Analyse discriminante

Véronique Tremblay

L'analyse discriminante

Date du début du siècle:

Fisher, Ronald A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics 7 (2). Wiley Online Library: 179–88.

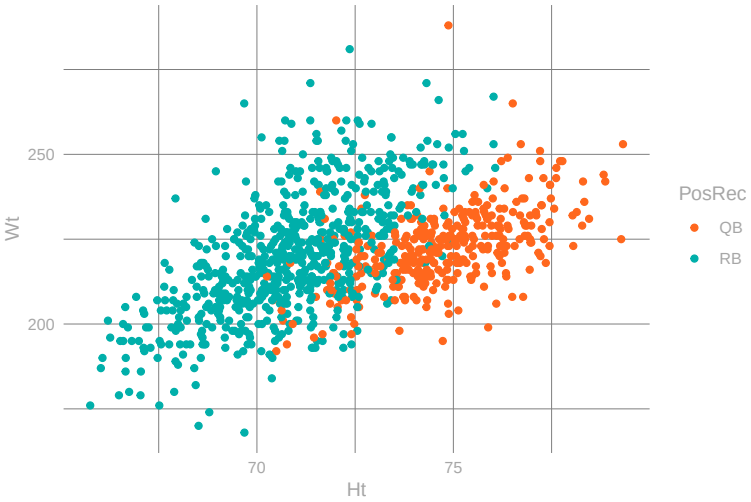
Soutien à la section 4.3 de Hastie, Tibshirani, and Friedman (2009)

L'analyse discriminante

Premet de faire de bonnes prédictions lorsque

- Y est nominale
- X est composé de variables normales multidimensionnelles (de même variance)

Qu'est-ce qu'on veut faire?



Décomposition de la variance

De vos cours d'analyse de la variance, vous savez (ou vous saurez) que

$$T = B + W$$

où

- T est la variabilité totale
- B est la variabilité intergroupe (*between*)
- W est la variabilité intragroupe (*within*)

Dans l'exemple

Variance entre les groupes (B)

498.2045	60.633902
60.6339	7.379439

Variance intra-groupe (W)

507.7955	432.5774
432.5774	998.6206

Variance totale (T)

1006.0000	493.2113
493.2113	1006.0000

Posons $\tilde{\mathbf{X}}$, le vecteur des variables aléatoires centrées et réduites.

Le score proposé par Fisher est une **combinaison linéaire des variables**, c'est-à-dire

$$Z = f(\tilde{X}_1, \dots, \tilde{X}_p) = \alpha^\top \tilde{\mathbf{X}}$$
$$Z = \alpha_1 \tilde{X}_1 + \cdots + \alpha_p \tilde{X}_p$$

$$Z = \alpha^\top \tilde{X}$$

On choisit α de façon à maximiser

$$\frac{\alpha^\top \mathbf{B}\alpha}{\alpha^\top \mathbf{W}\alpha} \quad \text{ou} \quad \frac{\alpha^\top \mathbf{B}\alpha}{\alpha^\top \mathbf{T}\alpha}$$

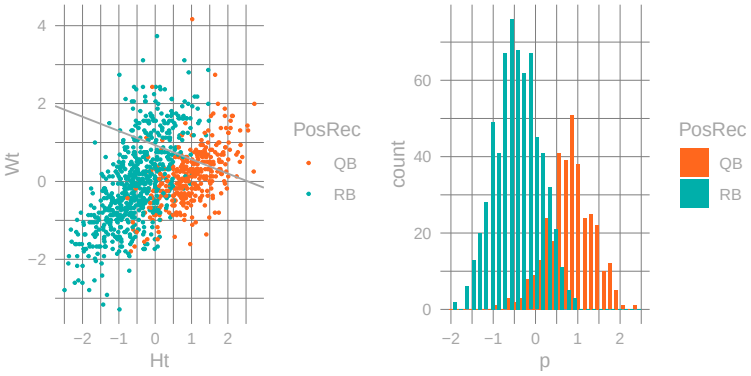
Ce qui revient à maximiser $\alpha^\top \mathbf{B}\alpha$ sous la contrainte que $\alpha^\top \mathbf{T}\alpha = 1$.

Du chapitre sur l'ACP, α est le vecteur propre normé associé à la plus grande valeur propre de $\mathbf{T}^{-1}\mathbf{B}$

```r
vp <- eigen(solve(S) %*% B)
vp
```

```
## eigen() decomposition
## $values
## [1]  5.837920e-01 -3.469447e-18
##
## $vectors
##            [,1]       [,2]
## [1,]  0.9310380 -0.1208134
## [2,] -0.3649222  0.9926752
```

Visualisation et règle de classification



	Moyenne de p	Écart-type de p
QB	0.8544327	0.5281928
RB	-0.4551772	0.5265065

Théorème de Bayes

$$\mathbb{P}(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

Proportion des joueurs à chaque position

QB	RB
0.347567	0.652433

Densités estimées pour un joueur

dQb	dRb
0.2424403	0.471075

En R

```
library(MASS)
model_lda <- lda(PosRec ~ Ht+Wt,
                 data = don_scale)
prediction_lda <- predict(model_lda)
```

Résumé

Références

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.