Apprentissage supervisé - Introduction (suite)

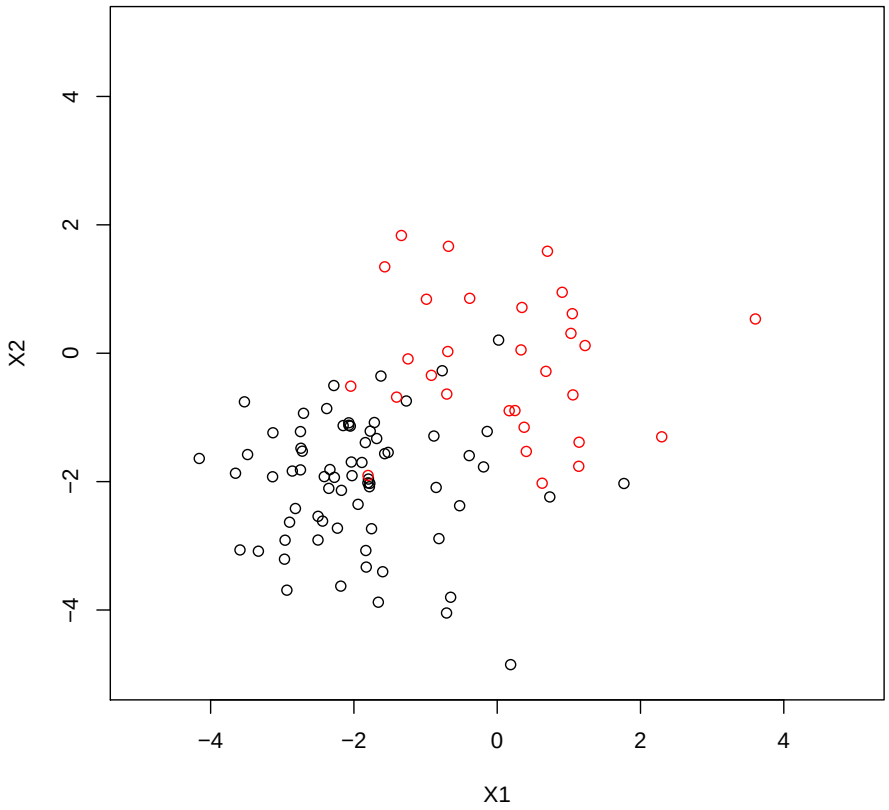
Mathieu Pigeon

UQAM

Problématique - Classification

- Dans une problématique de type *classification*, on dispose d'une base de données de taille n : $(g_i, \mathbf{x}_i)_{i=1,\ldots,n}$ avec $\mathbf{x}_i = \begin{bmatrix} x_{i1} & \cdots & x_{ip} \end{bmatrix}$.
- 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

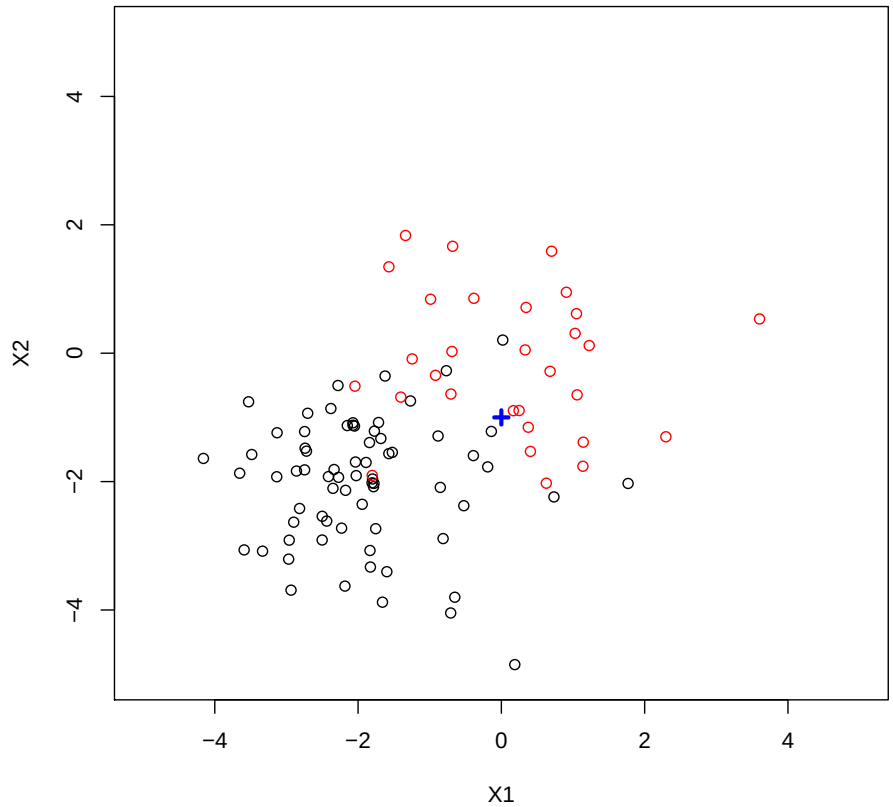


FIGURE – On cherche à classifier la nouvelle observation (croix bleue) → on cherche à prédire si cette observation sera rouge ou noire.

Problématique - Classification

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

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

- Cette fonction doit bien performer non seulement sur la base de données (base d'entrainement) mais également sur des observations non utilisées pour l'estimation de $\widehat{f}$ (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'entrainement

- On définit l'erreur d'entrainement comme étant

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

  où $\widehat{g}_i$ 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 raison 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

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

où $\{(g_i, \mathbf{x}_i)\}_{i=1,\ldots,m}$ 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 classifier correctement une observation :

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

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

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

Taux d'erreur du modèle Bayes

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

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

- On a

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

- 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

$$\begin{aligned} \mathbb{E}[Y] &= a'(\theta) \\ \text{Var}[Y] &= \phi a''(\theta) \\ \rightarrow \text{Var}[Y] &= \phi \mathcal{V}(\mathbb{E}[Y]), \end{aligned}$$

  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

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

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

Classification

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

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

  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

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

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

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

- 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.

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

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.