1 – Connaître le passé pour prédire l'avenir

Une base de données a été constituée en rassemblant des informations relevées sur un nombre N de patients.

Ces données, recueillies sur les patients passés, peuvent être utilisées pour aider à poser un diagnostic sur les patients futurs.

❧ Ces données sont $n = 6$ données biomécaniques :
- x_0 : angle d'incidence du bassin (en degrés)
- x_1 : angle d'orientation du bassin (en degrés)
- x_2 : angle de lordose lombaire (en degrés)
- x_3 : pente du sacrum (en degrés)
- x_4 : rayon du bassin (en millimètres)
- x_5 : distance de glissement de spondylolisthésis (en millimètres)

Elles sont représentées par des flottants au sein d'un tableau de N lignes et n colonnes, nommé `data`, qui peut être compris comme une matrice

$$(x_{i,k})_{\substack{0\leqslant i<N\\0\leqslant k<n}}.$$

❧ La base contient également une donnée qualitative décrivant l'état de santé du patient, codée par un entier au sein d'un vecteur $(y_i)_{0\leqslant i<N}$ de taille N, nommé `etat`.
- $y = 0$: état normal
- $y = 1$: hernie discale
- $y = 2$: spondylolisthésis

❧ Les données biomécaniques $(x_{i,k})_{0\leqslant k<n}$ du patient $0 \leqslant i < N$ figurent donc sur la ligne `data[i,:]` et son état de santé est codé par y_i = `etat[i]`.

❧ Le traitement statistique des données repose sur un modèle probabiliste : on considère des variables aléatoires

$$(X_0, \dots, X_{n-1}, Y)$$

définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P})$.

Chaque élément ω de Ω est censé représenter un individu ; pour $0 \leqslant k < n$, le réel $X_k(\omega) = x_k$ donne alors la valeur de la k-ième donnée biomécanique de cet individu et l'entier $Y(\omega) = y \in \{0;1;2\}$ code l'état de santé de cet individu.

2 – Modèle discret

Dans un premier temps, on suppose pour simplifier (et rester dans le cadre du programme de mathématiques !) que les variables aléatoires X_k sont **discrètes** et qu'elles prennent donc leurs valeurs dans un ensemble fini ou dénombrable $E \subset \mathbb{R}$.

Pour que notre modèle probabiliste soit utilisable, il faut connaître la loi conjointe des variables aléatoires, c'est-à-dire la famille

$$\big(\mathbf{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, Y = y)\big)_{(x_0,\dots,x_{n-1},y)\in E^n\times\{0;1;2\}}. \tag{1}$$

► Hypothèse naïve bayésienne

Par définition des probabilités conditionnelles,

$$\mathbf{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, Y = y) = \mathbf{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1} \mid Y = y)\,\mathbf{P}(Y = y). \tag{2}$$

L'**hypothèse naïve bayésienne** suppose en fait que

$$\mathbf{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1}, Y = y) = \left(\prod_{0\leqslant k<n} \mathbf{P}(X_k = x_k \mid Y = y)\right) \cdot \mathbf{P}(Y = y). \tag{3}$$

Cette hypothèse signifie que, sachant l'état de santé d'une personne (= en conditionnant par $[Y = y]$), les grandeurs biomécaniques décrites par les variables aléatoires $X_0, \dots, X_{n-1}$ varient indépendamment les unes des autres.

❧ Une telle hypothèse simplificatrice est raisonnable dès lors qu'on ne connaît aucune corrélation avérée entre les différentes grandeurs biomécaniques mesurées.

► **Apprentissage**

Les données enregistrées dans le tableau `data` et le vecteur `etat` vont alors nous permettre d'estimer
- la loi de Y,
- la loi conjointe de $(X_0, \dots, X_{n-1})$
- et les lois conditionnelles des X_k sachant $[Y = y]$ pour les différentes valeurs de $y \in \{0; 1; 2\}$.

❧ Plus précisément, pour tout $y \in \{0; 1; 2\}$, on note

$$I_y = \big\{0 \leqslant i < N \ : \ \texttt{etat[i]} = y\big\} \qquad \text{et} \qquad I = I_0 \sqcup I_1 \sqcup I_2 = \{0; 1; \dots; N-1\}. \tag{4}$$

Pour $y \in \{0; 1; 2\}$, on note N_y, le cardinal de I_y, de telle sorte que $N_0 + N_1 + N_2 = N$.

Selon la Loi des grands nombres, si les nombres N_y sont "assez grands", il est raisonnable de considérer que

$$\forall\, y \in \{0; 1; 2\}, \qquad \mathbf{P}(Y = y) \approx \frac{\#(I_y)}{\#(I)} = \frac{N_y}{N} \tag{5}$$

ainsi que, quels que soient $(x_0, \dots, x_{n-1}) \in E^n$,

$$\mathbf{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1}) \approx \frac{\#\big\{0 \leqslant i < N \ : \ \texttt{data[i,0]} = x_0, \dots, \texttt{data[i,n-1]} = x_{n-1}\big\}}{N} \tag{6}$$

et enfin que

$$\forall\, y \in \{0; 1; 2\},\ \forall\, x_k \in E, \qquad \mathbf{P}(X_k = x_k \mid Y = y) \approx \frac{\#\big\{0 \leqslant i < N \ : \ \texttt{etat[i]} = y \quad \text{et} \quad \texttt{data[i,k]} = x_k\big\}}{N_y}. \tag{7}$$

❧ J'insiste ! Invoquer la Loi des grands nombres n'est raisonnable que si les trois ensembles I_0, I_1 et I_2 sont assez nombreux...

► **Prédiction**

Comment ce modèle probabiliste peut-il servir d'aide au diagnostic ?

Pour un individu dont les données biomécaniques mesurées sont x_0, …, x_{n-1}, on peut maintenant déduire de ce qui précède une estimation de la probabilité conditionnelle des événements $[Y = 0]$, $[Y = 1]$ et $[Y = 2]$ connaissant les données biomécaniques mesurées sur le patient :

$$\mathbf{P}(Y = y \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1}) = \frac{\mathbf{P}(Y = y)}{\mathbf{P}(X_0 = x_0, \dots, X_{n-1} = x_{n-1})} \cdot \prod_{0 \leqslant k < n} \mathbf{P}(X_k = x_k \mid Y = y) \tag{8}$$

(en appliquant la Formule de Bayes et l'hypothèse naïve bayésienne (3)).

On prédit alors que l'état du patient est **vraisemblablement** la valeur y pour laquelle la probabilité conditionnelle

$$\mathbf{P}(Y = y \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1})$$

est maximale.

❧ On prendra soin de ne pas conclure de manière tranchée si la probabilité conditionnelle maximale n'est pas nettement supérieure aux deux autres probabilités conditionnelles !

❧ Dans l'expression (8), le dénominateur est toujours le même. Il n'est donc pas utile de le calculer pour comparer entre elles les probabilités conditionnelles.

3 – Modèle continu

La loi d'une **variable aléatoire continue** X (par opposition aux variables aléatoires discrètes étudiées précédemment) est caractérisée par une **densité** f_X, c'est-à-dire une fonction positive, intégrable sur $]-\infty, +\infty[$ et telle que

$$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1$$

et, par définition,

$$\forall\, a < b, \qquad \mathbf{P}(X \in [a, b]) = \int_a^b f_X(x)\, dx.$$

❧ Si les variables aléatoires X_0, …, X_{n-1} sont continues, alors la probabilité conditionnelle

$$\mathbf{P}(Y = y \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1})$$

n'a plus de sens (car on conditionne par une intersection d'événements négligeables).

► **Vraisemblance**

Revenons à notre modèle $(X_0, \dots, X_{n-1}, Y)$.

❧ On suppose ici que notre modèle est continu (et non plus discret), donc que la loi conditionnelle de X_k sachant $[Y = y]$ est caractérisée par une densité notée $[x \mapsto f_k(x \mid y)]$.

❧ L'hypothèse naïve bayésienne nous dit alors que la loi conditionnelle de $(X_0, \dots, X_{n-1})$ sachant que l'événement $[Y = y]$ est réalisé est caractérisée par la **vraisemblance**

$$\forall\, (x_0, \dots, x_{n-1}) \in \mathbb{R}^n, \qquad L(x_0, \dots, x_{n-1} \mid y) = \left(\prod_{0 \leqslant k < n} f_k(x_k \mid y) \right) \tag{9}$$

(L *stands for Likelihood*) au sens où

$$\mathbf{P}(X_0 \in [a_0, b_0], \dots, X_{n-1} \in [a_{n-1}, b_{n-1}] \mid Y = y) = \prod_{0 \leqslant k < n} \int_{a_k}^{b_k} f_k(x \mid y)\, dx. \tag{10}$$

❧ Par analogie avec le modèle discret étudié plus haut, on cherche maintenant pour quelle valeur de $y \in \{0; 1; 2\}$ le produit

$$L(x_0, \dots, x_{n-1} \mid y) \cdot \mathbf{P}(Y = y) \tag{11}$$

est maximal : c'est cette valeur y qui décrit le plus vraisemblablement l'état de santé du patient (en supposant comme on l'a fait plus haut que la quantité maximale soit assez nettement supérieure aux autres).

► Loi normale des erreurs

❧ Dans la mesure où on n'a aucune information particulière sur la manière dont sont réparties les valeurs prises par une variable aléatoire continue X, il est raisonnable de penser qu'une densité de X est de la forme

$$\forall\, x \in \mathbb{R}, \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp \frac{-(x - m)^2}{2\sigma^2}. \tag{12}$$

Les paramètres $m \in \mathbb{R}$ et $\sigma^2 > 0$ représentent respectivement l'espérance et la variance de la variable aléatoire X.

Un théorème indique en effet que cette densité, dite **densité de la loi normale des erreurs**, joue un rôle central en théorie des probabilités (Théorème-limite central).

❧ Nous supposerons dorénavant que les densités conditionnelles $f_k(\cdot \mid y)$ qui apparaissent dans l'expression de la vraisemblance $L(\cdot \mid y)$ sont toutes de cette forme.

► Estimation des paramètres de la vraisemblance

Les paramètres des différentes densités normales $f_k(\cdot \mid y)$ figurant dans l'expression de la vraisemblance $L(\cdot \mid y)$ peuvent être estimés à partir des données contenues dans le tableau `data`.

❧ On rappelle (4) que I_y désigne l'ensemble des individus $0 \leqslant i < N$ dont l'état est codé par y, c'est-à-dire tels que `etat[i]` $= y$, et que $N_y = \#(I_y)$.

❧ La moyenne $m_{k,y}$ de la densité $f_k(\cdot \mid y)$ de X_k sachant que $[Y = y]$ est estimée, selon la Loi des grands nombres, par la **moyenne empirique** :

$$m_{k,y} \approx \langle x_k \rangle_y \overset{\text{déf.}}{=} \frac{1}{N_y} \sum_{i \in I_y} x_{i,k}. \tag{13}$$

On reconnaît ici la moyenne des valeurs $x_{i,k}$ calculée en se restreignant aux individus i pour lesquels $y_i = y$.

❧ La variance $\sigma^2_{k,y}$ de cette densité $f_k(\cdot \mid y)$ sachant que $[Y = y]$ est estimée, de la même manière, par la **variance empirique** :

$$\sigma^2_{k,y} \approx \frac{1}{N_y} \sum_{i \in I_y} \big(x_{i,k} - \langle x_k \rangle_y\big)^2. \tag{14}$$

❧ Comme on l'a déjà fait remarquer plus haut, ces approximations ne sont des estimations raisonnables des paramètres m et σ^2 que si les entiers N_y sont "assez grands".

4 – Questions de code

► Q17.

Écrire deux fonctions `moyenne(x)` et `variance(x)` de complexité linéaire et retournant la moyenne empirique et la variance empirique calculées sur un vecteur `x` de taille quelconque.

► Q18.

Écrire une fonction `synthese(data, etat)` qui retourne un tableau

$$\big((m_{k,y},\ \sigma^2_{k,y})\big)_{\substack{0 \leqslant y < 3 \\ 0 \leqslant k < n}}$$

contenant les couples (moyenne empirique, variance empirique) calculés en appliquant les formules (13) et (14) pour chacun des n attributs du tableau `data` et chacun des 3 états possibles.

On supposera connue une fonction `separationParGroupes(data, etat)`, qui retourne une liste de trois tableaux de tailles respectives (N_0, n), (N_1, n) et (N_2, n). Ces trois tableaux contiennent les données du tableau `data` regroupées en fonction de l'état du patient.

► **Q19.**

Écrire une fonction `gaussienne(a, moy, var)` qui retourne l'expression (12) de la densité $f_X(a)$ pour une variable aléatoire X d'espérance $m = \texttt{moy}$ et de variance $\sigma^2 = \texttt{var}$.

► **Q20.**

Les données biomécaniques $(x_k)_{0 \leqslant k < n}$ d'un individu sont regroupées dans un vecteur `z`. Écrire une fonction `probabiliteGroupe(z, data, etat)` qui retourne le triplet des valeurs

$$\big(L(x_0, \ldots, x_{n-1} \mid y).\mathbf{P}(Y = y)\big)_{0 \leqslant y < 3}.$$

► **Q21.**

Pour attribuer un groupe à un individu, il faut déterminer la valeur maximale du triplet précédent.

Écrire une fonction `prediction(...)`, dont on précisera les arguments, qui renvoie le numéro du groupe auquel appartient l'individu dont les données sont regroupées dans le vecteur `z`.

► **Q23.**

On teste l'algorithme en l'appliquant aux données présentes dans le tableau `data` (qui a servi lors de l'apprentissage). On en déduit la **matrice de confusion**

$$C = \begin{pmatrix} 23 & 9 & 8 \\ 9 & 10 & 1 \\ 10 & 1 & 49 \end{pmatrix}$$

où $C_{i,j}$ représente le nombre d'individus dont l'état réel est i et l'état prédit par l'algorithme bayésien est j.

Que vaut N ? Quel est le taux de succès de l'algorithme bayésien ?

5 – Réponses aux questions

► **Q17.**

Calcul de la moyenne : on calcule la somme des termes et on divise par le nombre de termes. La complexité est évidemment linéaire (on effectue n additions et une division).

```
def moyenne(x):
    somme = 0
    for v in x:
        somme += v
    return v/len(x)
```

Calcul de la variance : on calcule *une fois* la moyenne du vecteur, puis la variance. Le calcul de la moyenne est de complexité linéaire. La suite du calcul demande n soustractions, n multiplications, n additions et une division : la complexité de l'ensemble est bien linéaire.

```
def variance(x):
    moy = moyenne(x)
    somme_carres = 0
    for v in x:
        somme_carres += (v-moy)**2
    return somme_carres/len(x)
```

► **Q18.**

On découpe le tableau `data` en groupes et on applique les fonctions précédentes sur chacun des groupes.

```
def synthese(data, etat):
    N, n = data.shape
    groupes = separationParGroupes(data, etat)
    nb_groupes = len(groupes)    #   3 groupes
    liste_moy_var = []
    for y in range(nb_groupes):
        liste_y = []
        for k in range(n):
            colonne = groupes[y][:,k]
            mu = moyenne(colonne)
            sigma2 = variance(colonne)
            liste_y.append([mu, sigma2])
         liste_moy_var.append(liste_y)
     return liste_moy_var
```

► **Q19.**

On applique la formule...

```
def gaussienne(a, moy, var):
    return exp(-(a-moy)**2/(2*var))/sqrt(2*pi*var)
```

► **Q20.**

Pour calculer le numérateur de (11), il faut classer les données par groupes avec `separationParGroupes`, en déduire le cardinal N_y de chaque groupe pour estimer $\mathbf{P}(Y = y)$ avec (5), estimer les paramètres m et σ^2 avec `synthese` et appliquer `gaussienne` pour finir.

```
def probabiliteParGroupes(z, data, etat):
    groupes = separationParGroupes(data, etat)
    nb_etats, n = groupes.shape
    N = len(data)
    L_synth = synthese(data, etat)
    L_prod = []
    for y in range(nb_etats):
        Ny = len(groupes[y])
        produit = Ny/N
        L_synth_y = L_synth[y]
        for k in range(n):
            moy, var = L_synth_y[k]
            produit *= gaussienne(z[k], moy, var)
         L_prod.append(produit)
     return array(L_prod)
```

On notera qu'on exécute *deux fois* la fonction `separationParGroupes` : une première fois de manière explicite, une seconde fois de manière implicite lors de l'exécution de `synthese`. C'est mal !

► **Q21.**

On cherche l'indice du maximum d'une liste, sans se préoccuper ici de vérifier si le maximum est nettement supérieur aux autres valeurs de la liste.

```
def prediction(a, data, etat):
    L_prod = probabiliteParGroupes(z, data, etat)
    nb_etats = len(L_prod)
    etat_max = 0
    prod_ref = L_prod[0]
    for i in range(1, nb_etats):
        prod_i = L_prod[i]
        if prod_i>prod_ref:
            etat_max, prod_ref = i, prod_i
     return etat_max
```

► **Q23.**

Le nombre N est le nombre de lignes du tableau `data`, c'est-à-dire le nombre de patients qui ont servi lors de l'apprentissage. C'est donc la somme des coefficients de la matrice de confusion : $N = 120$.

Le nombre de succès est le nombre de patients pour lesquels l'état prédit j par l'algorithme bayésien correspond à l'état réel i. C'est donc la somme des coefficients diagonaux de la matrice de confusion. Le taux de succès est donc égal à $82/120$, soit environ 69%.