**MÁQUINAS VECTOR SOPORTE**

Tenemos datos de diferentes centros educativos de un territorio. Se han recogido varias variables que representan las características de estos centros como, por ejemplo: si son un centro público o privado, el número de estudiantes matriculados, el gasto que hacen por estudiante, la ratio de estudiantes/aula, etc.

Con estos datos queremos clasificar a diferentes centros en centro Público (1) y centro Privado (0) en función del número de matriculados y la ratio estudiante/aula. Nuestra hipótesis es que el modelo clasificará como centro público aquellos casos donde haya más estudiantes matriculados y más estudiantes/aula, con lo cual, por encima del hiperplano quedarán los casos del centro público y por debajo los del centro privado.

from seaborn import load\_dataset, pairplot

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from sklearn.svm import SVC

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score

df = pd.read\_csv("C:/Users/alici/Desktop/Master/Aprendizaje Automático/Python/U.S. Colleges.dat", sep="\t")

print(df)

df=df.dropna()

print(df)

print(set(df['PUBLICO']))

#x=variables numericas

x=df[['MATRICULADOS', 'STUAULA']]

y=df['PUBLICO']

print(x)

print(y)

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, random\_state=100) #por defecto

print(x\_train) #por defecto: 75% de los datos para el entrenamiento y el 25% test

print(x\_test)

print(y\_train)

clf = SVC(kernel='linear')

clf.fit(x\_train, y\_train)

from seaborn import scatterplot

w = clf.coef\_[0]

b = clf.intercept\_[0]

print(w)

print(b)

x\_visual = np.linspace(2,10)  #eje x del grafico posterior

print(x\_visual)

y\_visual = -(w[0] / w[1]) \* x\_visual - b / w[1]

print(y\_visual)

print(scatterplot(data = x\_train, x='MATRICULADOS', y='STUAULA', hue=y\_train))

plt.plot(x\_visual,y\_visual)

plt.show()

#con datos train

Como podemos ver en el gráfico el hiperplano clasifica bien a pesar de unos pocos casos. Como imaginábamos, los casos que pertenecen al centro público quedan por encima del hiperplano y los que pertenecen al centro privado quedan por debajo. Se confirma nuestra hipótesis, pero antes asegurémonos de que nuestro modelo de MVS tiene un buen ajuste, porque podría no haber encontrado el mejor de los hiperplanos.

#con datos test

predictions = clf.predict(x\_test)

print(predictions[:10]) #10 primeras predicciones

#ajuste

print(accuracy\_score(y\_test, predictions))

plt.show()

Salida:

[0 0 0 0 0 0 1 0 0 0]

0.9333333333333333

Hemos predicho los 10 primeros casos y el resultado es que 9/10 los predice como pertenecientes al centro público y solo uno al centro privado.

Por último, obtenemos un ajuste de 0.93 (casi 1), es un ajuste muy bueno con lo cual, este modelo de MVS clasifica muy bien a los centros en centros públicos/privados en función al número de estudiantes matriculados y a la ratio estudiante/aula.