

Apprentissage supervisé - Introduction (suite)

Mathieu Pigeon

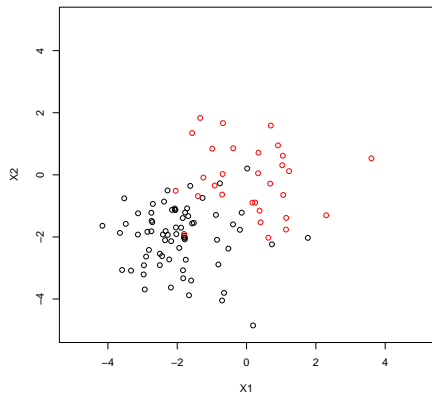
UQAM

- 1 Introduction
- 2 Métrique
- 3 Modèle Bayes
- 4 Régression binaire
- 5 Application

Problématique - Classification

- Dans une problématique de type *classification*, on dispose d'une base de données de taille n : $(g_i, \mathbf{x}_i)_{i=1, \dots, n}$ avec $\mathbf{x}_i = [x_{i1} \ \cdots \ x_{ip}]$.
- g_i est la variable réponse (catégorielle) et \mathbf{x}_i est un vecteur de variables explicatives (numériques ou catégorielles).
- On cherche à déterminer à quelle catégorie g^* une nouvelle observation dont les variables explicatives sont \mathbf{x}^* appartient.

Problématique - Classification



Problématique - Classification

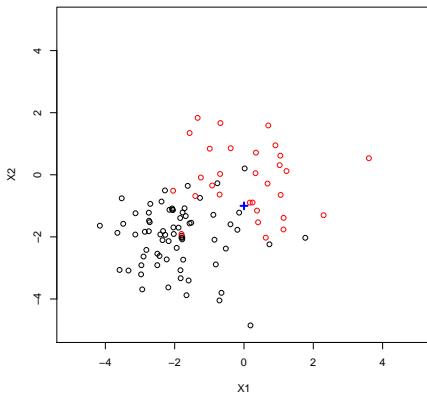


FIGURE – On cherche à classer la nouvelle observation (croix bleue) → on cherche à prédire si cette observation sera rouge ou noire.

Problématique - Classification

- On cherche à déterminer une fonction \hat{f} qui permettra de classifier une nouvelle observation :

$$\hat{f}(\mathbf{x}) = g_1 \text{ ou } g_2.$$

- Cette fonction doit bien performer non seulement sur la base de données (base d'entraînement) mais également sur des observations non utilisées pour l'estimation de \hat{f} (base de validation).
- Les concepts vus pour une problématique de régression s'appliquent à nouveau. Il faudra apporter quelques modifications pour tenir compte du fait que la variable réponse est maintenant catégorielle.

Problématique - Classification

Le lien entre les variables explicatives et la variable réponse n'est pas construit de la même façon :

- régression : $Y = f(\mathbf{X}) + \epsilon$
- classification : $\Pr(Y = g | \mathbf{X} = \mathbf{x})$, $g \in \mathcal{G}$, où \mathcal{G} est l'ensemble des catégories.

Erreur d'entraînement

- On définit l'erreur d'entraînement comme étant

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\hat{g}_i \neq g_i},$$

où \hat{g}_i est la classe prédite pour l'observation i .

- Il s'agit simplement de la proportion d'observations incorrectement classées dans l'échantillon disponible.
- On cherche à minimiser cette fonction.

Erreur de validation

- Pour les mêmes raisons qu'en régression, on cherchera plutôt à déterminer la performance du modèle à partir d'un échantillon non utilisé pour l'ajustement du modèle (base de validation).
- L'erreur de validation est donnée par

$$\frac{1}{m} \sum_{i=1}^m \mathbb{I}_{\hat{g}_i \neq g_i},$$

où $\{(g_i, \mathbf{x}_i)\}_{i=1, \dots, m}$ est une base de validation.

- Le « meilleur » modèle est celui qui minimise l'erreur de validation.

Modèle Bayes

- Le modèle Bayes (*Bayes classifier*) est le modèle qui maximise la probabilité de classer correctement une observation :

$$\hat{Y} = \arg \max_{g \in \mathcal{G}} \Pr(Y = g | \mathbf{X} = \mathbf{x}).$$

- La distribution de Y n'est pas connue : il n'est pas possible, en pratique, d'utiliser le modèle Bayes.
- Dans notre exemple introductif, on aura

$$\hat{Y} = \begin{cases} \text{rouge,} & \Pr(Y = \text{rouge} | X_1 = x_1, X_2 = x_2) \geq 0.5 \\ \text{noir,} & \text{sinon.} \end{cases}$$

Taux d'erreur du modèle Bayes

- La probabilité que le modèle Bayes classifie incorrectement une observation est donnée par

$$1 - \max_{g \in \mathcal{G}} \Pr(Y = g | \mathbf{X} = \mathbf{x}).$$

- On a

$$\begin{aligned} \Pr(\hat{Y} \neq Y) &= \mathbb{E} \left[\Pr(\hat{Y} \neq Y | \mathbf{X}) \right] \\ &= \mathbb{E} \left[1 - \max_{g \in \mathcal{G}} \Pr(Y = g | \mathbf{X}) \right] \\ &= 1 - \mathbb{E} \left[\max_{g \in \mathcal{G}} \Pr(Y = g | \mathbf{X}) \right]. \end{aligned}$$

- Le modèle Bayes étant le « meilleur » possible, son taux d'erreur est une borne minimale.

Problématique - Classification

- Rappel : classification : $\Pr(Y = g | \mathbf{X} = \mathbf{x}), g \in \mathcal{G}$.
- Si $Y \sim \text{Bernoulli}(p)$, alors

$$\Pr(Y = y) = p^y(1 - p)^{1-y} \text{ et } \mathbb{E}[Y] = p.$$

- Cadre des modèles linéaires généralisés (*generalized linear models* ou GLM).
- On considère (dans le cadre du cours) uniquement le cas où la variable réponse est binaire mais le modèle peut se généraliser au cas où la variable réponse est multi-catégorielle.

Modèles linéaires généralisés

- La distribution de Y est membre de la famille exponentielle linéaire

$$f_Y(y) = c(y, \phi) \exp \left(\frac{y\theta - a(\theta)}{\phi} \right)$$

$$g(\mathbb{E}[Y]) = \mathbf{X}^T \boldsymbol{\beta},$$

où $a()$ et $c()$ sont des fonctions, θ est le paramètre canonique, ϕ est le paramètre de dispersion, $\boldsymbol{\beta}$ est un vecteur de paramètres et $g()$ est la **fonction lien**.

- Les paramètres peuvent être estimés par maximum de vraisemblance (généralement numériquement).

Modèles linéaires généralisés

- On a alors

$$\mathbb{E}[Y] = a'(\theta)$$

$$\text{Var}[Y] = \phi a''(\theta)$$

$$\rightarrow \text{Var}[Y] = \phi \mathcal{V}(\mathbb{E}[Y]),$$

où $\mathcal{V}()$ est la **fonction de variance**.

- Pour le modèle logistique, on utilise généralement deux fonctions lien :
 - fonction logit : $g(x) = e^x / (1 + e^x)$ et
 - fonction probit : $g(x) = \Phi(x)$, où $\Phi()$ est la fonction de répartition d'une variable aléatoire Normale centrée et réduite.

Modèles linéaires généralisés

- Par exemple, avec un modèle logistique, la probabilité d'observer $Y = 1$ pour l'observation i est donnée par

$$\begin{aligned}\Pr(Y_i = 1 | \mathbf{X}_i = \mathbf{x}_i) &= \mathbb{E}[Y_i | \mathbf{X}_i = \mathbf{x}_i] \\ &= \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}} \in (0, 1).\end{aligned}$$

- On nomme généralement *Score* la composante $S = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.

Classification

- Pour utiliser ce modèle afin de faire une classification, il faut déterminer τ tel que

$$\hat{Y} = \begin{cases} g_1 & \Pr(Y = 1 | \mathbf{X} = \mathbf{x}) \geq \tau \\ g_2 & \text{sinon,} \end{cases}$$

où g_1 et g_2 sont les classes possibles.

- Ainsi, τ peut être considéré comme un hyper-paramètre du modèle qu'il faut déterminer.

Classification

- À partir de la base de données de validation, on détermine les vrais positifs, les vrais négatifs, les faux positifs et les faux négatifs obtenus pour une valeur donnée de τ .
- On définit la **sensibilité** comme étant la probabilité de détecter correctement un négatif, c'est-à-dire

$$\frac{\text{0 correctement détectés}}{\text{nombre total de 0}}.$$

- On définit la **spécificité** comme étant la probabilité de détecter correctement un positif, c'est-à-dire

$$\frac{\text{1 correctement détectés}}{\text{nombre total de 1}}.$$

- Le « meilleur » modèle sera celui qui permettra de capturer le plus possible de vrais positifs et de capturer le plus possible de vrais négatifs.

Assurance sommeil

- On considère la base de données *Ronfle.txt* disponible sur le site du cours. On tente d'expliquer le fait qu'une personne ronfle ou ne ronfle pas à l'aide de différentes variables explicatives.
- Les variables d'intérêt sont :
 - **Age** : variable continue ;
 - **Sexe** : 0 : homme, 1 : femme ;
 - **Ronfle** : 0 : non, 1 : oui ;
 - **Tabac** : variable catégorielle ;
 - **IMCDisc** : variable catégorielle ;
 - **AlcoolDisc** : variable catégorielle.

Préparation des données

```
Ronfle <- read.delim("~/Dropbox/Cours/ACT6100/Ronfle.txt")
```

```
### Création des catégories
```

```
Ronfle$Ronfle <- as.factor(Ronfle$Ronfle)
```

```
Ronfle$Sexe <- as.factor(Ronfle$Sexe)
```

```
Ronfle$Tabac <- as.factor(Ronfle$Tabac)
```

```
Ronfle$IMCDisc <- as.factor(Ronfle$IMCDisc)
```

```
Ronfle$AlcoolDisc <- as.factor(Ronfle$AlcoolDisc)
```

Ajustement du modèle initial

```
modele1 <- glm(Ronfle~Age + Sexe + Tabac + IMCDisc
               + AlcoolDisc,
               family = binomial(link = "logit"),
               data = Ronfle)
```

```
summary(modele1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.22839	2.01649	-1.601	0.10938	
Age	0.06055	0.02344	2.583	0.00980	**
Sexe1	-0.44253	0.74063	-0.598	0.55017	
Tabac1	-1.27973	0.58062	-2.204	0.02752	*
IMCDisc2	-0.82504	1.50769	-0.547	0.58422	
IMCDisc3	-0.63151	1.39831	-0.452	0.65154	
IMCDisc4	-1.19997	1.58286	-0.758	0.44839	

...

Modèle initial

- L'équation du modèle initial est donc

$$\begin{aligned}\text{logit}(P(Y = 1|X)) = & -3.23 - 0.44\text{Femme} + 0.06\text{Age} - 0.83\text{Normal} \\ & - 0.63\text{Surpoids} - 1.20\text{Obèse} - 1.28\text{Fumeur} \\ & + 0.99(1-4 \text{ verres}) + 2.24(+4 \text{ verres}).\end{aligned}$$

- Certaines variables étant non-significatives, on peut tenter d'alléger le modèle. Il est conseillé d'enlever les variables pas à pas, mais plus il y a de variables explicatives dans le modèle initial, plus ce travail sera long.

Sélection automatique des variables

```
library(MASS)
stepAIC(modele1, direction = "backward")
```

Start: AIC=126.4

Ronfle ~ Age + Sexe + Tabac + IMCDisc + AlcoolDisc

	Df	Deviance	AIC
- IMCDisc	3	109.21	121.21
- Sexe	1	108.77	124.77
<none>		108.41	126.41
- Tabac	1	113.70	129.70
- AlcoolDisc	2	117.77	131.77
- Age	1	115.97	131.97
...			

Ajustement du modèle final

```
modele2 <- glm(formula = Ronfle ~ Age + Tabac + AlcoolDisc,
               family = binomial(link = "logit"), data = Ronfle)
summary(modele2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.24010	1.39577	-3.038	0.002383	**
Age	0.06111	0.02334	2.618	0.008839	**
Tabac1	-1.17120	0.55651	-2.105	0.035332	*
AlcoolDisc1	1.21450	0.59950	2.026	0.042779	*
AlcoolDisc2	2.37953	0.71553	3.326	0.000882	***

Modèle final

- L'équation du modèle final est donc

$$\begin{aligned}\text{logit}(P(Y = 1|X)) = & -4.24 + 0.06\text{Age} - 1.17\text{Fumeur} \\ & + 1.21(1-4 \text{ verres}) + 2.38(+4 \text{ verres}).\end{aligned}$$

Sélection de l'hyper-paramètre

```
### Remplacer TAU par la valeur limite souhaitée,  
### NOM_BASE_DONNEES par le nom de la base de données,  
### NOM_MODELE par le nom du modèle et  
### NUM_VAR par le numéro de la colonne de la variable  
### réponse.
```

```
POS <- function(tau, D, M, NUM)  
{  
  PRED <- predict(M, type = "response")  
  VRAI <- (D[NUM] == 1)  
  PREDICTION <- (PRED > x)  
  chisq.test(VRAI, PREDICTION, correct = FALSE)$observed  
}
```

```
POS(TAU, NOM_BASE_DONNEES, NOM_MODELE, NUM_VAR)
```

Classification

En utilisant une base de données de validation, on obtient les résultats ci-dessous.

τ	Vrai pos.	Vrai nég.	Faux pos.	Faux nég.
0.1	34	10	55	1
0.3	28	34	31	7
0.4	22	45	20	13
0.5	14	55	10	21
0.6	9	59	6	26
0.7	7	65	0	28

TABLE – Résultats pour différentes valeurs de τ .

Classification

Ici, un point de rupture de $\tau^* = 0.5$ semble convenable. Pour $\tau^* = 0.5$, on a

- Sensibilité $= \frac{55}{55 + 10} = 0.85$;
- Spécificité $= \frac{14}{14 + 21} = 0.40$; et
- Proportion d'individus bien classés : $\frac{14 + 55}{14 + 55 + 10 + 21} = 0.69$.

Classification

Dans quelle catégorie sera classé un non-fumeur de 50 ans et ne buvant pas ?

À partir du modèle final, on obtient un score :

$$S = -4.2401 + (0.0611)(50) = -1.1851$$

que l'on transforme en probabilité :

$$\Pr(Y = 1) = \frac{e^{-1.1851}}{1 + e^{-1.1851}} = 0.2342.$$

Puisque $0.2342 < \tau^* = 0.5$, on classera le nouvel individu dans la catégorie des gens qui ne ronflent pas.