial pero vem pensar que seria més fàcil tractarla com si fos una polinomial de grau 3.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Aquesta és la gràfica, i com podem veure té forma de paràbola:



Un cop fet això, el que intentem calcular és el vector que, multiplicat per la matriu "x", ens doni la "y". En aquest apartat vam tenir molts problemes ja que, al utilizar la funció que vam fer al p3, ens donaven molts errors de dimensions. La solución que vam trobar va ser juntar dues matrius, la de LSTAT i la de RM per tenir una matriu de 2 dimensions i així no ens va donar cap problema.
Pel que fa als altres dos apartats, no hem sigut capaços de conseguir exactament el que se'ns demana. Hem escrit el codi que considerem hauria de funcionar segons les especificacions que ens donen, però hi ha quelcom que no acaba de funcionar i no hem sapigut localitzar l'error.

Exercici 3

El que hem fet a l'exercici 3 ha estat desenvolupar la regressió Ridge genèrica per Machine Learning a partir de la fòrmula vista a teoria:

$$E_R(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(\mathbf{x}_n, \mathbf{w}) - t_n)^2 + \lambda\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Aquest tipus de regressió ens costa d'entendre i assimilar, així que hem fet el que hem pogut. No hem estat capaços d'implementar-la amb el descens de gradient, ja que aquest concepte es veia en l'exercici 2 i no l'hem acabat de treure del tot, pel que no hem sigut capaços d'aplicar-ho a l'exercici 3. El codi està comentat al .ipynib per acabar d'entendre algunes especificacions.

Exercici 5

Per acabar el lab, se'ns demana la predicció del preu de 3 habitatges tenint en compte diversos atributs. Aquí és on es reflexa la importancia del machine learning. El que fem és calcular els pesos amb la closed form per poder obtenir posteriorment les prediccions desitjades.

Els preus obtinguts serien els que es mostren al executar el codi.

Els resultats semblen adients a les condicions (atributs) inicials.