Compte Rendu - Apprentissage et Reconnaissance de Formes

Michael Trazzi, Julien Denes

Mars 2018

TME 1 - Arbres de décision, sélection de modèles

Calcul de l'entropie

```
def entropie(vect):
    _, counts = np.unique(vect, return_counts=True)
    p_y = np.array(counts / len(vect))
    return (-np.sum(p_y * np.log(p_y)))
```

```
def entropie_cond(list_vect):
    n = len(list_vect)
    total_nb = sum(len(part) for part in list_vect)
    p = np.array((1,n))
    H = np.array((1, n))
    for i in range(n):
            p[i] = len(list_vect[i]) / total_nb
            H[i] = entropie(list_vect[i])
    return (np.sum(H * p))
```

Quelques expériences préliminaires

Q 1.4 Sur les données IMDB, nous avons testé des profondeurs d'arbres allant de 5 à 50. Plus la profondeur est grande, et plus on surapprend notre base de données d'apprentissage.
Q 1.5 Pour une profondeur de 5 (resp. 50), on obtient un score de 0.736429038587 (resp. 0.900152605189). On a bien un surapprentissage de nos données dans le cas de la profondeur de 50.
Q 1.6 Le score ainsi défini n'est pas un indicateur fiable : il indique uniquement notre capacité a surapprendre la base d'apprentissage, mais ne tient pas compte du pouvoir de généralisation.

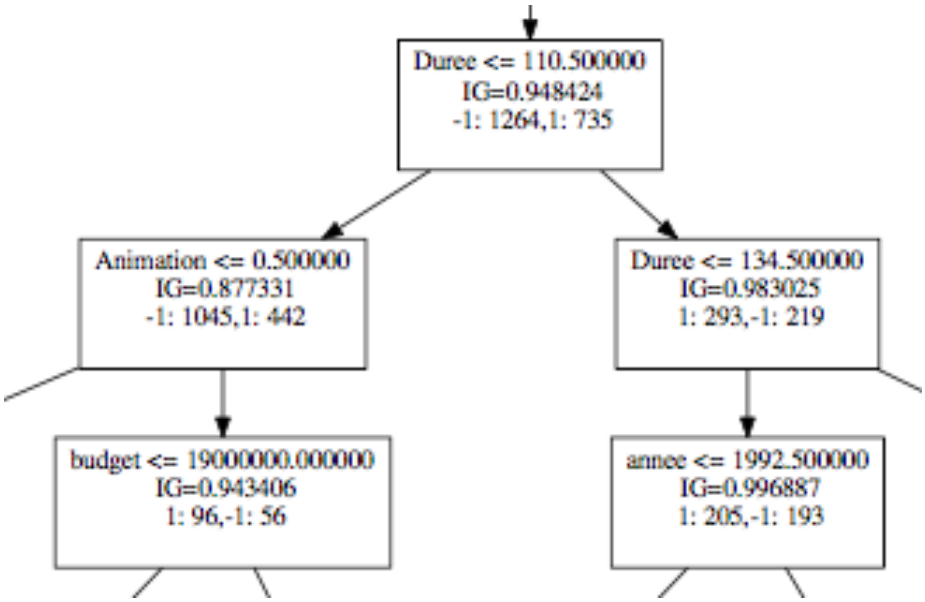

Figure 1: une partie de ce qu'on obtient avec profondeur 5

Sur et sous apprentissage

Q 1.7 (cf. Figures 2 et 3 pour les courbes)
Q 1.8 Avec peu d'exemples d'apprentissage, le score se stabilise tres rapidement en fonction de la profondeur. Avoir un arbre tres profond avec seulement 1000 exemples ne fait pas varier le score. A contrario, avec l'ensemble de la base d'apprentissage, le score continue a varier jusqu'a une prondeur de 25.
Q 1.9 Mes resultats ne me semblent pas fiables (seulement 5 points pour chaque courbe, base d'apprentissage de seulement quelques milliers d'exemples). Pour les ameliorer il faudrait prendre plus d'exemples (10 000 000), plus de répartitions (au moins 5) et plus de points (100 profondeurs différentes).

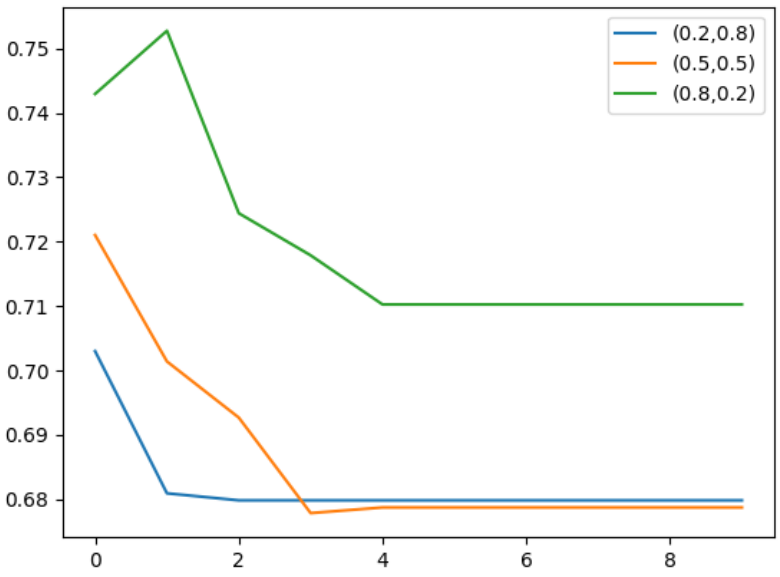

Figure 2: Evolution du score en fonction de (profondeur-1)/5 avec toute la base d'apprentissage

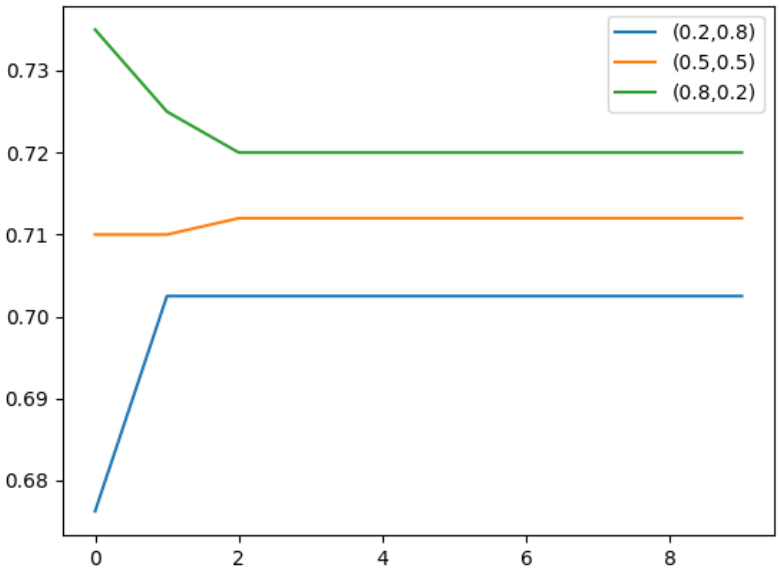

Figure 3: Idem avec peu d'exemples (1000 au lieu de 4000)

TME 2 - Estimation de densité - Expérimentations

Méthode des histogrammes

Méthode a noyaux

Différence entre faible et forte discrétisation

Role des parametres des méthodes a noyaux

Choix automatique des meilleurs parametres

Estimation de la qualité du modele

TME 3 - Descente de gradient

Optimisation de fonctions

Régression logistique

TME 4 - Perceptron

Implémentation

Données USPS

Données 2D et projection