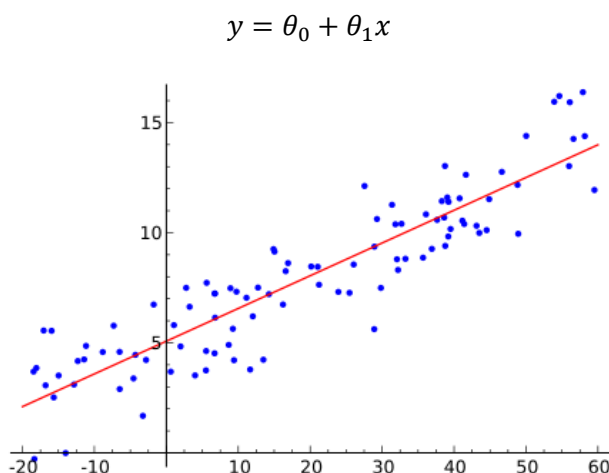


Sklop 2: Linearna regresija

Uvod

Linearna regresija je postopek za modeliranje povezave med odvisno in neodvisnimi spremenljivkami. Rezultat je model, ki na podlagi vrednosti vhodnih, neodvisnih spremenljivk napove vrednost izhodne, odvisne.

Če modeliramo eno neodvisno spremenljivko govorimo o preprosti linearni regresiji. Če pa modeliramo več neodvisnih spremenljivk, pa govorimo o multipli linearni regresiji (*ang. multiple linear regression*). Na sliki je predstavljen primer preproste linearne regresije za napovedovanje ene odvisne spremenljivke z eno neodvisno.



Slika: Linearna regresija za napovedovanje ene odvisne spremenljivke z eno neodvisno.

Pri multipli regresiji z ustreznim postopkom določimo vrednosti parametrov modela θ_j tako, da lahko napovemo vrednost y za poljuben vzorec, za katerega poznamo vrednosti njegovih značilk x . Model zapišemo v obliki:

$$y^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots + \theta_n x_n^{(i)}$$

Pri čemer je $y^{(i)}$ vrednost napovedi i -tega vzorca, $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}]$ so značilke i -tega vzorca, vrednosti $\theta = [\theta_0, \theta_1, \dots, \theta_n]$ so parametri modela, n število značilk.

Predstavitev problema

Podjetje, ki se ukvarja z dobavo kurilnega olja strankam želi na podlagi podatkov obstoječih strank napovedati, koliko kurilnega olja bo potrebovalo za nove stranke.

Predstavitev podatkov

Posamezni vzorci podatkov predstavljajo stranke podjetja za dobavo kurilnega olja. V podatkih obstoječih strank je zajeta debelina izolacije, povprečna zunanja temperatura, število stanovalcev, povprečna starost stanovalcev, velikost hiše in količina naročenega olja. Vsi podatki so numerični.

Naloga

Z multiplo linearno regresijo določite skupno količino kurilnega olja za nove stranke, katerih podatki se nahajajo v datoteki HeatingOilTest na podlagi učne množice obstoječih strank iz datoteke HeatingOilTrain. Določite tudi povprečno količino olja za nove stranke.

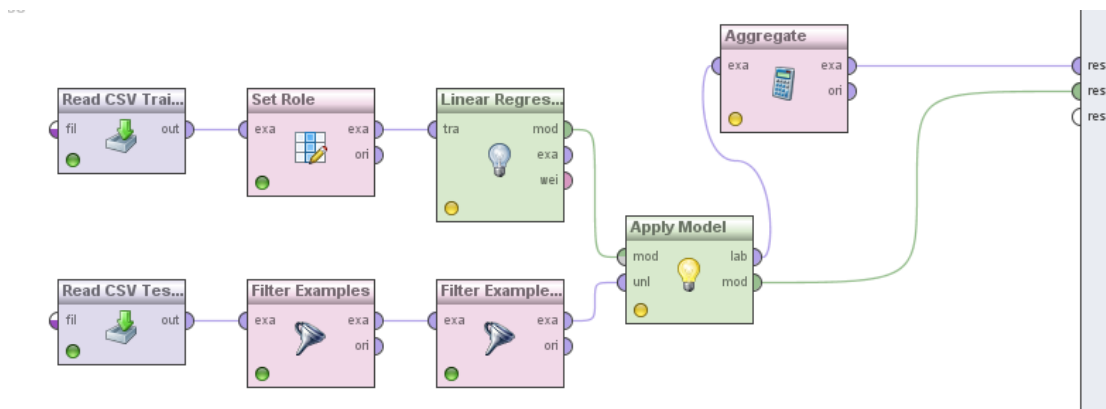
Naložite podatke HeatingOilTrain.csv in HeatingOilTest.csv z operatorjema *Read CSV*. Učni množici podatkov nastavite vlogo **label** značilki Heating_Oil.

Oglejte si naložene podatke testne in učne množice s pogledom Meta Data View. Bodite pozorni na razpon podatkov (Range). Testni podatki vsebujejo vzorce izven razpona podatkov učne množice. Vzorce testne množice je potrebno filtrirati. Z operatorjem *Filter Examples* iz testne množice odstranite vse vzorce, ki so izven območja vrednosti učne množice. V operatorju *Filter Examples* si v ta namen kot pogoj izberite opcijo *attribute_value_filter* in vtipkajte parameter *string* $Avg_Age \geq 15.1$, ter $Avg_Age \leq 72$.

S testnimi podatki naučite model linearne regresije z operatorjem *Linear Regression*.

Model uporabite na podatkih z operatorjem *Apply Model*. Opazujte model in napovedi.

Za izračun skupne količine napovedi potrebnega kurilnega olja ter povprečno vrednost le tega uporabite in ustrezno nastavite še operator *Aggregate*.



Enačba multiple linearne regresije za napoved porabe kurilnega olja:

Skupna raba kurilnega olja iz testnih podatkov: _____

Povprečna raba kurilnega olja na hišo iz testnih podatkov: _____

Viri

-Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining. Practical Machine Learning. Tools and Techniques. Third Edition

- <http://rapid-i.com/>

- <https://sites.google.com/site/dataminingforthemasses/>

- http://en.wikipedia.org/wiki/Linear_regression

Vaja pripravljena v okviru KCUI delavnice na UL-FE