Apprentissage en grande dimension (Partie #4)

Thomas Verdebout

Université de Lille

Contenu du cours

1. Introduction.
2. Apprentissage supervisé: régression.
3. Apprentissage supervisé: classification.
4. Apprentissage non-supervisé

2: Apprentissage supervisé: analyse discriminante.

L'analyse discriminante et la classification ont pour objectifs de séparer en différents groupes des objets appartenant à diverses populations (par la construction de "discriminants", c'est-à-dire de quantités numériques qui différeront autant que possible d'une population à l'autre), et de définir des règles de classification qui permettront de "prédire" à quelle population appartient un nouvel objet.

Remarque: les procédures que nous allons construire (qui tentent bien entendu de minimiser les cas de misclassification) devront aussi prendre en compte les probabilités à priori qu'une nouvelle observation à classifier provienne de la population i ($i = 1, \ldots, m$), et des coûts de misclassification, qui peuvent ne pas être symétriques (exemple).

Considérons deux populations π_1 et π_2, qui sont associées à des lois de probabilité (sur $\mathbb{R}^p$) absolument continues de densité f_1 et f_2, respectivement.

Ces deux populations diffèrent par leur position, leur dispersion, ou toute autre caractéristique. Le problème que nous considérons est le suivant:

sur base de la valeur $\mathbf{x} = (x_1, \ldots, x_p)'$ prise par un p-v.a. $\mathbf{X} = (X_1, \ldots, X_p)'$ provenant de π_1 ou de π_2 (i.e., $\mathbf{X} \sim f_1$ ou $\mathbf{X} \sim f_2$), comment parier de manière raisonnable sur la population dont X est issu?

$\rightsquigarrow$ Une règle de classification consiste à donner une partition (R_1, R_2) de l'ensemble χ des valeurs possibles pour x, telle que

- $\mathbf{X}$ sera classifié en π_1 si $x \in R_1$ et
- $\mathbf{X}$ sera classifié en π_2 si $x \in R_2$.

Si on se donne des probabilités à priori p_1, p_2 que $\mathbf{X}$ provienne respectivement de π_1 et π_2, la probabilité P de misclassification de $\mathbf{X}$ est donnée par

$$P = p_{2|1} \times p_1 + p_{1|2} \times p_2,$$

où

$$p_{i|j} = P[\mathbf{X} \in R_i | \mathbf{X} \in \pi_j] = \int_{R_i} f_j(\mathbf{x})\, d\mathbf{x}.$$

Une procédure de classification (R_1, R_2) optimale veillera à minimiser P.

Si en outre différents coûts de misclassification doivent être pris en compte (notons $c_{1|2}$ et $c_{2|1}$ respectivement les coûts si on classifie en π_1 un objet provenant de π_2 et si on classifie en π_2 un objet provenant de π_1), une procédure de classification (R_1, R_2) optimale devra cette fois minimiser le coût de misclassification moyen

$$\begin{aligned} E_C &= E[\text{coût}|\mathbf{X} \in \pi_1]\, p_1 + \mathrm{E}[\text{coût}|\mathbf{X} \in \pi_2]\, p_2 \\ &= \left(0 \times p_{1|1} + c_{2|1}\, p_{2|1}\right) p_1 + \left(c_{1|2}\, p_{1|2} + 0 \times p_{2|2}\right) p_2 \\ &= c_{2|1}\, p_{2|1}\, p_1 + c_{1|2}\, p_{1|2}\, p_2. \end{aligned}$$

Le résultat suivant décrit la procédure de classification (R_1, R_2) optimale dans ce contexte général:

Théorème: *la procédure de classification (R_1, R_2) qui minimise le coût de misclassification moyen est donnée par*

$$R_1 = \left\{ \mathbf{x} \in \chi \,\middle|\, \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c_{1|2}\, p_2}{c_{2|1}\, p_1} \right\} \quad et \quad R_2 = \chi \backslash R_1 .$$

Preuve. Le coût moyen E_C sera minimal si l'intégrande de

$$\begin{aligned} E_C &= c_{2|1}\, p_{2|1}\, p_1 + c_{1|2}\, p_{1|2}\, p_2 \\ &= c_{2|1}\, (1 - p_{1|1})\, p_1 + c_{1|2}\, p_{1|2}\, p_2 \\ &= c_{2|1}\, p_1 + (c_{1|2}\, p_{1|2}\, p_2 - c_{2|1}\, p_{1|1}\, p_1) \\ &= c_{2|1}\, p_1 + \int_{R_1} (c_{1|2}\, f_2(\mathbf{x})\, p_2 - c_{2|1}\, f_1(\mathbf{x})\, p_1)\, d\mathbf{x} \end{aligned}$$

est négative pour tout $x \in R_1$. □

Remarque. Il suffit de connaître les rapports $\frac{f_1}{f_2}$, $\frac{c_{1|2}}{c_{2|1}}$ et $\frac{p_2}{p_1}$ pour déterminer la règle de classification optimale.

Quelques cas particuliers:

- Si les probabilités p_1, p_2 sont égales ou sont inconnues,

$$R_1 = \left\{ x \in \chi \,\middle|\, \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c_{1|2}}{c_{2|1}} \right\} \quad \text{et} \quad R_2 = \chi \backslash R_1.$$

- Si les coûts $c_{1|2}$, $c_{2|1}$ sont égaux ou non spécifiés,

$$R_1 = \left\{ x \in \chi \,\middle|\, \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \right\} \quad \text{et} \quad R_2 = \chi \backslash R_1.$$

- Si p_1, p_2, $c_{1|2}$, $c_{2|1}$ sont non spécifiés,

$$R_1 = \left\{ x \in \chi \,\middle|\, \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \right\} \quad \text{et} \quad R_2 = \chi \backslash R_1.$$

Considérons d'abord le cas où $\pi_i = \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, avec $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 (=: \boldsymbol{\Sigma})$.

$\rightsquigarrow$ **Proposition**: soit $\mathbf{a} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Alors la procédure de classification optimale classifie $\mathbf{x}$ en π_1 si

$$\mathbf{a}'\mathbf{x} \geq \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right]$$

et en π_2 sinon.

Preuve. □

Remarque. Si $\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1} = 1$, il faut donc classifier $\mathbf{x}$ en π_1 ssi

$$\mathbf{a}'\mathbf{x} \geq \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

On voit facilement par la définition de a que

$$\mathbf{a}'\boldsymbol{\mu}_1 \geq \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \quad \text{et} \quad \mathbf{a}'\boldsymbol{\mu}_2 \leq \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Il faut donc classifier $\mathbf{x}$ en π_1 si la projection de $\mathbf{x}$ sur l'axe a est plus proche de celle de $\boldsymbol{\mu}_1$ que de celle de $\boldsymbol{\mu}_2$.

De manière équivalente, il faut classifier $\mathbf{x}$ en π_1 ssi

$$2\left[\mathbf{a}'\mathbf{x} - \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right] = d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_2) - d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_1) \geq 0.$$

Dans le cas général, la règle de classification compare donc les projections de $\mathbf{x}$, de $\boldsymbol{\mu}_1$ et de $\boldsymbol{\mu}_2$ sur l'axe a et classifie $\mathbf{x}$ en π_1, en fonction des coûts et des probabilités à priori, si

$$2\left[\mathbf{a}'\mathbf{x} - \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right] = d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_2) - d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_1) \geq 2 \ln \left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right].$$

La fonction $\mathbf{x} \mapsto \mathbf{a}'\mathbf{x} - \frac{1}{2}\mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right]$ sur laquelle est fondée la règle de classification est appelée fonction discriminante linéaire de Fisher . On parlera *d'analyse discriminante linéaire.*

En pratique, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ et $\boldsymbol{\Sigma}$ sont inconnus et il faut les estimer sur base d'un échantillon ("training sample") d'observations indépendantes

$$\mathbf{X}_1, \ldots, \mathbf{X}_{m_1} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad \mathbf{Y}_1, \ldots, \mathbf{Y}_{m_2} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$$

En utilisant les notations dans la Section 3.1.4, la règle empirique consiste alors à classifier $\mathbf{x}$ en π_1 plutôt qu'en π_2 ssi

$$(\bar{\mathbf{X}} - \bar{\mathbf{Y}})'\mathbf{S}_{\text{pool}}^{-1}\,\mathbf{x} \geq \frac{1}{2}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})'\mathbf{S}_{\text{pool}}^{-1}(\bar{\mathbf{X}} + \bar{\mathbf{Y}}) + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right].$$

Si on considére plutôt le cas où $\pi_i = \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$, sans faire l'hypothèse d'égalité des covariances, on obtient alors (en procédant comme ci-dessus) le résultat suivant (exercice):

$\rightsquigarrow$ **Proposition**: soit

$$k := \ln\left[\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right] + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2).$$

Alors la procédure de classification optimale classifie $\mathbf{x}$ en π_1 si

$$-\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x} + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\mathbf{x} \geq \frac{k}{2} + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right]$$

et en π_2 sinon.

Remarque. Pour des raisons évidentes, on parlera cette fois de règle de classification (et donc d'analyse discriminante) quadratique

.

Si on dispose d'un échantillon d'observations indépendantes

$$\mathbf{X}_1, \ldots, \mathbf{X}_{m_1} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad \mathbf{Y}_1, \ldots, \mathbf{Y}_{m_2} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),$$

la règle empirique classifie dans ce cas $\mathbf{x}$ en π_1 ssi

$$-\frac{1}{2}\mathbf{x}'(\mathbf{S}_x^{-1} - \mathbf{S}_y^{-1})\mathbf{x} + (\bar{X}'\mathbf{S}_x^{-1} - \bar{Y}'\mathbf{S}_y^{-1})\mathbf{x} \geq \frac{\hat{k}}{2} + \ln\left[\frac{c_{1|2}\, p_2}{c_{2|1}\, p_1}\right],$$

où

$$\hat{k} := \ln\left[\frac{|\mathbf{S}_x|}{|\mathbf{S}_y|}\right] + (\bar{\mathbf{X}}'\mathbf{S}_x^{-1}\bar{\mathbf{X}} - \bar{\mathbf{Y}}'\mathbf{S}_y^{-1}\bar{\mathbf{Y}}).$$

Nous considérons le cas de m populations π_i ($i = 1, \ldots, m$) , associées à des lois de probabilité (sur $\mathbb{R}^p$) absolument continues de densité f_i ($i = 1, \ldots, m$).

Nous cherchons à déterminer une règle de classification qui vise à parier sur la population dont est issue la réalisation **x** d'un p-v.a. **X** (que l'on suppose provenir d'une des π_i).

Comme dans le cas à deux populations, une telle règle est complétement déterminée par une partition

$$(R_i,\ i = 1, \ldots, m)$$

de l'ensemble χ des valeurs possibles pour **x**, et consistera à classifier **X** en π_i ssi $\mathbf{x} \in R_i$.

Si on note

- $c_{i|j}$ le coût de classification en π_i d'un objet de π_j,
- p_j la probabilité à priori que $\mathbf{X}$ provienne de π_j et
- $p_{i|j} := P[\mathbf{X} \in R_i | \mathbf{X} \in \pi_j] = \int_{R_i} f_j(\mathbf{x})\, d\mathbf{x}$ la probabilité conditionnelle qu'un objet soit classifié en π_i sachant qu'il provient de π_j,

une procédure de classification $(R_i,\ i = 1, \ldots, m)$ sera dite optimale si elle minimise le coût de misclassification moyen, qui s'écrit ici (avec $c_{j|j} = 0$)

$$E_C = \sum_{j=1}^{m} \mathrm{E}[\text{coût} | \mathbf{X} \in \pi_j]\, p_j = \sum_{j=1}^{m} \left[\sum_{i=1}^{m} c_{i|j}\, p_{i|j} \right] p_j.$$

On peut alors montrer le résultat suivant:

Théorème: la procédure de classification qui minimise le coût de misclassification moyen consiste à classifier $\mathbf{x}$ en la population π_i pour laquelle

$$h_i(\mathbf{x}) := \sum_{j=1}^{m} c_{i|j}\, p_j\, f_j(\mathbf{x})$$

est minimal.

Remarques:

- Ceci étend bien le résultat vu pour $m = 2$.
- Si les coûts $c_{i|j}$ sont égaux ou non spécifiés, la règle optimale classifie $\mathbf{x}$ en la population π_i telle que $p_i f_i(\mathbf{x}) = \max_j\{p_j f_j(\mathbf{x})\}$.

Preuve.

$$
\begin{aligned}
E_C &= \sum_{j=1}^{m} \left[\sum_{i=1}^{m} c_{i|j}\, p_{i|j} \right] p_j \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} c_{i|j}\, p_j \int_{R_i} f_j(\mathbf{x}) d\mathbf{x} \\
&= \sum_{i=1}^{m} \int_{R_i} h_i(\mathbf{x}) d\mathbf{x}.
\end{aligned}
$$

□

Considérons le cas où $\pi_i = \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, m$, sans inclure de coûts de misclassification.

$\rightsquigarrow$ **Proposition**:

$$d_j(\mathbf{x}) := -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln p_j.$$

Alors la procédure de classification optimale classifie x en π_i ssi $d_i(\mathbf{x}) = \max_j \{d_j(\mathbf{x})\}$.

Preuve. Le théorème affirme qu'il est optimal de classifier $\mathbf{x}$ en π_i ssi $\ln(p_i f_i(\mathbf{x})) = \max_j \{\ln(p_j f_j(\mathbf{x}))\}$, où

$$\ln(p_j f_j(\mathbf{x})) = \ln p_j - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j),$$

ce qui établit le résultat. $\square$

En pratique, on classifiera $\mathbf{x}$ en π_i ssi

$$\hat{d}_i(\mathbf{x}) = \max_j \{\hat{d}_j(\mathbf{x})\},$$

où

$$\hat{d}_j(\mathbf{x}) := -\frac{1}{2} \ln |\mathbf{S}_j| - \frac{1}{2}(\mathbf{x} - \bar{\mathbf{X}}_j)' \mathbf{S}_j^{-1} (\mathbf{x} - \bar{\mathbf{X}}_j) + \ln p_j.$$

Bien entendu, $\bar{\mathbf{X}}_j$ et $\mathbf{S}_j$ désignent ici les estimateurs non biaisés usuels de $\boldsymbol{\mu}_i$ et $\boldsymbol{\Sigma}_i$ calculés à partir de m échantillons indépendants

$$(\mathbf{X}_{j1}, \ldots, \mathbf{X}_{j,n_j}),$$

où $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{j,n_j}$ sont i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$.

Dans le cas particulier où $\mathbf{\Sigma}_i = \mathbf{\Sigma}$ pour tout $i = 1, \ldots, m$, la règle de classification revient à classifier $\mathbf{x}$ en π_i ssi $d_i(\mathbf{x}) = \max_j\{d_j(\mathbf{x})\}$, où

$$d_j(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_j'\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln p_j,$$

ou de manière équivalente, à classifier $\mathbf{x}$ en π_i ssi
$d_i(\mathbf{x}) = \max_j\{d_j(\mathbf{x})\}$, où

$$d_j(\mathbf{x}) := \boldsymbol{\mu}_j'\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln p_j.$$

Remarque. $d_j(\mathbf{x})$ peut être estimé en remplaçant $\boldsymbol{\mu}_j$ par $\bar{\mathbf{X}}_j$ et $\mathbf{\Sigma}$ par

$$\mathbf{S}_{\text{pool}} := \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \ldots + (n_m - 1)\mathbf{S}_m}{n_1 + n_2 + \ldots + n_m - m}.$$

Au niveau population, cette règle de classification revient encore à classifier $\mathbf{x}$ en π_i ssi $d_i(\mathbf{x}) = \max_j\{d_j(\mathbf{x})\}$, où

$$d_j(x) := -\frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mu_j'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln p_j$$

$$= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) + \ln p_j = -\frac{1}{2}d^2_{\Sigma}(\mathbf{x}, \boldsymbol{\mu}_j) + \ln p_j$$

(intuition dans le cas $p_j = 1/m$).

Pour l'analogue empirique, ces nouveaux $d_j(x)$ seront bien entendu estimés par

$$\hat{d}_j(\mathbf{x}) := -\frac{1}{2}d^2_{\mathbf{S}_{\text{pool}}}(\mathbf{x}, \bar{\mathbf{X}}_j) + \ln p_j.$$

Classer le saumon aléoute et canadien. Les pêcheurs de commerce aléoutes n'ont pas le droit d'attraper trop de saumons canadiens, et vice-versa. Comment faire pour classer les saumons afin de sortir indemne de cette situation peu heureuse?

Les poissons ont un cycle de vie fort intéressant. Ils naissent dans des courants d'eau fraîche, nagent jusqu'à l'océan et après quelques années ils retournent à leur lieu de naissance afin d'y mourir paisiblement.

Comme ils sont récoltés alors qu'ils sont dans l'océan, on ne peut pas *a priori* décider à quel groupe ils appartiennent.

Cependant, ils jouissent d'anneaux de croissance qui sont notoirement connus pour être plus grands pour les poissons canadiens lors de leur croissance en eau fraîche.

Alaskan		Canadien	
frais	marine	frais	marine
108	368	129	420
131	355	148	371
105	469	179	407
86	506	152	381
...			
...			
94	491	153	352
87	480	108	339

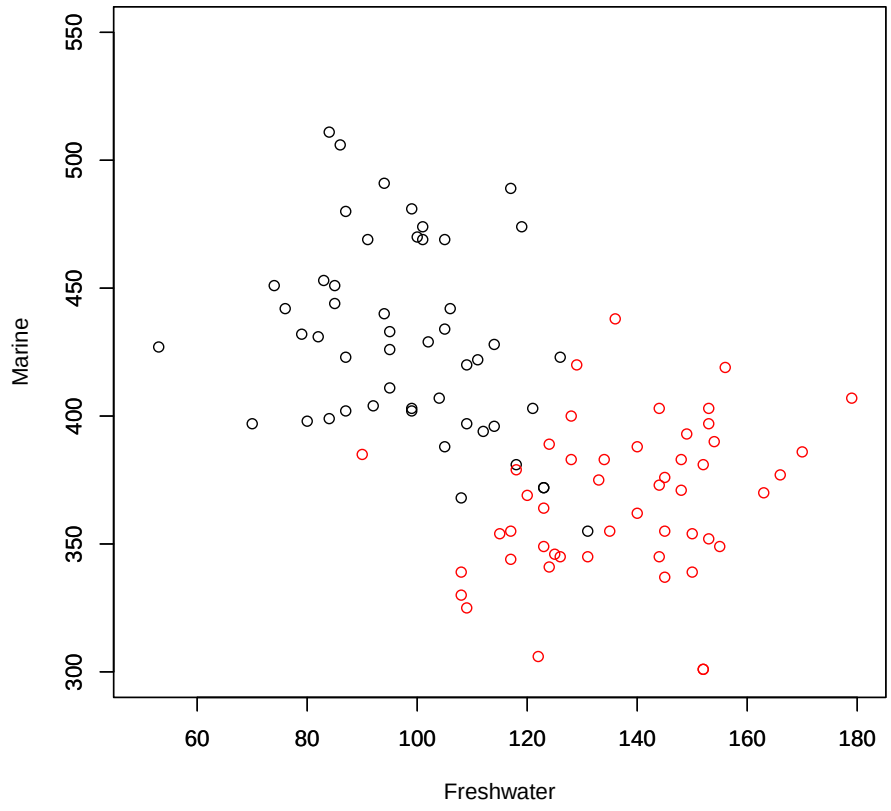

Figure: Alaskan fish (black) and Canadian fish (red) growthring diameters for a sample of 50.

Nous avons pris les 30 premières données afin d'estimer $\mathbf{S}_x$, $\mathbf{S}_y$ et $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$. Alors la règle de discrimination quadratique est appliquée aux 20 observations restantes. Dans ce cas particulier, nous n'avons pas fait de misclassification.

```
# estimate mean and covariance from the first 30 observations
Sx=var(sal[1:30,1:2])
Sy=var(sal[1:30,3:4])
k=log(det(Sx))-log(det(Sy))
+t(mx)%*%solve(Sx)%*%mx-t(my)%*%solve(Sy)%*%my
mx=c(mean(sal[1:30,1]),mean(sal[1:30,2]))
my=c(mean(sal[1:30,3]),mean(sal[1:30,4]))

# define discrimiant function
disc<-function(x){
if(-1/2*t(x)%*%(solve(Sx)-solve(Sy))%*%x
+(t(mx)%*%solve(Sx)-t(my)%*%solve(Sy))%*%x-k/2>0)  1 else 0}
```

Afin d'obtenir une meilleure estimation de la qualité de notre procédure de classification, nous pouvons utiliser l'approche suivante connue sous le nom de *leave-one-out*.

- Prendre toutes les observations sauf une pour estimer le modèle (p.ex. $\mathbf{S}_x$, $\mathbf{S}_y$, ...)
- Classifier l'observation laissée de côté
- Répéter cette procédure pour chaque observation.

Nous pouvons alors estimer les probabilités de misclassification.

En ayant recours à la discrimination linéaire, J&W ont appliqué cette procédure aux données saumoniennes.

		prédit	
		π_1	π_2
vrai	π_1	44	6
	π_2	1	49

Remarque. Il s'avère plus rare que des saumons canadiens (de naissance) sont misclassifiés comme aléoutes que vice-versa. La procédure n'est donc pas correcte dans un certain sens, et croire aveuglément en une certaine procédure est dangereux! Surtout pour les pêcheurs de saumons...

3: Apprentissage supervisé: classification avec régression logistique.

On considère de nouveau ici de la classification pour laquelle on observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ des copies i.i.d. de (X, Y) avec une réponse Y binaire (qui prend les valeurs 0 ou 1).

Ce qui nous intéresse ici, c'est

$$p(X) = \mathrm{P}(Y = 1|X).$$

Si on utilise un simple modèle de régression du type $p(X) = \beta_0 + \beta_1 X$, il est typique que l'on obtienne des probabilités estimées négatives ou plus grande que 1.

Pour éviter ce problème, il faut modéliser $p(X)$ à l'aide d'une fonction qui donnera des estimations entre 0 et 1.

Beaucoup de fonctions font ceci mais en régression logistique, on utilise

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{(1 + \exp(\beta_0 + \beta_1 X))}$$

Notons que

$$\frac{p(X)}{(1 - p(X))} = \exp(\beta_0 + \beta_1 X);$$

les valeurs $\frac{p(X)}{(1-p(X))}$ sont appelées les "odds" et prennent leurs valeurs dans $(0, \infty)$.

On a donc

$$\log(\frac{p(X)}{(1-p(X))}) = \beta_0 + \beta_1 X$$

qui sont donc les log-odds ou logit.

The logistic regression model has logits that are linear functions of X.

The coefficients β_0 and β_1 are unknown and have to be estimated.

On peut utiliser les estimateurs maximu de vraisemblance de β_0 et β_1. La fonction de vraisemblance conditionnelle est donnée par

$$\ell(\beta_0, \beta_1) := \prod_{i=1}^{n} \mathrm{P}[Y_i = y_i | X_i = x_i] = \prod_{i=1}^{n} (p(x_i))^{y_i} (1 - p(x_i))^{1-y_i}$$

telle que

$$\begin{aligned} \log(\ell(\beta_0, \beta_1)) &= \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)) \\ &= \sum_{i=1}^{n} \log(1 - p(x_i)) + y_i(\beta_0 + \beta_1 x_i) \end{aligned}$$

Malheureusement, pas de forme explicite pour l'estimateur maximum de vraisemblance.

Analyse discriminante avec R:

```
## The Stock Market Data
library(ISLR2)
names(Smarket)
dim(Smarket)
summary(Smarket)
pairs(Smarket)
cor(Smarket)
cor(Smarket[, -9])
attach(Smarket)
plot(Volume)
## Logistic Regression
glm.fits <- glm(
    Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Vo
    data = Smarket, family = binomial
  )
summary(glm.fits)
```

```
Linear Discriminant Analysis
###
library(MASS)
lda.fit <- lda(Direction ~ Lag1 + Lag2, data = Smarke
    subset = train)
lda.fit
plot(lda.fit)
## Quadratic Discriminant Analysis
###
qda.fit <- qda(Direction ~ Lag1 + Lag2, data = Smarke
    subset = train)
qda.fit
```