Datamining 4: La regression
Les arbres de regression

M2 STD

November 26, 2014

L'idée de base

Imaginons qu'on ait segmenté Y en k classes homogènes. On peut alors appliquer n'importe quelle méthode de classification vue au chapitre précédent pour estimer notre nouvelle variable "qualitative" (ordonnée en fait). On peut ainsi utiliser les arbres de classification (entre autres)....

L'avantage des arbres

Modulo une toute petite variante sur les arbres on peut aussi les utiliser en regression sans partir d'une segmentation initiale qui pourrait etre arbitraire !

Algorithme des arbres de regressions

Tout se passe globalement comme pour les arbres de classifications:

- On peut avoir des variables explicatives quantitatives et qualitatives
- A chaque étape (noeud) on cherche la variable qui discrimine le mieux (plutot comme dans CART: i.e. par division en 2)
- On continue jusqu'as ce qu'un certain critère soit atteint
- On "ellague" par validation
- On affecte une "valeur" dans chaque feuille

Algorithme des arbres de regressions

Les différences se trouvent dans les étapes “en rouge”.

- On peut avoir des variables explicatives quantitatives et qualitatives
- A chaque étape (noeud) on cherche la variable qui discrimine le mieux (plutot comme dans CART: i.e. par division en 2)
- On continue jusqu'as ce qu'un certain critère soit atteint
- On “ellague” par validation
- On affecte une “valeur” dans chaque feuille

Pouvez vous intuiter les méthodes choisies pour les derniers puis deuxième points ?

Arbres de regression les deux différences

- La valeur dans une feuille sera ainsi la valeur moyenne
- La variable la plus discriminante sera trouvée en effectuant un test d'égalité des moyennes (avec variance inégales)

Arbres de regression : inconvénients

- On estime f la fonction de regression comme une fonction constante par morceaux (faire des forets d'arbres permet de regler un peu ce probleme)
- Attention aux interprétations intempestives (lorsque des variables explicatives sont très corrélées)

Exemple

Taux de criminalité a boston en fonction de 13 autres variables

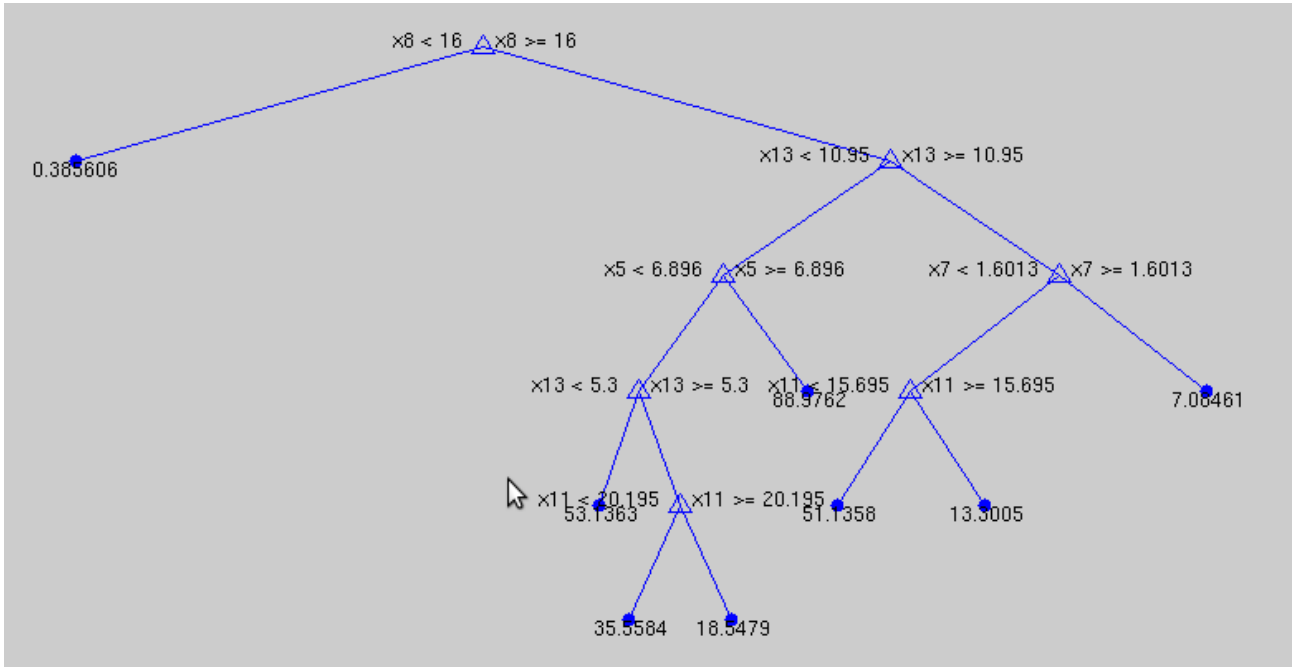

Figure: La sortie matlab

Exemple

Taux de criminalité a boston en fonction de 13 autres variables

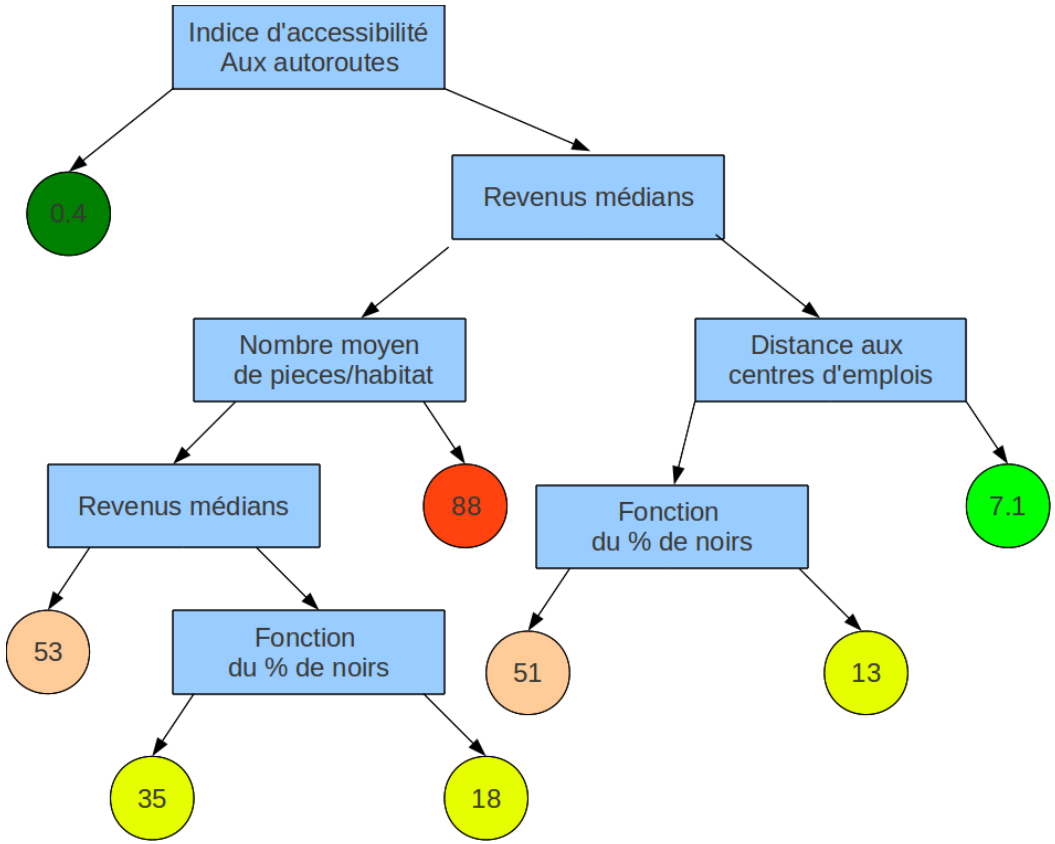

Figure: apres travail

Comparaison avec une regression linéaire

Du point de vue des variable selectionnées

```
Variables explicatives : en vert et oranges les significativement positives et
négatives dans une régression linéaire ; en gras celles qui jouent dans l'arbre

1. ZN          proportion of residential land zoned for lots over 25,000 sq.ft.
2. INDUS       proportion of non-retail business acres per town
3. CHAS        Charles River dummy variable (= 1 if tract bounds river; 0
otherwise)
4. NOX         nitric oxides concentration (parts per 10 million)
5. RM          average number of rooms per dwelling
6. AGE         proportion of owner-occupied units built prior to 1940
7. DIS         weighted distances to five Boston employment centres
8. RAD         index of accessibility to radial highways
9. TAX       full-value property-tax rate per $10,000
10. PTRATIO    pupil-teacher ratio by town
11. B          1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town)
12. LSTAT      % lower status of the population
13. MEDV       Median value of owner-occupied homes in $1000's
%
```

Figure: apres travail