Compte rendu du TP2

Membre: Hugo 16241007 Arnaud 16241047

Dans cette séance, on compare les performances des 3 discriminateurs en observant leur taux de réusitte et leur barre d'erreur pour des datas dont distributions suivent la loi normal.

D'après ce qu'on a pris en cours, le discriminateur $\widehat{d_{\mu,\Gamma}}$ avec les deux paramètres connus est l'optimal, car il présente le vrai cas. Et le discriminateur quadratique utilise les coniques pour classifier les objets. C'est la forme générale de celui de linéaire. Le discriminateur linéaire utilise le maximum de vraisemblance pour décider les paramètres. Mais seulement dans le cas où on a la distribution normal et de l'égalité de covariance, ce discriminateur est performant.

Pour la première partie, on génere des base de donnés en deux dimension et on veut les classifier aux 2 classes. Et on a observé 3 cas differents : $\mu_1 \neq \mu_2$, $\Gamma_1 = \Gamma_2$; $\mu_1 = \mu_2$, $\Gamma_1 \neq \Gamma_2$; $\mu_1 \neq \mu_2$, $\Gamma_1 \neq \Gamma_2$ pour mieux comprendre comment ils se distribuent et pour nous faire réfléchir comment on peut les classifer. Dans 1.4/ on est dans le cas où $\mu_1 \neq \mu_2$, $\Gamma_1 \neq \Gamma_2$ donc les deux distributions n'ont pas le lien. Dans ce cas là on fait ce qui ressemble a la Figure 10 dans le CM6, on peut voir que dans la Figure de < $Max_{1,2}$ des ddp> ,le frontière est où la densité de probabilité des deux distribution sont égaux et ça vérifie ce qu'on appris dans le cours sur la Figure 10(un minimum de proba d'erreur).

Pour la deuxième partie, on essaie de comparer les 3 discriminateur pour les distributions en deux dimension aux 3 classes dans la base de generalisation.

Premièrement pour le discriminateur avec μ et Γ connus(D1), on utilise la formule $\sigma_{\tau_g} = \sqrt{\frac{\mu_{\tau_g}(1-\mu_{\tau_g})}{Pgen}}$ pour estimer la précision du discriminateur. Et puis on ajoute un autre discriminateur linéaire(D2) et on trouve que en comparaison avec D1, le tau_g de D2 est plus petit et σ_{τ_g} est plus grand. Ça represente une perte de performance signative. Et pour moi c'est raisonable, parce que pour 2 distribution arbitraires en 2 dimension aux 2 classes, c'est très difficile de les classifier par des lignes droites. Puis on rajoute un autre discriminateur quadratique(D3), et cette fois ci , on observe non seulement le tau_g et σ_{τ_g} mais aussi la figure qui montre comment ils organisent leurs frontières. Et la performance est D1>D3>D2. C'est aussi raisonnable parce que par leurs definitions, D1 connait la plus d'information, D3 estime tous les Γ mais D2 suppose que les Γ sont égaux.

Puis on passe au cas où les 3 distributions de base avec leur frontières sont presque en parallèle(exemple 2). Et on peut voir que il n'y a pas de différences de performance signatives. Dans le cours CM6 on voit que quand $\Gamma_1 = \Gamma_2$ alor D3 va faire une frontière de forme de droite. Donc dans ce cas là, les 3 discriminateur sont même compétents. <qui peut le plus peut le moin> maintenant signifie que D3 est plus compétents dans le cas plus complex donc il est aussi compétent pour le cas simple.

Finalement quand on passe au cas où la dimension de l'espace est 38. On voit que maintenant une phénomene différente : le tau_g pour D2 est le plus petit parmi les autre ! Pour D1 et D2, on a tau_g=1 et $\sigma_{\tau_g} = 0$. Je pense que c'est raisonable parce que on va seulment classifier les base de donnés en 3 classes, mais ils ont 38 dimensions. On a l'impression que si la distribution est un peu sérrée, et c'est difficile de les classifier sur une surface , alors on pass à la dimension 3, et on peut trouver des surfaces qui casse le cube qui pourrait bien classifier les donnés. Donc si on pass a la dimension 38, c'est plus facile de trouver des frontière, ce qui peut expliquer la performance du discriminateur linéaire. Mais ce que je ne comprend pas c'est la perte de performance de D3 qui doit toujours d'être mieux que D2. Pour moi je trouve c'est bizzare.