CS498AML_HW3_panz2

PanZhang

9/22/2018

	0N	1N	2N	3N	4N	0c	1c	2c	3c	4c
Dataset I	4.5424707	0.3834503	0.1755630	0.1417837	0.1608384	4.5431190	0.3846135	0.1778153	0.1444405	0.1608384
Dataset II	4.5424707	0.6410932	0.7156285	0.9083929	1.1156579	4.5495390	0.6486421	0.7506211	0.9419728	1.1156579
Dataset III	4.5424707	1.2903725	1.9672404	2.6508411	3.6532797	4.5574730	1.3234622	2.1197481	3.0273799	3.6532797
Dataset IV	4.5424707	0.7999427	0.8280826	0.9849498	1.1940000	4.5661987	0.8406142	1.2070898	1.2711920	1.1940000
Dataset V	4.5424707	1.9177678	3.3317221	4.5482572	5.1392667	4.9199280	2.8356794	4.6514345	4.9712473	5.1392667

```
os.getcwd()
os.chdir('/Users/panzhang/Desktop/CS498AML/HW3/')
data1=pd.read_csv('dataI.csv',sep=',')
data2=pd.read_csv('dataII.csv',sep=',')
data3=pd.read_csv('dataIII.csv',sep=',')
data4=pd.read_csv('dataIV.csv',sep=',')
data5=pd.read_csv('dataV.csv',sep=',')
iris=pd.read_csv('iris.csv',sep=',')
iris.columns=["X1","X2","X3","X4"]

#data reconstruction and MSE calculation
def new_data(data,s,NC,d_nless=iris):
    num_data, dim = data.shape
    if (NC == "N"):
        mean_data = d_nless.mean(axis=0)
        X=np.cov(iris.T)
    else:
        mean_data = data.mean(axis=0)
        X = np.cov(data.T)
    e, EV = np.linalg.eigh(X)
    idx = e.argsort()[::-1]
    e = e[idx]
    EV = EV[:, idx]
    new_df = pd.DataFrame(columns=["X1","X2","X3","X4"])
    s=s
    error=0
    for i in range(num_data):
        #x=np.zeros(shape=(1,dim))
        #i=0
        x=[0,0,0,0]
        for j in range(s):
            x = x + np.dot(EV[:,j].T,(data.iloc[i]-mean_data))*EV[:,j]
        new_df.loc[i]= x + mean_data
        error = error + pow(np.linalg.norm(iris.loc[i] - new_df.loc[i]),2)
    mse=error/num_data
    return (new_df,mse)

# 50 MSE numbers
mse = np.zeros((5,10))
for i in range(5):
    data="data" + str(i + 1)
    for j in range(5):
        mse[i,j] = new_data(eval(data), j, "N")[1]
        mse[i,j+5] = new_data(eval(data), j, "C")[1]
mse
#np.savetxt('panz2-numbers.csv', mse, delimiter = ',')

#reconstruction of the dataset of version II, expanded onto 2 principal components,
    where mean and principal components are computed from the dataset of version II
data2_recon=new_data(data2,2,"C")[0]
data2_recon.to_csv("panz2-recon.csv", index=False, header=True)
```