DATA SCIENCE AND MINING

Lab & Assignment : Supervised Learning Linear Discriminant Analysis and Logistic Regression

22 mars 2015

AKOUEMO FEUJIO eric Frank

TABLE DES MATIÈRES

1 IMPLÉMENTATION DU CODE

L'implémentation du code est dans l'archive jointe à ce document.

Nous avons fait l'effort de le commenté le plus possible et surtout de décrire les étapes ou grandes lignes du code .

2 RÉSULTATS OBTENUES

2.1 The ROC curve

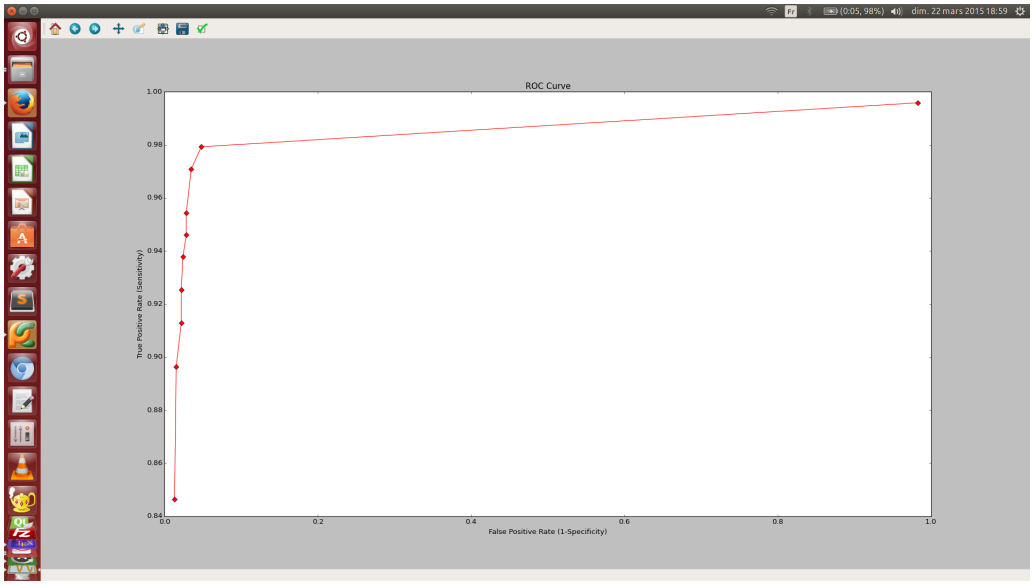

Figure 1 : The ROC curve

2.2 Choix du seuil basé sur la courbe tracée

Which threshold would you choose ?

D'après le tracé de la courbe on peut remarqué que le seuil le plus avantageux est 0.7 ou 0.8 mais avec une préférence pour 0.8 car il offre un meilleur de vrai positif pour un taux de faux positif qui reste largement acceptable.

Is the default one the optimal one ?

Non ce n' est pas le seuil par défaut(à savoir 0.5) qui est le plus avantageux comme on a pu le constater.

2.3 Les paramètres θ et notre interprétation de leur signification

Which parameter has which theta ? what does that mean ?

Le fichier des données contient plusieurs entrées et le vecteur correspond à :

θ		1	2	3	4	5	6	7	8	9	
Ligne du fichier `breast_cancer.txt`	1e+06	5	1	1	1	2	1	3	1	1	0

NB : il faut noter que le vecteur θ commence à l'indice 0

What if you include the id ?How does that change your model ?

si on inclut Id dans ce cas il faudra ajouter une dimension dans le vecteur theta pour qu'il représente ce paramètre du fichier .

Cela ne change pas le modèle , mais pourrait influencer les résultats obtenues.

What happens if you take the top 5 of theta in absolute magnitude and use only those (and the corresponding features) ?

Si on se limite au 5 premiers paramètre du vecteur theta alors on ferra une prédiction qui ne sera représentative que de ces paramètres.Les résultats obtenus seraient **indépendants** des autres paramètres (à savoir les 4 derniers)