

Memòria pràctica 8

Josep Marc Mingot//, Gabriel Reines

27 de desembre de 2012

Resum

Índex

1	Introducció	3
2	Visualització de dades	3
3	Reducció de dimensionalitat	3
4	Classificadors	3
4.1	LDA, QDA	3
4.2	k-NN	3
4.3	SVM	3
4.4	Xarxes Neuronals	3
4.5	Arbres	3
4.6	Random Forest (opcional)	3
4.7	Unió de classificadors	3
5	Conclusions	3
A	Primer Apèndix	3

1 Introducció

In this example we embed parts of the examples from the `kruskal.test` help page into a \LaTeX document:

```
> data(airquality)
```

which shows that the location parameter of the Ozone distribution varies significantly from month to month. Finally we include a boxplot of the data:

2 Visualització de dades

El primer pas en qualsevol problema de classificació és entendre les dades que se'ns han donat. Per aquest motiu en aquesta secció pretenem donar a través de l'anàlisi visual, una serie de propietats sobre les variables que influeixen en el problema.

Comencem estudiant la correlació entre elles. Al tenir 116 variables l'eina més adequada per visualitzar l'autocorrelació entre elles és visualitzar gràficament la matriu de correlació. En la figura ?? observem la matriu on cada casella representa un valor i com més fosc més proper a 1. Hi ha variables que estan completament correlacionades (per exemple `neighborhood intensity feature 2` amb `neighborhood intensity feature 8`). Per eliminar-les, fixem un nivell a partir del qual considerem que les variables estan correlacionades i n'eliminem aquella que té una mitja de correlació més alta amb les altres. Després d'aquest procés i fixant un tall a 0.85 passem de 116 variables a 88 variables. Podem observar la matriu d'autocorrelació neta en la figura ??.

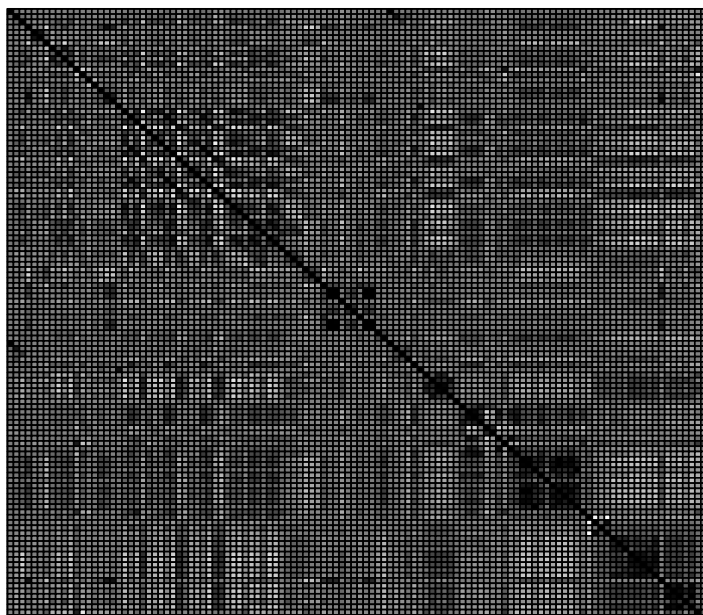


Figura 1: Autocorrelació de les 117 variables.

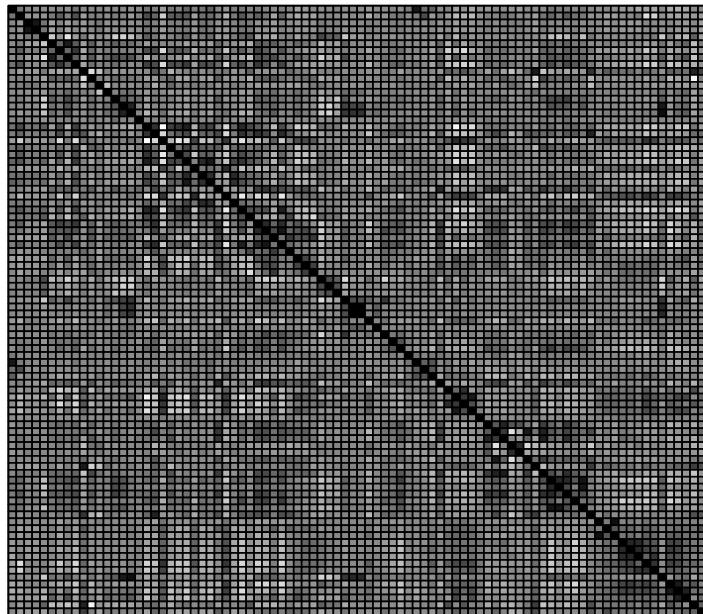


Figura 2: Autocorrelació de les 88 variables.

3 Reducció de dimensionalitat

4 Classificadors

4.1 LDA, QDA

4.2 k-NN

4.3 SVM

4.4 Xarxes Neuronals

4.5 Arbres

4.6 Random Forest (opcional)

4.7 Unió de classificadors

5 Conclusions

A Primer Apèndix

Referències

- [1] Miika Ahdesmaki and Korbinian Strimmer, *Feature selection in omics prediction problems using CAT scores and False Nondiscovery Rate Control*, The Annals of Applied Statistics, Vol. 4, 2010.