classificacao

November 3, 2016

Projeto da disciplina de Data Mining
PESC - Programa de Engenharia de Sistemas e Computação
COPPE / UFRJ
Autor: Rafael Lopes Conde dos Reis
E-mail: condereis@cos.ufrj.br
GitHub: https://github.com/condereis/data-mining

1 Resumo

O trabalho consiste em classificar os dados de uma base desconhecida. Nesta planilha existe uma coluna que indica a classe das amostras, algumas das quais são desconhecida, sendo indicadas com "?". A tarefa consiste em realizar a dessas classificação destas amostras, fornecendo como saída seu ID e a respectiva classe. Deverá ser usado o Naive Bayes como baseline e dois (ou mais) classificadores distintos para realizar a mesma tarefa. Por fim, deverá ser feia uma análise dos resultados obtidos.

2 Pacotes Utilizados

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        from sklearn.cluster import KMeans
        from sklearn.decomposition import PCA

        %matplotlib inline
```

3 Análise Exploratória

Primeiramente foi feita uma análise exploratória das variáveis, plotando suas distribuições e um resumo estatístico das mesmas.

```
In [2]: raw = pd.read_excel('../data/classificacao/classificacao-entrada.xls', deci
        raw.tail()
```

Out[2]:

	id	X1	X2	X3	X4	X5	X6	X
10788	10798	17	0.999410	0.999364	0.999789	0.999744	0.999023	0.99897
10789	10799	17	0.999288	0.999228	0.999348	0.999287	0.998888	0.99882
10790	10800	17	0.999278	0.999370	0.999058	0.999150	0.999058	0.99915
10791	10801	17	0.999388	0.999523	0.999366	0.999502	0.999366	0.99950
10792	10802	17	0.999430	0.999634	0.999703	0.999908	0.998936	0.99914

	X8	X9	X10	X11	X12	Classe
10788	0.999789	0.999744	0.999330	0.999284	-0.917199	N
10789	0.999501	0.999441	0.999041	0.998981	-0.492163	N
10790	0.999365	0.999457	0.999211	0.999303	-1.043727	N
10791	0.999520	0.999655	0.999366	0.999502	-1.019158	N
10792	0.999703	0.999908	0.998936	0.999141	-0.917199	?

In [3]:
```
raw.describe()
```

Out[3]:

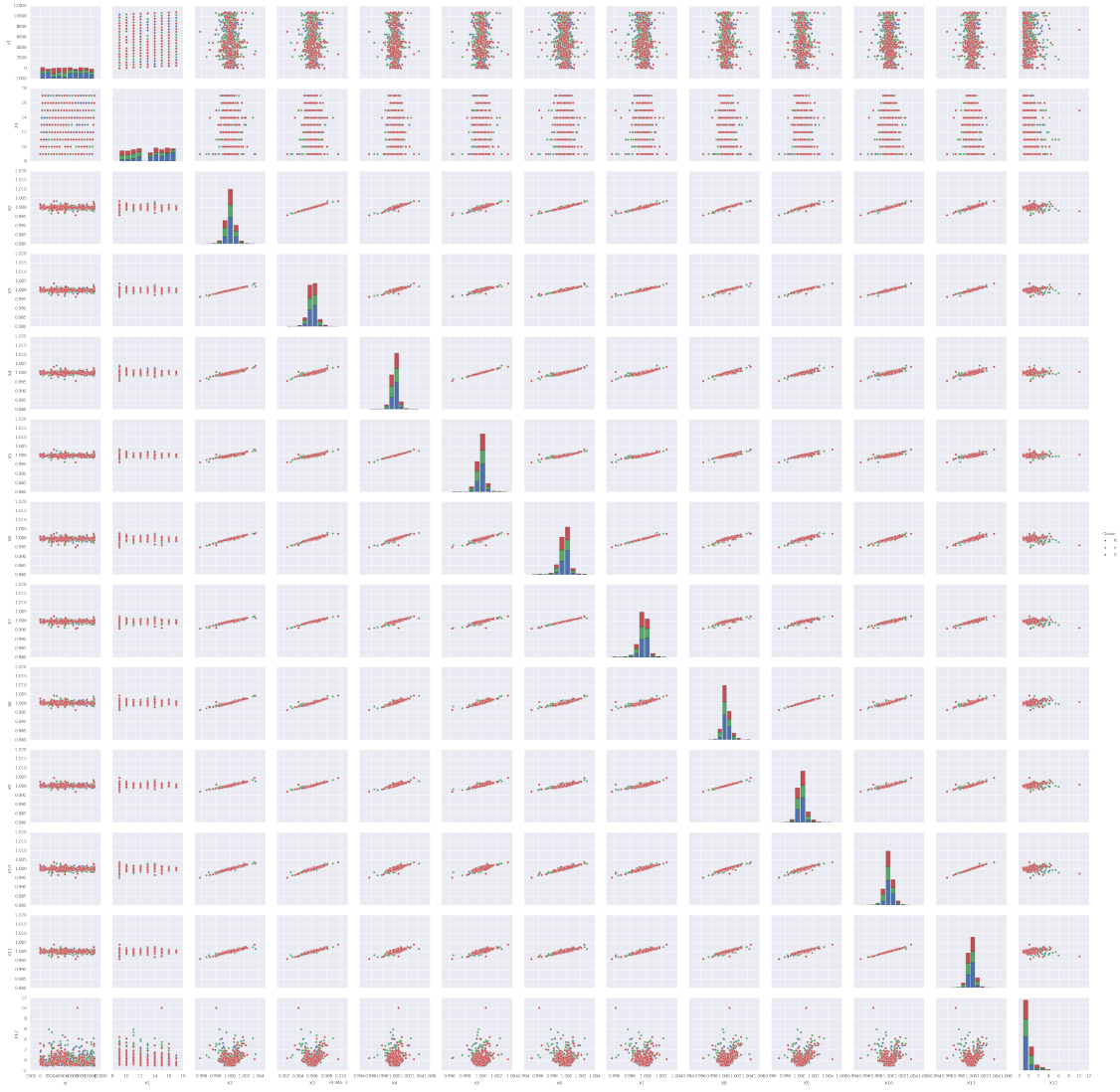
	id	X1	X2	X3	X4
count	10793.000000	10793.000000	10793.000000	10793.000000	10793.000000
mean	5406.000000	13.004262	1.000001	1.000001	0.999999
std	3115.815062	2.581256	0.000739	0.000653	0.000663
min	10.000000	9.000000	0.993092	0.993643	0.993583
25%	2708.000000	11.000000	0.999619	0.999661	0.999677
50%	5406.000000	13.000000	1.000002	0.999999	1.000001
75%	8104.000000	15.000000	1.000395	1.000342	1.000338
max	10802.000000	18.000000	1.007497	1.007155	1.006971

	X5	X6	X7	X8	X9
count	10793.000000	10793.000000	10793.000000	10793.000000	10793.000000
mean	0.999999	0.999690	0.999689	1.000314	1.000314
std	0.000568	0.000740	0.000654	0.000735	0.000650
min	0.993711	0.992300	0.992941	0.994045	0.994172
25%	0.999721	0.999347	0.999392	0.999911	0.999948
50%	0.999998	0.999752	0.999750	1.000248	1.000242
75%	1.000283	1.000097	1.000051	1.000661	1.000616
max	1.006781	1.006844	1.006781	1.008614	1.008299

	X10	X11	X12
count	10793.000000	10793.000000	10793.000000
mean	0.999999	0.999999	-0.000342
std	0.000787	0.000708	1.000223
min	0.992452	0.993354	-1.062153
25%	0.999608	0.999649	-0.673970
50%	0.999999	1.000000	-0.295615
75%	1.000394	1.000356	0.351766
max	1.007755	1.007692	14.230537

In [4]:
```
train = raw[raw.Classe != '?']
test = raw[raw.Classe == '?']
```

In [5]: sns.pairplot(train.sample(frac=0.1), hue='Classe');



In [ ]: