EE-311—Apprentissage et intelligence artificielle

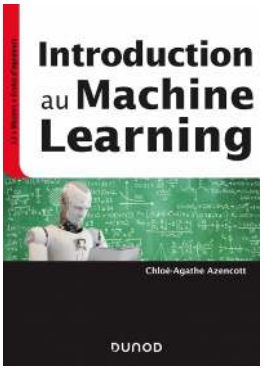
10. Modèles de densités et inférence Bayésienne

Michael Liebling

https://moodle.epfl.ch/course/view.php?id=16090

17 mai 2024 (compilé le 21 juin 2024)

Ouvrage de référence et source



Ces transparents sont basés en grande partie sur le texte de Chloé-Agathe Azencott "Introduction au Machine Learning", Dunod, 2019
ISBN 978-210-080153-4

L'auteure a mis le texte (sans les exercices) à disposition ici : http://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf

Avertissement : Bien que ces transparents partagent la notation mathématique, la structure de l'exposition (en partie), et certains exemples avec le livre, ils ne constituent qu'un complément et non un remplacement ou une source unique pour la couverture des matières du cours. À ce titre, ces transparents ne se substituent pas au texte.

Objectifs de cette leçon

1. Formaliser le concept de classe grâce à des modèles probabilistes
2. Définir des règles de décision, sur la base de tests de rapport de vraisemblance :
   - 2.1 décision par maximum de vraisemblance
   - 2.2 décision par maximum a posteriori
   - 2.3 décision par minimisation du risque de Bayes
3. Deux techniques d'estimation de densités de probabilité
   - 3.1 par maximum de vraisemblance (MLE : maximum likelihood estimator)
   - 3.2 par estimateur de Bayes

Partie 1 : Formalisme probabiliste pour la classification

Modèles génératifs pour la classification binaire

L'approche statistique de la classification formalise le concept de classe grâce à des modèles probabilistes.

Nous allons dès lors considérer que :

- les n observations $\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n$ sont la réalisation d'une variable aléatoire $X \in \mathcal{X}$.
- leurs étiquettes $y_1, y_2, \ldots, y_n$ sont la réalisation d'une variable aléatoire
  - classification binaire : $Y \in \{0, 1\}$
  - classification multi-classe : $Y \in \{1, 2, \ldots, C\}$ (où C est le nombre total de classes)

Modélisation générative

La **modélisation générative** consiste à considérer une loi de probabilité jointe $\mathbb{P}(X, Y)$ pour l'ensemble des variables entrant dans le modèle (donc à la fois X et Y).

Interprétation : La modélisation générative répond à la question :

Comment les données que l'on observe auraient-elles pu être générées ?

En les modélisant comme la réalisation d'une variable aléatoire, on a une façon efficace de représenter l'information complexe qu'elles contiennent.

Exemple : modélisation d'un test médical

Les résultats d'un test de dépistage médical seront plus facilement représentés si on considère qu'un résultat est modélisé par une loi de Bernoulli plutôt que par un processus biochimique complexe (dont on ne connaîtrait pas forcément tous les paramètres).

Source (Roche rapid antigen test) : `https://diagnostics.roche.com/`

Photo : ML

Attention cependant :

- La loi aléatoire ne modélisera *que* la distribution des résultats (mais pas le test lui-même) : en tant qu'individu, le lancer de la pièce n'est pas une alternative viable à faire un test (sérieux) !
- En pratique, les données ne sont pas toujours le fruit d'un processus aléatoire.

Définition des concepts d'inférence et prédiction

Probabilité d'appartenance à une classe : la probabilité qu'une observation $\vec{x}$ appartienne à la classe c est déterminée par

$$\mathbb{P}(Y = c | X = \vec{x})$$

Inférence : un problème d'inférence consiste à déterminer les lois de probabilité $\mathbb{P}(Y = c | X = \vec{x})$ à partir de nos observations et hypothèses

Prédiction (ou décision) : un problème de prédiction utilise les lois de probabilité pour déterminer la classe y d'une observation $\vec{x}$

Règle de décision simple (principe de base)

Étant données les probabilités d'appartenance à une classe suite à l'observation de $\vec{x}$:

- $\mathbb{P}(Y = 1|X = \vec{x})$
- $\mathbb{P}(Y = 0|X = \vec{x})$

nous considérons la règle de décision simple (classification binaire) :

$$\hat{y} = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = \vec{x}) > \mathbb{P}(Y = 0|X = \vec{x}) \\ 0 & \text{sinon} \end{cases}$$

et dans le cas de la classification multi-classe :

$$\hat{y} = \underset{c=1,\ldots,C}{\operatorname{argmax}} \mathbb{P}(Y = c|X = \vec{x})$$

Convention d'écriture des probabilités

Nous écrirons parfois

$$\mathbb{P}(\vec{x})$$

au lieu de

$$\mathbb{P}(X = \vec{x})$$

quand il n'y a pas d'ambiguité.

Loi de Bayes

Théorème 4.1 (Loi de Bayes)

$$\mathbb{P}(Y = c|\vec{x}) = \frac{\mathbb{P}(Y = c)\,\mathbb{P}(\vec{x}|Y = c)}{\mathbb{P}(\vec{x})}$$

Avec

- $\mathbb{P}(Y = c|\vec{x})$: la distribution a posteriori des étiquettes (*après* avoir observé une réalisation $\vec{x}$)
- $\mathbb{P}(Y = c)$: la distribution a priori des étiquettes (connue *avant* d'avoir observé $\vec{x}$)
- $\mathbb{P}(\vec{x}|Y = c)$: la vraisemblance (que l'on observe la réalisation $\vec{x}$ de X sachant que la classe est c)
- $\mathbb{P}(\vec{x})$: la probabilité marginale que $\vec{x}$ soit observée (indépendamment de sa classe)

Probabilité marginale

$\mathbb{P}(\vec{x})$, la probabilité marginale que $\vec{x}$ soit observée (indépendamment de sa classe) peut s'écrire sous la forme (classification binaire) :

$$\mathbb{P}(\vec{x}) = \mathbb{P}(\vec{x}|Y = 0)\,\mathbb{P}(Y = 0) + \mathbb{P}(\vec{x}|Y = 1)\,\mathbb{P}(Y = 1)$$

ou (cas multi-classe) :

$$\mathbb{P}(\vec{x}) = \sum_{c=1}^{C} \mathbb{P}(\vec{x}|Y = c)\,\mathbb{P}(Y = c)$$

Exemple : dépistage cancer col de l'utérus

Un test de dépistage du cancer du col de l'utérus a les caractéristiques suivantes :

- test positif suggère présence du cancer
- test négatif suggère absence de cancer
- sensibilité : 70% (parmi les personnes vraiment atteintes, la proportion pour lesquelles le test est positif)
- spécificité : 98% (parmi les personnes non-atteintes, la proportion pour lesquelles le test est négatif)

L'incidence de la maladie est d'environ 1 femme sur 10'000.

Quelle est la probabilité qu'une personne testée soit atteinte d'un cancer si le test est positif ? (positive predictive value)

Solution (dépistage cancer de l'utérus)

Notation :

- résultat du test : X variable aléatoire binaire
  (1 : test positif, 0 test négatif)
- statut de la personne : Y variable aléatoire binaire
  (1 : atteinte, 0 non-atteinte)

La question : "Quelle est la probabilité qu'une personne testée soit atteinte d'un cancer si le test est positif ?" revient à inférer

$\mathbb{P}(Y = 1|X = 1)$ i.e. : 'atteinte' $(Y = 1)$ sachant que 'test positif' $(X = 1)$

que l'on calcule selon la loi de Bayes par

$$\mathbb{P}(Y = 1|X = 1) = \frac{\mathbb{P}(Y = 1)\,\mathbb{P}(X = 1|Y = 1)}{\mathbb{P}(X = 1)}$$

Solution (suite)

Or on reconnaît :

- sensibilité : 70% (la proportion de personnes vraiment atteintes pour lesquelles le test est positif) $\equiv \mathbb{P}(X = 1|Y = 1)$
- spécificité : 98% (la proportion de personnes non-atteintes pour lesquelles le test est négatif)
  $\equiv \mathbb{P}(X = 0|Y = 0) = 1 - \mathbb{P}(X = 1|Y = 0)$
- L'incidence de la maladie est d'environ 1 femme sur 10'000
  $\equiv \mathbb{P}(Y = 1) = 1 - \mathbb{P}(Y = 0)$
- Marginale (test positif) :
  $\mathbb{P}(X = 1) = \mathbb{P}(X = 1|Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X = 1|Y = 1)\mathbb{P}(Y = 1)$

Ce qui nous donne :

$$\begin{aligned}\mathbb{P}(Y = 1|X = 1) &= \frac{\mathbb{P}(Y = 1)\mathbb{P}(X = 1|Y = 1)}{\mathbb{P}(X = 1)} \\ &= \frac{10^{-4} \times 0.7}{(1 - 0.98) \times (1 - 10^{-4}) + 0.7 \times 10^{-4}} \\ &= 0.0035 = 0.35\%\end{aligned}$$

Application naïve de la règle de décision simple lors du dépistage

Avec la probabilité calculée ci-dessus, si on appliquait la règle de décision simple évoquée plus haut :

$$\hat{y} = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|X = \vec{x}) > \mathbb{P}(Y = 0|X = \vec{x}) \\ 0 & \text{sinon} \end{cases}$$

vu que la probabilité qu'une personne ne soit pas atteinte si le test est positif est :

$$\begin{aligned}\mathbb{P}(Y = 0|X = 1) &= 1 - \mathbb{P}(Y = 1|X = 1) \\ &= 0.9965 = 99.65\%\end{aligned}$$

on prédirait la classe négative (pas atteint) pour presque tous les tests positifs !

$\Rightarrow$ **le dépistage est utilisé pour identifier les personnes qui devraient faire un test plus fiable, pas pour faire un diagnostique.**

Dépistage peu fiable également si résultat négatif ?

La probabilité qu'une personne soit atteinte si le test est négatif est très basse :

$$\begin{aligned}\mathbb{P}(Y=1|X=0) &= \frac{\mathbb{P}(Y=1)\,\mathbb{P}(X=0|Y=1)}{\mathbb{P}(X=0)} \\ &= \frac{\mathbb{P}(Y=1)\,\mathbb{P}(X=0|Y=1)}{\mathbb{P}(X=0|Y=0)\,\mathbb{P}(Y=0)+\mathbb{P}(X=0|Y=1)\,\mathbb{P}(Y=1)} \\ &= \frac{10^{-4}\times(1-0.7)}{0.98\times(1-10^{-4})+(1-0.7)\times 10^{-4}} \\ &= 3.06\times 10^{-5} = 0.003\%\end{aligned}$$

Par conséquent, il est raisonable de faire un diagnostic 'non-atteinte' sur la base d'un test de dépistage négatif.
(Comme vu au slide précédent, on se gardera par contre de faire un diagnostic sur l'unique base d'un test positif et on prescrira un test plus fiable (mais potentiellement plus coûteux, lent, etc.))

Vertus de tester un grand nombre en limitant le nombre de test couteux/lents

Ne vaudrait-il pas la peine d'utiliser un test plus fiable dès le départ ?

Lors d'une campagne de dépistage on mesurera $\mathbb{P}(X=0) = \mathbb{P}(X=0|Y=0)\,\mathbb{P}(Y=0)+\mathbb{P}(X=0|Y=1)\,\mathbb{P}(Y=1) = 0.98\times(1-10^{-4})+(1-0.7)\times 10^{-4} = 0.9799 = 97.99\%$ de test négatifs, ce qui permet de libérer de test plus poussés les personnes avec ce résultat (avec bonne confiance).

L'intérêt d'adopter un test de dépistage (avec une sensibilité et une spécificité donnée) sera donc une fonction :
- de la prévalence (actuelle, estimée) de la maladie dans la population,
- des conséquence d'un faux négatif (pour la personne, pour le système de santé),
- des moyens à disposition pour des test plus poussés
- des coût induits par des test plus poussés.

Partie 2 : Règles de décision

Règles de décision (via calcul du rapport de vraisemblance)

Tests du *rapport de vraisemblance* pour décisions par :

1. maximum de vraisemblance
2. maximum a posteriori (note : sera dérivé en premier ici)
3. minimisation du risque de Bayes

↑ par ordre croissant de généralité, mais on va les dériver dans l'ordre 2. → 1. → 3.

Rappel **Loi de Bayes**

$$\mathbb{P}(Y = c|\vec{x}) = \frac{\mathbb{P}(Y = c)\,\mathbb{P}(\vec{x}|Y = c)}{\mathbb{P}(\vec{x})}$$

- $\mathbb{P}(Y = c|\vec{x})$: distribution a posteriori des étiquettes (après avoir observé $\vec{x}$)
- $\mathbb{P}(Y = c)$: la distribution a priori des étiquettes (avant d'avoir observé $\vec{x}$
- $\mathbb{P}(\vec{x}|Y = c)$: **la vraisemblance (que l'on observe la réalisation $\vec{x}$ de X sachant que la classe est c)**←−
- $\mathbb{P}(\vec{x})$ la probabilité marginale que $\vec{x}$ soit observée (indépendamment de sa classe)

Dérivation des tests du rapport de vraisemblance

Idée : on veut des tests qui font intervernir les vraisemblances, i.e. $\mathbb{P}(\vec{x}|Y=c)$, $c=0,1$, les vraisemblances que l'on observe la réalisation $\vec{x}$ de X sachant que la classe est soit $c=0$ ou $c=1$.

Dérivation : On part de la règle de décision simple :

$$\hat{y} = \begin{cases} 1 & \text{si } \mathbb{P}(Y=1|X=\vec{x}) > \mathbb{P}(Y=0|X=\vec{x}) \\ 0 & \text{sinon,} \end{cases}$$

qui consiste donc à prédire la classe la plus probable étant donnée l'observation, ce qui correspond à séléctionner la classe $\hat{y}$ qui maximise la valeur a posteriori $\mathbb{P}(Y=\vec{y}|\vec{x})$.
Par la loi de Bayes on peut ré-écrire cette règle en faisant intervenir les vraisemblances $\mathbb{P}(\vec{x}|Y=c)$, $c=0,1$:

$$\hat{y} = \begin{cases} 1 & \text{si } \frac{\mathbb{P}(\vec{x}|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(\vec{x})} > \frac{\mathbb{P}(\vec{x}|Y=0)\mathbb{P}(Y=0)}{\mathbb{P}(\vec{x})} \\ 0 & \text{sinon} \end{cases}$$

puis finalement, en simplifiant $\mathbb{P}(\vec{x})$ on obtient ...

Décision par maximum a posteriori (formulation par vraisemblances et distribution d'étiquettes a priori)

Décision par maximum a posteriori

- Dans le cas binaire la règle de décision

$$\hat{y} = \begin{cases} 1 & \text{si } \mathbb{P}(\vec{x}|Y=1)\,\mathbb{P}(Y=1) > \mathbb{P}(\vec{x}|Y=0)\,\mathbb{P}(Y=0) \\ 0 & \text{sinon} \end{cases}$$

est appelée *règle de décision* par maximum a posteriori.

- Dans le cas multi-classe, cette règle s'écrit

$$\hat{y} = \underset{c=1,\ldots,C}{\operatorname{argmax}}\, \mathbb{P}(\vec{x}|Y=c)\,\mathbb{P}(Y=c)\,.$$

Note : Cette formulation sélectionne la classe qui maximise la probabilité a posteriori, $\mathbb{P}(Y=c|\vec{x})$, mais ne fait intervenir que $\mathbb{P}(Y=c)$ (distribution a priori des étiquettes = avant d'avoir observé la réalisation) et $\mathbb{P}(\vec{x}|Y=c)$ (vraisemblance que l'on observe la réalisation $\vec{x}$ de X sachant que la classe est c)

Rapport de vraisemblance

Définition 4.2 (Rapport de vraisemblance)
On représente par $\Lambda(\vec{x})$ le *rapport de vraisemblance* :

$$\Lambda(\vec{x}) = \frac{\mathbb{P}(\vec{x}|Y=1)}{\mathbb{P}(\vec{x}|Y=0)}$$

2. (voir diapo suivante pour 1.) Règle de décision par maximum a posteriori : formulation comme test sur le rapport de vraisemblance

Avec la définition du rapport de vraisemblance, la **règle de décision par maximum a posteriori** s'écrit :

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) > \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \\ 0 & \text{sinon.} \end{cases}$$

1. Règle de décision par maximum de vraisemblance (= avec hypothèse d'égalité des distributions a priori)

Avec l'hypothèse que les distributions a priori sont égales, $\mathbb{P}(Y=0) = \mathbb{P}(Y=1)$ (c'à-d, le rapport $\frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} = 1$) on peut définir :

Définition 4.3 (Décision par maximum de vraisemblance) La règle de décision

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) > 1 \\ 0 & \text{sinon} \end{cases}$$

est appelée *règle de décision par maximum de vraisemblance*.

Alternative : on préfère souvent exprimer cette règle sous forme de log :

$$\hat{y} = \begin{cases} 1 & \text{si } \log \Lambda(\vec{x}) > 0 \\ 0 & \text{sinon} \end{cases}$$

Exemple : déterminer le sexe de poissons à partir de leur longueur

Données : un échantillon d'une population de poissons (de même espèce) avec des mâles et des femelles.
But : déterminer leur sexe uniquement à partir de leur longueur.
Modèle :

- Y variable aléatoire binaire : 0 pour mâle, 1 pour femelle
- X variable aléatoire continue : longueur (en cm)

On suppose :
• Longueurs des femelles est normalement distribuée, centrée en 6 cm, et écart-type 1 cm :

$$\mathbb{P}(x|Y=1) \sim \mathcal{N}(6,1)$$

• Longueurs des mâles est normalement distribuée, centrée en 4 cm, et écart-type 1 cm :

$$\mathbb{P}(x|Y=0) \sim \mathcal{N}(4,1)$$

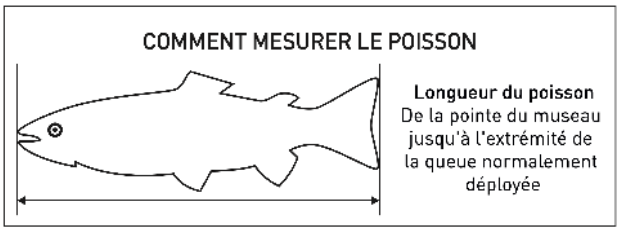
Guppy *(Poecilia reticulata)*

"Poisson d'eau douce tropicale, originaire d'Amérique du Sud." [1]
Un mâle en haut, deux femelles en bas [probablement. . . [2]]

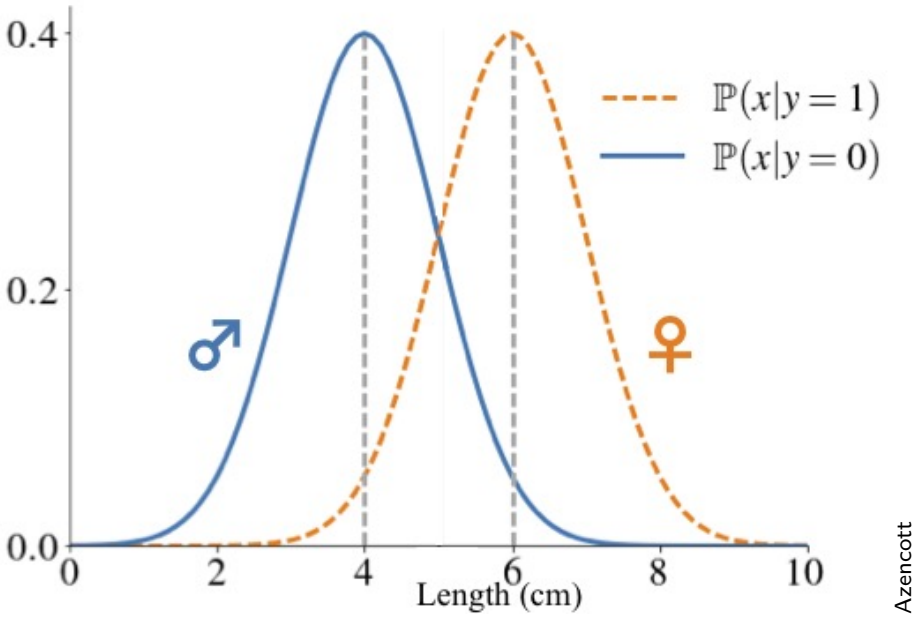
Source : `https://upload.wikimedia.org/wikipedia/commons/a/a2/Guppy_pho_0048.jpg`

1. Wikipedia : `https://fr.wikipedia.org/wiki/Guppy`
2. `https://mrfishkeeper.com/male-and-female-guppies/`

Distribution des longueurs (pour poissons de chaque sexe)



`https://www.vd.ch/fileadmin/user_upload/themes/environnement/faune_nature/fichiers_pdf/peche/01_Decouvrez_la_peche_dans_le_canton_de_VD/Extrait_Reglement_Leman.pdf`



Prédiction
(sous hypothèse que proportion mâles/femelles est 50/50)

Le rapport de vraisemblance s'écrit :

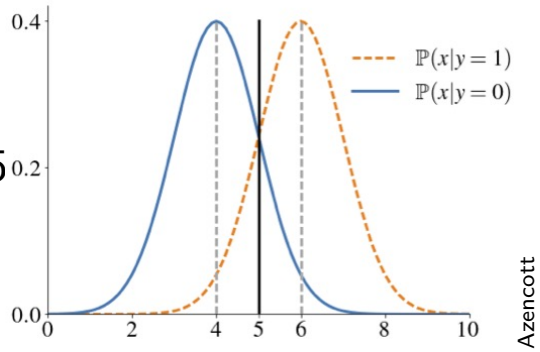
$$\Lambda(x) = \frac{\mathbb{P}(x|Y=1)}{\mathbb{P}(x|Y=0)} = \frac{e^{-(x-6)^2/2}}{e^{-(x-4)^2/2}}$$

et son logarithme vaut donc

$$\log \Lambda(x) = -\frac{1}{2}(x-6)^2 + \frac{1}{2}(x-4)^2 = 2(x-5)$$

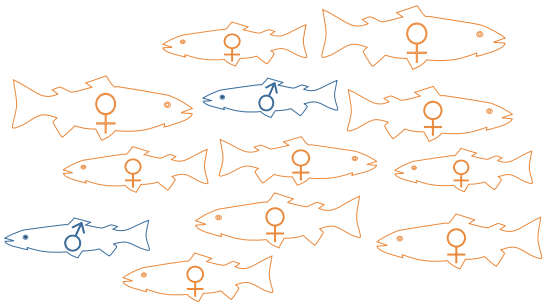
Par la règle de décision par maximum de vraisemblance on obtient

$$\hat{y} = \begin{cases} 1 \text{ si } \log \Lambda(x) > 0, \text{ donc si } x > 5 \\ 0 \text{ sinon} \end{cases}$$



Prédiction par maximum a posteriori

Information supplémentaire à propos de l'échantillon :
$\#(\text{femelles}) = 5 \times \#(\text{mâles})$
Le rapport des distributions a priori :

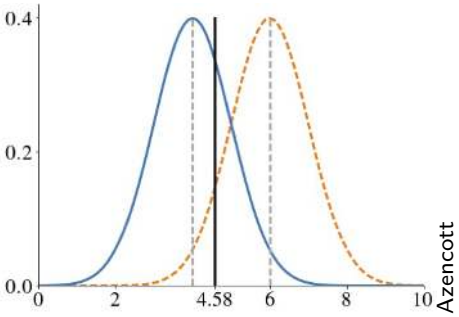


$$\frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} = \frac{1}{5}$$

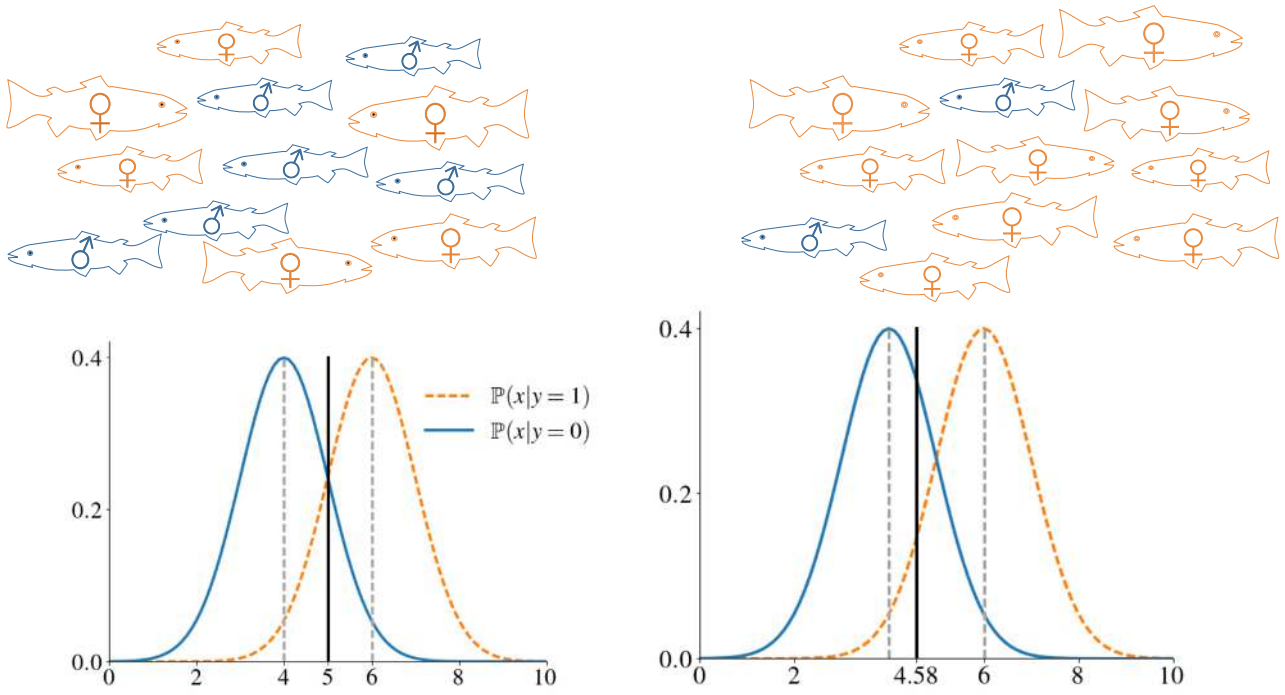
La règle du maximum a posteriori (log)

$$\hat{y} = \begin{cases} 1 \text{ si } \log \Lambda(x) > \log\left(\frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)}\right) \text{ c'à-d. si } x > 5 - \log(5)/4 \approx 4.58 \\ 0 \text{ sinon} \end{cases}$$

Connaissance a priori entraîne le déplacement de la valeur seuil :



Règles de décision : maximum de vraisemblance et maximum a posteriori



(A) Si les deux classes sont également probables, les poissons dont la taille est inférieure à 5 cm sont étiquetés mâles et les autres femelles.

(B) Si un poisson a cinq fois plus de chances a priori d'être femelle, les poissons dont la taille est inférieure à 4,58 cm sont étiquetés mâles et les autres femelles.

FIGURE 4.1 – Règle de décision pour le sexe d'un guppy en fonction de sa taille.

©Azencott

Théorie de la décision bayésienne

Les règles de décision vues :

- maximum de vraisemblance
- maximum a posteriori

s'inscrivent dans le cadre plus général de la **théorie de la décision**. Dans ce cadre,

- la variable aléatoire Y définie sur $\mathscr{Y}$ représente non pas une étiquette, mais une vérité cachée, ou un état de la nature
- la variable aléatoire X définie sur $\mathscr{X}$ représente les données observées ; De plus on considère :
- une variable A, définie sur un espace $\mathscr{A}$ qui représente l'ensemble des décisions (actions) qui peuvent être prises.

On se donne une fonction de coût :

$$L : \mathscr{Y} \times \mathscr{A} \to \mathbb{R}$$

Étant donné un état caché véritable y et une action a, la fonction de loss $L(y, a)$ quantifie le prix à payer pour avoir choisi l'action a alors que l'état caché véritable était y.

Décisions et météo

Dois-je prendre ou non mon parapluie ce matin ?

Dois-je prendre mon parapluie ?

Exemple : *Dois-je prendre ou non mon parapluie ce matin ?*
Nous pouvons modéliser ce problème de la façon suivante :

- $\mathscr{A}$ contient deux actions :
  prendre mon parapluie et *ne pas prendre mon parapluie*
- $\mathscr{Y}$ contient les vérités :
  il ne pleut pas, *il pleut un peu*, *il pleut fort*, *il y a beaucoup de vent*
- $\mathscr{X}$ espace décrivant les informations sur lesquelles je peux m'appuyer (prévisions météorologiques, couleur du ciel quand je pars de chez moi)

Je peux choisir la fonction de coût suivante :

	pas de pluie	pluie faible	pluie forte	vent
parapluie	1	0	0	2
pas de parapluie	0	2	4	0

et choisir l'action a qui minimise la probabilité d'erreur, i.e. l'espérance d'une fonction coût (suite...)

Décision de Bayes

Définition 4.4 (Décision de Bayes) La règle de décision qui consiste à choisir l'action a^* qui minimise l'espérance de la fonction de coût est appellée règle de décision de Bayes :

$$a^*(\vec{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \mathbb{E}[L(y, a)] = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y|\vec{x}) L(y, a)$$

Notes :

- On parlera aussi du principe de minimisation de la perte espérée (minimum expected loss en anglais.)
- En économie : on préfère au concept de fonction de coût celui d'utilité (peut être simplement définie comme l'opposé d'une fonction de coût). La minimisation deviendra une maximisation de l'utilité espérée (maximum expected utility)

Contraste avec la minimisation du risque empirique

La minimisation de la perte espérée est à contraster avec la minimisation du risque empirique (avec hypothèse $\approx$ action)

$$\mathscr{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\vec{x}^i), y^i).$$

dans laquelle on remplace la distribution $\mathbb{P}(X|Y)$ par sa distribution empirique obtenue en partageant de manière égale la masse de probabilité entre les n observations

$$\mathbb{P}(X = \vec{x}, Y = y|\mathscr{D}) = \frac{1}{n} \sum_{i=1}^{n} \delta(y, y^i) \delta(\vec{x}, \vec{x}^i)$$

Par contraste, dans le cadre bayésien, on paramétrise la distribution $\mathbb{P}(X, Y)$ par un paramètre $\vec{\theta}$ optimisé sur $\mathscr{D}$.

• Cadre empirique : hypothèses sur la distribution des données potentiellement simplistes

• Cadre bayésien : distribution apprise sans considérer le processus de décision dans lequel elle sera utilisée

Risque de Bayes

Alors que la décision de Bayes consiste à choisir, pour une observation donnée, l'espérance de la fonction de coût, on définit le risque de Bayes comme l'espérance globale de la fonction de coût :

Définition 4.5 (Risque de Bayes) Le *risque de Bayes* est l'espérance du coût sous la règle de décision de Bayes :

$$r = \int_{\vec{x} \in \mathcal{X}} \sum_{y \in \mathcal{Y}} L\left(y, a^*\left(\vec{x}\right)\right) \mathbb{P}\left(\vec{x}, y\right) \mathrm{d}\vec{x}$$

Note : Définir une stratégie qui minimise le risque de Bayes est équivalent à appliquer la règle de décision de Bayes.

Classification par la règle de décision de Bayes

Cadre de la classification binaire

- $y \in \mathcal{Y}$ représente la véritable classe d'une observation
- $a \in \mathcal{A}$ représente sa classe prédite
- classification binaire : $\mathcal{A} = \mathcal{Y} = \{0, 1\}$

La fonction de coût :

$$\begin{aligned} L : \mathcal{Y} \times \mathcal{Y} &\to \mathbb{R} \\ c, k &\mapsto \lambda_{ck} \end{aligned}$$

$\Rightarrow \lambda_{ck}$: coût de prédire la classe k quand la classe véritable est c.
La règle de décision de Bayes est équivalente à la règle de décision :

$$\hat{y} = \begin{cases} 1 & \text{si } \underbrace{\lambda_{11}\mathbb{P}\left(Y=1|\vec{x}\right)}_{\substack{\text{vraie classe} = 1 \\ \text{classe prédite} = 1}} + \underbrace{\lambda_{01}\mathbb{P}\left(Y=0|\vec{x}\right)}_{\substack{\text{vraie classe} = 0 \\ \text{classe prédite} = 1}} \\ & \qquad \leq \underbrace{\lambda_{10}\mathbb{P}\left(Y=1|\vec{x}\right)}_{\text{vraie} = 1\,;\ \text{prédite} = 0} + \underbrace{\lambda_{00}\mathbb{P}\left(Y=0|\vec{x}\right)}_{\text{vraie} = 0\,;\ \text{prédite} = 0} \\ 0 & \text{sinon} \end{cases}$$

3. Règle de décision de Bayes sous forme de test d'un rapport de vraisemblance et pour le cas multi-classe

La règle de décision de Bayes de la slide précédente peut se récrire sous la forme d'un test du rapport de vraisemblance :

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) = \frac{\mathbb{P}(\vec{x}|Y=1)}{\mathbb{P}(\vec{x}|Y=0)} > \frac{(\lambda_{01}-\lambda_{00})\mathbb{P}(Y=0)}{(\lambda_{10}-\lambda_{11})\mathbb{P}(Y=1)} \\ 0 & \text{sinon.} \end{cases}$$

Hint pour la dérivation : utiliser la loi de Bayes et le fait que $(\lambda_{11} - \lambda_{10}) < 0$.

Cadre de la classification multi-classe

Règle de décision de Bayes dans le cas multi-classe :

$$\hat{y} = \underset{k=1,\ldots,C}{\operatorname{argmin}} \sum_{c=1}^{C} \lambda_{ck} \mathbb{P}\left(Y = c|\vec{x}\right)$$

Interprétation : pour une classe candidate k, on somme λ_{ck} (le coût de prédire k lorsque la classe véritable est c) pour toutes les étiquettes $c = 1, \ldots, C$ en pondérent avec la probabilité a posteriori que l'étiquette est c lorsqu'on observe $\vec{x}$.

Coût 0/1 avec la règle de décision de Bayes

On retrouve le coût 0/1 (Section 2.4) en utilisant $\lambda_{ck} = 1 - \delta\left(k, c\right)$.

La règle de décision de Bayes devient

$$\hat{y} = \begin{cases} 1 & \text{si } \mathbb{P}\left(Y = 0|\vec{x}\right) \leq \mathbb{P}\left(Y = 1|\vec{x}\right) \\ 0 & \text{sinon} \end{cases}$$

et ainsi la règle de décision de Bayes est équivalente à la règle décision par maximum a posteriori. Ceci est vrai aussi dans le cas multi-classe.

Le coût 0/1 n'est pas la seule fonction de coût possible, même pour un problème de classification binaire. En particulier, toutes les erreurs de classification ne sont pas nécessairement également coûteuses. Par exemple, prédire qu'une patiente atteinte d'un cancer est en bonne santé peut être largement plus problématique que l'inverse.

Règle de décision par régions de décision

Les règles de décisions peuvent aussi s'exprimer en termes de régions de décision (cf. section 2.1) : la règle de décision consiste simplement à étiqueter l'observation $\vec{x}$ en accord avec la région de décision à laquelle elle appartient :

$$\hat{y} = \begin{cases} 1 & \text{si } \vec{x} \in \mathscr{R}_1 \\ 0 & \text{sinon.} \end{cases}$$

Dans le cas multi-classe, cette règle revient à

$$\hat{y} = \sum_{c=1}^{C} \delta_{\vec{x} \in \mathscr{R}_c}$$

Equivalence de la règle de décision par régions de décision avec la règle de décision de Bayes

Cette règle de décision est équivalente à la règle de décision de Bayes si l'on définit comme fonction discriminante la fonction :

$$\begin{aligned} g(\vec{x}) = &\left(\lambda_{10}\mathbb{P}(Y=1|\vec{x}) + \lambda_{00}\mathbb{P}(Y=0|\vec{x})\right) \\ &- \left(\lambda_{11}\mathbb{P}(Y=1|\vec{x}) + \lambda_{01}\mathbb{P}(Y=0|\vec{x})\right) \end{aligned}$$

cela permet de définir la fonction de décision (voir cours 2) :

$$\hat{y} = f(\vec{x}) = \begin{cases} 0 & \text{si } g(\vec{x}) \leq 0 \\ 1 & \text{si } g(\vec{x}) > 0. \end{cases}$$

ou, dans le cas multi-classe :

$$g_k(\vec{x}) = -\sum_{c=1}^{C} \lambda_{ck}\mathbb{P}(Y=c|\vec{x}).$$

qu'on utilise dans la fonction de décision multi-classe (voir cours 2) :

$$f(\vec{x}) = \arg\max_{k=1,\ldots,C} g_k(\vec{x}).$$

Règle de décision dans le cas du coût 0/1

Dans le cas du coût 0/1, la fonction discriminante vaut

$$f(\vec{x}) = \mathbb{P}(Y = 1|\vec{x}) - \mathbb{P}(Y = 0|\vec{x})$$

et la règle de décision de Bayes est bien équivalente à la décision par maximum a posteriori vue auparavant :

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) > \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \\ 0 & \text{sinon} \end{cases}$$

Résumé : Règles de décision

Rapport de vraisemblance : $\Lambda(\vec{x}) = \frac{\mathbb{P}(\vec{x}|Y=1)}{\mathbb{P}(\vec{x}|Y=0)}$

3. Règle de décision de Bayes :

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) > \frac{(\lambda_{01}-\lambda_{00})\mathbb{P}(Y=0)}{(\lambda_{10}-\lambda_{11})\mathbb{P}(Y=1)} \\ 0 & \text{sinon.} \end{cases}$$

$\rightarrow$ un coût est associé à chaque décision

2. Règle de décision par maximum a posteriori

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) > \frac{\mathbb{P}(Y=0)}{\mathbb{P}(Y=1)} \\ 0 & \text{sinon.} \end{cases}$$

$\Leftrightarrow$ Bayes avec coût 0/1 associé à chaque décision
$\rightarrow$ tient compte de la distribution de probabilités a priori des étiquettes

1. Règle de décision par maximum vraisemblance

$$\hat{y} = \begin{cases} 1 & \text{si } \Lambda(\vec{x}) > 1 \\ 0 & \text{sinon.} \end{cases}$$

$\Leftrightarrow$ Décision par maximum a posteriori avec distribution a priori des étiquettes équiprobables

Partie 3 : Estimation de densités

Modélisation paramétrique

Jusque là, nous avons considéré que $\mathbb{P}(X|Y)$ était donnée, mais... ce n'est pas toujours le cas :

parfois, il faut modéliser cette distribution !

Modélisation paramétrique Lorsque nous modéliserons une distribution nous la contraindrons à appartenir à une famille bien précise de lois de probabilités, avec paramètres $\vec{\theta}$ à valeurs dans un espace Θ de dimension finie

Description du problème d'estimation de densité et notation

On suppose disposer d'un échantillon :

$$\mathscr{D} = \vec{x}^1, \vec{x}^2, \ldots, \vec{x}^n,$$

n observations d'une variable aléatoire X à valeurs sur $\mathscr{X}$.

Nous supposons que la distribution de X a une forme connue, paramétrisée par le paramètre θ.

Comment estimer θ ?

Estimation par maximum de vraisemblance

Définition 4.6 (Estimateur par maximum de vraisemblance)
L'estimateur par maximum de vraisemblance (*maximum likelihood estimator* ou *MLE*) de θ est le vecteur $\hat{\theta}_{\text{MLE}}$ qui maximise la vraisemblance, autrement dit la probabilité d'observer $\mathscr{D}$ étant donné θ :

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\, \mathbb{P}\left(\mathscr{D}|\theta\right)$$

Exemple : en fonction du jeu de données $\mathscr{D}$ observé, on s'attend à un θ différent qui le modélise :

Procédure d'estimation MLE (sur la base de n observations iid)

Si l'on suppose qu'on a n observations indépendentes et identiquement distribuées (iid), on peut décomposer la vraisemblance comme :

$$\mathbb{P}(\mathscr{D}|\theta) = \prod_{i=1}^{n} \mathbb{P}\left(X = \vec{x}^i|\theta\right)$$

Pour simplifier les calculs, on choisira souvent de maximiser non pas directement la vraisemblance mais son logarithme :

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \log \mathbb{P}\left(X = \vec{x}^i|\theta\right)$$

Exemple d'estimation MLE : jeu de pile ou face

Nous modélisons l'observation "pile" ou "face" comme la réalisation d'une variable aléatoire X, définie sur l'univers $\mathscr{X} = \{0, 1\}$ (0 pour pile, 1 pour face) suivant une loi de probabilité $\mathbb{P}$.

Choix classique, la loi de Bernoulli (lancé d'une pièce avec probabilités de pile ou face non-équilibrées) :

$$\mathbb{P}(X = x) = \begin{cases} p & \text{si } x = 1 \\ (1-p) & \text{si } x = 0 \end{cases}$$

De manière équivalente, on peut écrire

$$\mathbb{P}(X = x) = p^x (1-p)^{1-x}$$

Estimation MLE pour jeu de pile ou face (suite)

On suppose que $\mathscr{D} = \{x^1, x^2, \ldots, x^n\}$ est constitué de n observations iid.

L'estimateur par maximum de vraisemblance de p est :

$$\begin{aligned}\hat{p}_{\text{MLE}} &= \underset{p\in[0,1]}{\text{argmax}} \sum_{i=1}^{n} \log \mathbb{P}\left(X = x^i | p\right) \\ &= \underset{p\in[0,1]}{\text{argmax}} \sum_{i=1}^{n} \log \left(p^{x^i} \left(1-p\right)^{1-x^i}\right) \\ &= \underset{p\in[0,1]}{\text{argmax}} \sum_{i=1}^{n} x^i \log p + \left(n - \sum_{i=1}^{n} x^i\right) \log\left(1-p\right)\end{aligned}$$

Estimation MLE pour jeu de pile ou face (suite et fin)

La fonction $L : p \mapsto \sum_{i=1}^{n} x^i \log p + \left(n - \sum_{i=1}^{n} x^i\right) \log\left(1-p\right)$ est concave, nous pouvons donc la maximiser en annulant sa dérivée :

$$\frac{\partial L}{\partial p} = \sum_{i=1}^{n} x^i \frac{1}{p} - \left(n - \sum_{i=1}^{n} x^i\right) \frac{1}{1-p}$$

ce qui nous donne

$$\left(1-\hat{p}_{\text{MLE}}\right) \sum_{i=1}^{n} x^i - \hat{p}_{\text{MLE}} \left(n - \sum_{i=1}^{n} x^i\right) = 0$$

et donc

$$\hat{p}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x^i$$

L'estimateur par maximum de vraisemblance de p est tout simplement la moyenne de l'échantillon.

Estimation MLE pour jeu de pile ou face (interprétation)

Pour trouver

$$\hat{p}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x^i$$

pour un jeu d'observation $\mathscr{D}$ on compte le nombre moyen de face :

Sensibilité, spécificité, et prévalence d'un test de dépistage

Dans quelle mesure un test rapide Corona est-il significatif ?

La fiabilité du test Corona de Roche dans le cadre d'une utilisation comme autotest a été évaluée dans des études indépendantes menées à l'Hôpital de la Charité de Berlin et à l'université d'Heidelberg[1] *avec des tests PCR comme méthode de référence. Les données obtenues à cette occasion ont étayé la décision de l'OFSP sur l'ajout du SARS-CoV-2 Rapid Antigen Test Nasal à la liste des tests validés selon les standards de dépistage à la mi-mars. Dans le cadre de ces études, le test a atteint une sensibilité globale de 82.5% et une spécificité de 100%. Ces études ont par ailleurs montré qu'un autotest coronavirus permet d'identifier les personnes présentant une charge virale élevée de manière fiable, autant qu'un test rapide coronavirus effectué par un personnel professionnel (sensibilité de 96.6% dans les deux groupes).* `https://diagnostics.roche.com/ch/fr/article-listing/sars-cov-2-rapid-antigen-test-nasal-self-testing.html`

Estimation MLE des paramètres d'une loi aléatoire pour modéliser un test de dépistage du cancer du col de l'utérus

Dans l'exemple décrit plus haut, nous avons considéré connues $\mathbb{P}(X|Y=0)$ et $\mathbb{P}(X|Y=1)$ (vraisemblance d'un résultat de test sachant la classe). Comment les estimer de manière expérimentale ?

Ingrédients (choses qu'on a à disposition) :

- un jeu $\mathscr{D}_0$ de n_0 personnes non-atteintes, parmi lesquelles t_0 ont un test négatif
- un jeu $\mathscr{D}_1$ de n_1 personnes atteintes, parmi lesquelles t_1 ont un test positif
- on sait que la prévalence de la maladie est $\mathbb{P}(Y=1) = p_r$

Estimation de la vraisemblance (suite)

Modèle choisi pour modéliser les vraisemblances (modélise le manque de fiabilité des tests) :

- $\mathbb{P}(X|Y=0) \sim$ Bernoulli paramètre p_0
- $\mathbb{P}(X|Y=1) \sim$ Bernoulli paramètre p_1

Problème d'estimation revient donc à trouver p_0 et p_1.

La loi de Bayes nous dit que la probabilité qu'une personne dont le test est positif soit atteinte est :

$$\mathbb{P}(Y=1|X=1) = \frac{\mathbb{P}(X=1|Y=1)\,\mathbb{P}(Y=1)}{\mathbb{P}(X=1)}$$

Par le choix de notre modèle de Bernoulli, nous avons

$$\mathbb{P}(X=x|Y=0) = p_0^x\,(1-p_0)^{1-x} \text{ et}$$
$$\mathbb{P}(X=x|Y=1) = p_1^x\,(1-p_1)^{1-x}\text{, ainsi,}$$

Estimation MLE test de dépistage (suite et fin)

Avec les lois du slide précédent, nous avons :

$$\mathbb{P}(Y = 1|X = 1) = \frac{p_1 p_r}{p_1 p_r + p_0 (1 - p_r)}$$

En remplaçant p_0 and p_1 par leurs estimateurs par maximum de vraisemblance (moyenne de l'échantillon)

$$\hat{p}_0 = \hat{p}_{0,\text{MLE}} = 1 - \frac{t_0}{n_0} \quad \text{spécificité estimée du test}$$

$$\hat{p}_1 = \hat{p}_{1,\text{MLE}} = \frac{t_1}{n_1} \quad \text{sensibilité estimée du test}$$

(application numérique : $\frac{t_0}{n_0} = 0.98$, $\frac{t_1}{n_1} = 0.70$, et $p_r = 10^{-5}$)

Estimateur de Bayes

Point de départ : On suppose que la valeur du paramètre θ qui caractérise notre modèle n'est pas complètement inconnue.

On suppose, par exemple, qu'en tant qu'expert·e·s du domaine d'application, on a une bonne idée des valeurs qu'il peut prendre.

But : utiliser cette information qu'on a à bon escient, par exemple afin de palier à un nombre d'observations faible.

Approche : nous allons modéliser θ à son tour comme une variable aléatoire, et définir sa distribution a priori $\mathbb{P}(\theta)$.

Estimateur de Bayes : définition

Définition 4.7 (Estimateur de Bayes) Étant donnée une fonction de coût L, l'estimateur de Bayes $\hat{\theta}_{\text{Bayes}}$ de θ est défini par

$$\hat{\theta}_{\text{Bayes}} = \underset{\hat{\theta}}{\operatorname{argmin}}\, E\left[L\left(\theta, \hat{\theta}\right)\right].$$

Si l'on utilise pour L l'erreur quadratique moyenne, alors :

$$\hat{\theta}_{\text{Bayes}} = \underset{\hat{\theta}}{\operatorname{argmin}}\, \mathbb{E}\left[\left(\theta - \hat{\theta}\right)^2\right]$$

Estimateur de Bayes (lorsque le coût est quadratique)

Si on considère $\hat{\theta}$ déterministe et un coût quadratique, nous avons :

$$\begin{aligned}
\hat{\theta}_{\text{Bayes}} &= \underset{\hat{\theta}}{\operatorname{argmin}}\, \mathbb{E}\left[\left(\theta - \hat{\theta}\right)^2\right] \\
&= \underset{\hat{\theta}}{\operatorname{argmin}}\, \hat{\theta}^2 - 2\hat{\theta}\mathbb{E}[\theta] + \mathbb{E}[\theta^2] \\
&= \underset{\hat{\theta}}{\operatorname{argmin}}\, \underbrace{\left(\hat{\theta} - \mathbb{E}[\theta]\right)^2}_{\text{min. quand } \hat{\theta}=\mathbb{E}[\theta]} \underbrace{-\mathbb{E}[\theta]^2 + \mathbb{E}[\theta^2]}_{\text{ne dépendent pas de } \hat{\theta}} \\
&= \mathbb{E}[\theta]
\end{aligned}$$

Cette espérance est prise sur la distribution de θ et de X (distribution a posteriori de θ), qui nous sert à estimer θ. Ainsi :

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}\left[\theta|X\right] = \int \theta\, \mathbb{P}\left(\theta|X\right) \mathrm{d}\theta$$

Note : Quand la distribution a priori du paramètre est uniforme, l'estimateur de Bayes est équivalent à l'estimateur par maximum de vraisemblance.

Illustration : estimateur de Bayes des paramètres du test de dépistage

On considère que la valeur des deux paramètres p_0 et p_1 des lois Bernoulli qui régissent les résultats des tests dans notre modèle, sont des réalisation d'une variable aléatoire de type Bêta :

$$p_0 \sim \text{Beta}(\alpha_0, \beta_0) \quad p_1 \sim \text{Beta}(\alpha_1, \beta_1)$$

Loi de densité de la loi Bêta (définie pour $\alpha, \beta > 0$, et $0 \leq u \leq 1$) :

$$f_{\alpha,\beta}(u) = \frac{u^{\alpha-1}(1-u)^{\beta-1}}{B(\alpha,\beta)} \text{ avec } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Commençons par p_0 : Pour reprendre les notations générales, on considère que p_0 est le θ et pour calculer son estimateur de Bayes, $\tilde{p}_0 = \hat{p}_{0,\text{Bayes}} = \hat{\theta}_{\text{Bayes}}$, il nous faut connaître la loi

$$\mathbb{P}(\theta|X), \text{ qui est, dans ce cas : } \mathbb{P}(p_0|\mathscr{D}_0)$$

(Rappel : $\mathscr{D}_0$ est le jeu de données composé de n_0 personnes non-atteintes, parmi lesquelles t_0 ont un test négatif)

Calcul de la probabilité qu'un paramètre p_0 ait généré les données mesurées $\mathscr{D}_0$

Sous-problème : on a mesuré le jeu de données $\mathscr{D}_0$ (avec des résultats de test positifs et négatifs, dans une population non atteinte), $\mathbb{P}(p_0|\mathscr{D}_0)$ représente la probabilité que p_0 était le paramètre de la loi de Bernoulli qui a produit les résultats du test de dépistage $\mathscr{D}_0$, où l'on a observé, dans une population non-atteinte de n_0 personnes, t_0 test négatifs et $n_0 - t_0$ tests positifs.

Note importante : il faut bien noter ici que tout jeu mesuré experimentalement pourrait être le résultat d'un générateur aléatoire avec n'importe quel p_0 donné (par exemple : population n'est pas atteinte, tous les tests sont positifs alors que $p_0 = 0.5$) *mais ça serait très improbable !* C'est précisément cette probabilité qu'on cherche à déterminer ici, afin qu'on puisse assigner le p_0 qui serait le plus à même d'expliquer (le plus vraisemblable) les mesures.

Calculons cette probabilité que p_0 soit le bon paramètre

Par la loi de Bayes, on a :

$$\mathbb{P}(p_0|\mathscr{D}_0) = \frac{\mathbb{P}(\mathscr{D}_0|p_0)\,\mathbb{P}(p_0)}{\mathbb{P}(\mathscr{D}_0)}.$$

avec :

$$\mathbb{P}(p_0) = \frac{p_0^{\alpha_0-1}(1-p_0)^{\beta_0-1}}{B(\alpha_0,\beta_0)}$$

la probabilité que le paramètre soit p_0, supposant qu'il suit une loi Bêta(α_0,β_0)

$$\mathbb{P}(\mathscr{D}_0|p_0) = \prod_{i=1}^{n_0} p_0^{x^i}(1-p_0)^{1-x^i}$$

la probabilité a priori que si le paramètre était p_0 on ait mesuré $\mathscr{D}_0$: on suppose que chaque résultat de dépistage x^i dans $\mathscr{D}_0$ est iid $\sim$ Bernoulli (p_0)

$= \ldots$ (voir page suivante)

Calcul de la probabilité a priori

$$\ldots = \mathbb{P}(\mathscr{D}_0|p_0) = \prod_{i=1}^{n_0} \underbrace{p_0^{x^i}}_{\text{avec } x^i=1 \text{ dans } n_0-t_0 \text{ cas}} \underbrace{(1-p_0)^{1-x^i}}_{\text{avec } x^i=0 \text{ dans } t_0 \text{ cas}}$$
$$= p_0^{n_0-t_0}(1-p_0)^{t_0}$$

on obtient donc :

$$\mathbb{P}(p_0|\mathscr{D}_0) = \frac{\mathbb{P}(\mathscr{D}_0|p_0)\,\mathbb{P}(p_0)}{\mathbb{P}(\mathscr{D}_0)}$$
$$= \underbrace{\frac{1}{\mathbb{P}(\mathscr{D}_0)\,B(\alpha_0,\beta_0)}}_{\text{constante de normalisation}} p_0^{n_0-t_0+\alpha_0-1}(1-p_0)^{t_0+\beta_0-1}$$
$$= f_{(n_0-t_0+\alpha_0),(t_0+\beta_0)}(p_0)$$

on voit que $\mathbb{P}(p_0|\mathscr{D}_0)$ suit une distribution beta de paramètres $(n_0-t_0+\alpha_0),(t_0+\beta_0)$.

Note : la nouvelle constante de normalisation garantit qu'on a bien une intégrale unité.

Estimateur de Bayes des paramètres qui modèlisent le test de dépistage

. . . et en injectant ces resultat dans la formule de l'estimateur de Bayes, on obtient finalement :

$$\begin{aligned}\tilde{p}_0 = \hat{p}_{0,\text{Bayes}} = \hat{\theta}_{\text{Bayes}} &= \mathbb{E}\left[\theta|X\right] \\ &= \mathbb{E}\left[p_0|\mathscr{D}_0\right] = \int \theta\, \mathbb{P}\left(p_0|\mathscr{D}_0\right) \mathrm{d}p_0 \\ &= \frac{n_0 - t_0 + \alpha_0}{(n_0 - t_0 + \alpha_0) + (t_0 + \beta_0)} \\ &= \frac{n_0 - t_0 + \alpha_0}{n_0 + \alpha_0 + \beta_0}\end{aligned}$$

où l'on a utilisé le fait que l'espérance d'une loi bêta est $\frac{\alpha}{\alpha+\beta}$.

Note : on peut faire une dérivation similaire pour $\tilde{p}_1$.

Comparaison de l'estimateur de Bayes avec l'estimateur maximum de vraisemblance

Rappel : l'estimateur par maximum de vraisemblance était :

$$\hat{p}_0 = 1 - \frac{t_0}{n_0}$$

cela nous permet d'écrire l'estimateur de Bayes comme suit :

$$\tilde{p}_0 = \frac{n_0}{n_0 + \alpha_0 + \beta_0}\hat{p}_0 + \frac{\alpha_0 + \beta_0}{n_0 + \alpha_0 + \beta_0}\frac{\alpha_0}{\alpha_0 + \beta_0}$$

L'estimateur de Bayes est adapté, quelle que soit la taille de $\mathscr{D}_0$:

- si n_0 est grand l'estimateur de Bayes $\tilde{p}_0$ est proche de l'estimateur par maximum de vraisemblance $\hat{p}_0$
- si n_0 est petit l'estimateur de Bayes $\tilde{p}_0$ est proche de $\frac{\alpha_0}{\alpha_0+\beta_0}$ = l'espérance de la distribution a priori sur p_0.

$\Rightarrow$ plus on a données, plus on leur fait confiance et plus on peut potentiellement s'éloigner de l'espérance a priori du paramètre (dont on restera proche avec peu de données).

Résumé du cours 10

1. Formalisation du concept de classe : on s'appuie su rla loi de Bayes
2. Définition des règles de décision sur la base de tests de rapport de vraisemblance (par ordre croissant de généralité) :
   - 2.1 décision par maximum de vraisemblance
   - 2.2 décision par maximum a posteriori
   - 2.3 décision par minimisation du risque de Bayes
3. Dérivation de 2 techniques d'estimation de densités de probabilité
   - 3.1 par maximum de vraisemblance (MLE : maximum likelihood estimator)
   - 3.2 par estimateur de Bayes (plus grande flexibilité)

Guide de lecture pour ce cours

Chloé-Agathe Azencott "Introduction au Machine Learning",
Dunod, 2019, ISBN 978-210-080153-4
Chapitre 4 : Inférence bayésienne