# Dataset preprocessing and interpretation

1 Abandon the rows that contain ’ ?’



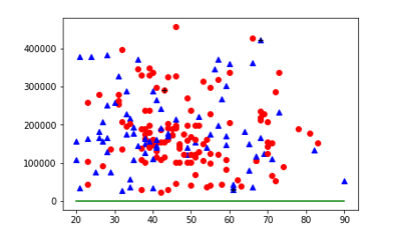
2 Dealing with discrete (categorical) features:

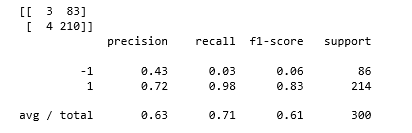


Use digital to label the different feature.

3 10 fold cross validation

# Implement a linear soft-margin SVM





The y axis (fnlwgt) is so big compared with x axis (age), so it seems the decision boundary is a line of y=0. Actually, the line is y = 89.54\*x + 475357.11 (w=89.54, b=475357.11).

To saving time, I set 300 data to test.

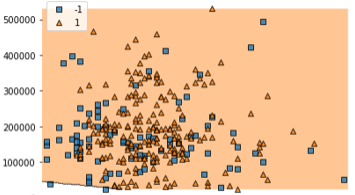
The data we use is non-linear separable. Hence, if the value of C is big enough and limited iteration times, the performance will be optional.

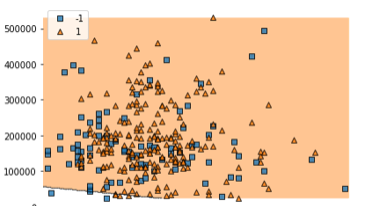
C=0.1 69.2%

C=1 71.0%

C=10 71.3%

C=100 70.6%

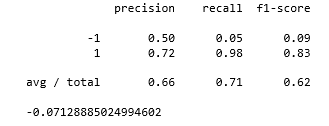
C=100

c=0.01

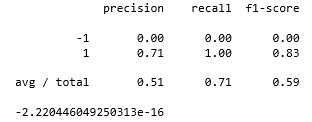
The data we use is non-linear separable. Hence, if the value of C is big enough and limited iteration times, the performance will be optional. I used sklearn Grid-Search to get the best C, and the result is 5 or larger than 5. The accuracy using 10-fold cross validation is nearly 72%.

# Implement a kernel SVM

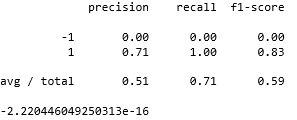
Linear Kernel



RBF



Poly



I use K-NN to get accuracy. A nd the result is 0.8342. 