Mathématiques pour l'intelligence artificielle
Cours 3

Raphaël BEN ZAKOUR / Yoann DELORD

Septembre 2022

M2 Informatique S3

Professeur : Thomas Schatz

Faculté des Sciences
Aix*Marseille Université

Table des matières

1 Algèbre linéaire

1.1 Espaces associés à une matrice (Suite)

SVD et matrices orthogonales

Soit $Q \in \mathbb{R}^{n \times n}$ une matrice orthognale, il existe une factorisation telle que :

$$Q = U\Sigma V^T \tag{1}$$

Avec U, V des matrices orthogonales non uniques et Σ une matrice diagonale unique. Or nous pouvons écrire Q comme $Q = Q \times I_n \times I_n = I_n \times I_n \times Q$. Comme I_n et Q sont orthognales, par identification nous pouvons voir que les valeurs singulières d'une matrice orthogonale de taille n sont toutes égales à 1.

Comprendre une transformation linéaire

Nous pouvons nous demander à quoi correspond une transformation linéaire quelconque en regardant seulement sa matrice.
Prenons la matrice A suivante :

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

A première vue, il est compliqué d'en tirer quoi que ce soit sur la transformation qu'elle effectuerait sur un vecteur quelconque. Regardons maintenant sa décomposition approximative en valeurs singulières.

$$A = U\Sigma V^T \text{avec}$$

$$U \# \begin{pmatrix} -0,21 & -0,89 & 0,41 \\ -0,52 & -0,25 & -0,82 \\ -0,83 & 0,39 & 0,41 \end{pmatrix} \Sigma \# \begin{pmatrix} 16,85 & 0 & 0 \\ 0 & 1,07 & 0 \\ 0 & 0 & 0 \end{pmatrix} V \# \begin{pmatrix} -0,48 & -0,57 & -0,66 \\ 0,78 & 0,08 & -0,62 \\ 0,47 & -0,82 & 0,41 \end{pmatrix}$$

Soient $\lambda_1 = 16.85, \lambda_2 = 1.07$ et $\lambda_3 = 0$ les valeurs singulières approchées de A.

Nous pouvons lire de la décomposition les élements suivants :

- Il y a 2 valeurs singulières non nulles donc d'une part le rang de A est 2, et d'autre part la transformation effectuée par la matrice A se réalise dans seulement 2 dimensions.
- L'amplitude de la transformation est proportionnelle aux valeurs singulières. Or λ_1 est plus grande que λ_2 d'un facteur 16. Ce qui signifie qu'A provoque une transformation 16 fois plus importante dans une dimension que dans la seconde.
- Cette transformation se fait dans des directions particulières données par la matrice V. La direction associée à λ_1 est donnée par le vecteur v_1 (premier vecteur colonne de la matrice V) et v_2 pour λ_2. C'est à dire qu'A projette le vecteur d'entrée sur les vecteurs $\lambda_1 \cdot v_1$ et $\lambda_2 \cdot v_2$, d'où l'étirement 16 fois plus important dans la direction v_1.
- Le résultat de la transformation par A d'un vecteur quelconque est exprimé dans le système de coordonnées donné par la matrice U

Pour faire le lien avec le schéma sur les espaces associés à une matrice :

$\lambda_1, \lambda_2 \neq 0 \iff u_1, u_2$ forment une base de l'image de A et v_1 v_2 une base de la co-image de A^T.

$\lambda_3 = 0 \iff u_3$ forme une base du co-noyau de A^T et v_3 une base du noyau de A.

Pour rappel :

- L'espace défini par l'image de A est complémentaire et orthogonale à l'espace défini par le co-noyau de A^T.
- L'espace défini par la co-image de A^T est complémentaire et orthogonale à l'espace défini par le noyau de A.

Les espaces sont complémentaire car leur union engendre l'espace, ici $\mathbb{R}^3$, et sont orthogonaux car leur intersection est vide.

Démonstration du processus de Gram-Schmidt

Nous allons montrer par récurrence le processus de Gram-Schmidt qui permet à partir une famille de k vecteurs indépendants $(a_1, a_2, \ldots, a_k) \in E^k$, de trouver une famille de vecteurs orthonormaux telle que pour tout $r \leq k$, $(q_1, q_2, \ldots, q_r)$ soit une base orthonormale de $\text{Vect}(a_1, a_2, \ldots, a_r)$.

Soit $P(k)$ la propriété telle que $\forall\, i \in \{1, \ldots, k\}, j \in \{1, \ldots, k\}$, si $i \neq j$ alors $< q_i | q_j >= 0$.

Pour $k = 1$ la propriété est triviale car la valeur logique d'une proposition où il n 'y a rien à satisfaire est vraie, en effet pour $k = 1$ il n'existe pas de pairs (q_i, q_j) telle que i soit différent de j.

Supposons maintenant que la propriété soit vraie au rang k, montrons qu'elle l'est aussi au rang $k + 1$. Pour ce faire nous avons avoir besoin de la propriété suivante du produit scalaire :

$$< au + bv | w >= a < u | w > + b < v | w > \text{ (propriété de bilinéarité)}$$

Avec a, b des réels et u, v, w des vecteurs de même dimension.
Nous supposons donc posséder k vecteurs orthogonaux q_k et nous voulons montrer que le vecteur q_{k+1} résultant du procédé de Gram-Schmidt est orthognal à n'importe quel vecteur q_r de $(q_1, \ldots, q_k)$. Déterminons le produit scalaire de q_r avec q_{k+1} :

$$\begin{aligned}
< q_r | q_{k+1} > &= < q_r | a_{k+1} - \sum_i^k < q_i | a_{k+1} > q_i > \\
&= < q_r | a_{k+1} > - \sum_i^k < q_r | < q_i | a_{k+1} > q_i > \quad \text{(bilinéarité)} \\
&= < q_r | a_{k+1} > - \sum_i^k < q_r | q_i >< q_i | a_{k+1} > \quad \text{(bilinéarité)}
\end{aligned}$$

Or $< q_r | q_i >$ est nul pour tout $i \neq r$ par hypothèse de récurrence d'où

$$\begin{aligned}
< q_r | q_{k+1} > &= < q_r | a_{k+1} > - < q_r | q_r >< q_r | a_{k+1} > \\
&= < q_r | a_{k+1} > -1 < q_r | a_{k+1} > \\
&= 0
\end{aligned}$$

Ainsi $< q_r | q_{k+1} >$ est nul donc par construction de q_{k+1}, ce vecteur est donc orthogonal à n'importe quel vecteur $q_r \in (q_1, \ldots, q_k)$.

La propriété P est vraie par récurrence.

1.2 Théorème spectral

Enoncé : Si S est une matrice **symétrique**, réelle de taille $m \times m$, alors il existe une matrice orthogonale **réelle** Q de taille $m \times m$ et une matrice diagonale **réelle** Λ de taille $m \times m$ telles que $S = Q\Lambda Q^t$.

Un vecteur propre v et une valeur propre λ d'une matrice S respectent la relation $Sv = \lambda v$. Dans le cas de la SVD, un vecteur singulier à droite v, singulier à gauche u et une valeur singulière σ d'une matrice S respectent la relation $Sv = \lambda u$

Matrice définie positive

Définition : Une matrice symétrique réelle est dite définie positive, notée $S \succ 0$ ssi pour toute matrice colonne u, $u^T S u > 0$.
Une matrice symétrique réelle est dite semi-définie positive, notée $S \succeq 0$ ssi pour toute matrice colonne u, $u^T S u \geq 0$.

Une relation d'ordre peut-être établie sur les matrices symétriques réelles. Soient S_1 et S_2 des matrices symétriques réelles, alors :

$$S_1 \prec S_2 \iff S_2 - S_1 \succ 0$$
$$S_1 \preceq S_2 \iff S_2 - S_1 \succeq 0$$

Cette relation est partielle, elle n'est pas totale car elle ne permet pas d'ordonner n'importe quel couple de matrices symétriques réelles (la différence de deux matrices peut ne pas être une matrice définie positive).

Une matrice A est définie positive ssi ses valeurs propres sont > 0 et semi définie positive ssi ≥ 0.

Puissance et exponentielle de matrice

Soit S une matrice carré qui admet la décomposition en valeurs propres suivante :

$$S = Q\Lambda Q^T$$

Le calcul de S^2 peut se faire suivant cette décomposition à savoir $S^2 = (Q\Lambda Q^T)(Q\Lambda Q^T) = Q\Lambda I\Lambda Q^T = Q\Lambda^2 Q^T$. On généralise cette formule à la puissance k

$$S^k = Q\Lambda^k Q^T \text{ avec } \Lambda^k = diag(\lambda_1^k, \ldots, \lambda_n^k).$$

L'exponentielle d'un nombre se définit comme $\exp x \mapsto \sum_{k=0}^{+\infty} \frac{x^k}{k!}$ et se généralise aux matrices comme tel :

$$\exp S = \sum_{k=0}^{+\infty} \frac{S^k}{k!} = Qe^{\Lambda}Q^T \text{ avec } e^{\Lambda} = diag(e_1^{\lambda}, \ldots, e_n^{\lambda}).$$

1.3 Diagonalisation

Soit $A \in \mathbb{R}^{n\times n}, \lambda \in \mathbb{C}, \lambda$ est une valeur propre de A ssi il existe $v \in \mathbb{C}^n$, non nul tel que $Av = \lambda v$, autrement dit si l'image de v par A est dans la même direction que v. Un tel v est appelé un vecteur propre de A associé à la valeur propre λ.

Il existe une décomposition de A telle que :

$$AT = T\Lambda \text{ avec } T = [v_1, \ldots, v_n], \Lambda = diag(\lambda_1, \ldots, \lambda_n)$$

T est une matrice de rang n car les vecteurs propres sont libres entre eux, donc T est inversible et T^{-1} existe. Ainsi nous pouvons écrire A comme

$$A = T\Lambda T^{-1}$$

Dans le cas où l'on trouve une valeur propre complexe, son conjugué est aussi une valeur propre et les vecteurs propres associés à ces deux valeurs propres sont également conjugués.

1.4 Forme canonique de Jordan

La forme canonique de Jordan est une transformation de similarité de n'importe quelle matrice $A \in \mathbb{R}^{n\times n}$ vers une matrice de la forme :

$$T^{-1}AT = \Lambda = J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{pmatrix}$$

où les blocs J_i de taille n_i s'explicitent comme

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_i \times n_i} \text{ et } n = \sum_{i=1}^{p} n_i$$

Lorsque l'on ne peut pas utiliser la décomposition de Jordan, il est possible d'utiliser la décomposition de Schur, qui décompose une matrice A en une matrice semblable M triangulaire supérieure.

1.5 Déterminant et trace

Soit A une matrice carré $n \times n$, le déterminant de A se décrit formellement comme

$$\det(A) = \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma_i} \right)$$

Avec $\operatorname{sgn}(\sigma)$ la parité du nombre d'élément dans une décomposition de σ en une séquence de transpositions (échange de deux éléments).

Formule du déterminant et de la trace :

— $\det(AB) = \det(A)\det(B)$

— $\operatorname{Tr}(A) = \sum_{i=1}^{n} a_{i,i}$

— $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$

— $\operatorname{Tr}(A_1, A_2, \ldots, A_k) = \operatorname{Tr}(A_2, A_3, \ldots, A_k, A_1) = \operatorname{Tr}(A_k, A_1, A_2, \ldots, A_{k-1})$

— A et B similaires $\Longrightarrow \operatorname{Tr}(A) = \operatorname{Tr}(B)$, $\det(A) = \det(B)$

— Le det d'une matrice par bloc est le produit des déterminants de chacun des blocs.

Il existe des propriétés de matrices qui sont invariantes aux changement de bases. Par exemple : le rang, le polynôme caractéristique, le déterminant, les valeurs propres ...

1.6 Théorème de Cayley-Hamilton

Enoncé : Quelque soit une matrice $A \in \mathbb{R}^{n \times n}$, on a $X_A(A) = 0$ avec $X_A(\lambda) = \det(\lambda I - A)$ le polynôme caractéristique de A.

Corollaire : pour tout entier naturel p, $A^p \in Vect(I, A, A^2, ..., A^{n-1})$

Polynôme caractéristique

Le polynôme caractéristique d'une matrice $A, n \times n$ est défini comme $X_A(\lambda) = \det(\lambda I - A)$. Calculons par exemple le polynôme caractéristique de la matrice $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$:

$$X_A(\lambda) = \begin{vmatrix} \lambda \cdot 1 - 1 & 0 - 2 \\ 0 - 3 & \lambda \cdot 1 - 4 \end{vmatrix} = \begin{vmatrix} \lambda - 1 & -2 \\ -3 & \lambda - 4 \end{vmatrix} = (\lambda - 1)(\lambda - 4) - 6 = \lambda^2 - 5\lambda - 2$$

Les racines de ce polynôme (les λ qui annulent le polynôme) sont les valeurs propres de A.

1.7 Espace de fonctions linéaires

L'espace des fonctions linéaires que l'on note $\mathcal{L}(E, F)$ est constitué de l'ensemble des applications linéaires à valeur de E vers F avec $\dim(E) = n$ et $\dim(F) = m$. Cet espace est de dimension fini, donc bien plus petit que l'espace des fonctions continues.
Notons que l'espace des matrices à valeurs réelles de taille $m \times n$ que l'on note $\mathcal{M}_{m,n}(\mathbb{R})$, est de dimension $n \times m$ et est isomorphe à l'espace des fonctions linéaires $\mathcal{L}(E, F)$, c'est à dire qu'il existe une bijection entre ces deux espaces.

1.8 Norme matricielle

La norme d'une matrice est une fonction $||\cdot|| : \mathbb{R}^{m\times n} \to \mathbb{R}$ qui vérifie les propriétés suivantes :

— $||A|| \geq 0$, $\forall\, A \in \mathbb{R}^{m\times n}$ et $||A|| = 0 \iff A = 0$

— $||\alpha A|| = |\alpha|||A||$, $\forall\, A \in \mathbb{R}^{m\times n}$ et α un scalaire

— $||A + B|| \leq ||A|| + ||B||$, $\forall\, A, B \in \mathbb{R}^{m\times n}$

La norme matricielle est induite par la norme vectorielle, par exemple pour la norme 2 :

$$||A||_2 = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{||Ax||_2}{||x||_2} = \max_{x\in\mathbb{R}^n, ||x||_2=1} ||Ax||_2$$

2 Probabilités

2.1 Rappels

L'**univers**, noté Ω, est l'ensemble des issues pouvant être obtenues lors d'une expérience aléatoire.

Un évènement $A \subseteq \Omega$ est un sous-ensemble des résultats possibles pour une expérience.

L'espace des probabilités, noté $\mathcal{F}$, permet la mesure quantitative d'une expérience aléatoire.

La mesure de probabilité est une fonction à valeurs réelles étant défini comme :

$$\begin{aligned} P : \mathcal{F} &\to \mathbb{R} \\ P(A) &\geq 0, \forall A \in \mathcal{F} \\ P(\Omega) &= 1 \end{aligned}$$

Si A_1, A_2, ... l'ensemble des évènements disjoints, $A_i \subset A_j = \emptyset$ quand $i \neq j$ alors :

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

2.2 Probabilités conditionnelles et théorème de Bayes

Pour tous les évènements A, B tel que $P(B) \neq 0$, on définit :

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

On obtient le **théorème de Bay es** en appliquant la probabilité conditionnelle :

$$\begin{aligned} P(B|A) = \frac{P(B \cap A)}{P(A)} &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(B)P(A|B)}{P(A)} \end{aligned}$$

Dans le cas où l'on prend 3 évènements A, B, C, on peut énoncer le **théorème de Bayes conditionné** :

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$$

2.3 Loi des probabilités totales

Soient $B_1, \ldots, B_n$ n évènements disjoints où l'union est l'univers. Alors pour tout évènement A :

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i) \end{aligned}$$

On peut aussi écrire le théorème de Bayes comme :

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Exemple :
Un coffre au trésor **A** contient 100 pièces d'or, un autre coffre au trésor **B** contient 60 pièces d'or et 40 pièces d'argent.
On choisit aléatoirement uniformément un coffre puis une pièce.
Si cette pièce choisie est en or, quelle est la probabilité que l'on ait choisi le coffre **A** ?
On pose le problème comme la probabilité de prendre le coffre A en sachant que la pièce est en or.
Notations : G pour or (gold), A et B pour les coffres A et B respectifs.
En utilisant le théorème de Bayes et la loi des probabilités totales on peut écrire :

$$P(A|G) = \frac{P(A)P(G|A)}{P(A)P(G|A) + P(B)P(G|B)} = \frac{0.5 * 0.1}{0.5 * 0.1 + 0.5 * 0.6} = \mathbf{0.625}$$

2.4 Chain rule

Enoncé : Pour tout n évènements $A_1, ..., A_n$, la probabilité jointe peut être exprimée par le produit des probabilités conditionnelles.

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1)...P(A_n|A_{n-1} \cap A_{n-2} \cap ... \cap A_1)$$

2.5 Indépendance

Deux évènements A et B sont indépendants si :

$$P(AB) = P(A)P(B)$$

On note $A \perp B$**.**

A partir de là, si $A \perp B$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Cela implique que si deux évènements sont indépendants alors observer un évènement n'aura pas d'effets sur l'autre et inversement.
En général : $A_1, ..., A_n$ sont mutuellement indépendants si :

$$P\left(\bigcap_{i \subseteq S} A_i\right) = \prod_{i \subseteq S} P(A_i)$$

pour tout $S \subseteq \{1, ..., n\}$

2.6 Variables aléatoires

$X = k$ est l'évènement que la variable aléatoire X prend en k.

Variables aléatoires discrètes :
Val(X) est un espace
P(X = k) peut être non nul

Variables aléatoires continues : Val(X) est un interval.
$P(X = k) = 0$ pour tout k : $P(a \leq X \leq b)$ peut être non nul.

2.7 Fonction de masse

Prenons une variable aléatoire discrète X, une fonction de masse associe les valeurs de X à une probabilité.
$p_x(x) := P(X = x)$
Pour une fonction de masse valide, il faut $\sum_{x \in Val(x)} p_x(x) = 1$

2.8 Fonction de répartition

Une fonction de répartition associe une variable aléatoire continue à une probabilité (i. e. $\mathbb{R} \to [0, 1]$)

$$F_X(x) := P(X \leq x)$$

une fonction de répartition doit respecter les règles suivantes :
$\lim_{x \to -\infty} F_X(x) = 0$
$\lim_{x \to \infty} F_X(x) = 1$
Si $a \leq b$ alors $F_X(a) \leq F_X(b)$

On note aussi $P(a \leq X \leq b) = F_X(b) - F_X(a)$.

2.9 Variable aléatoire à densité

Une fonction aléatoire à densité d'une variable aléatoire continue est la dérivée de la fonction de la répartition.

$$f_X(x) := \frac{dF_X(x)}{dx}$$

On a donc :

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$

La fonction est valide si :

pour tous les les réelles x, $f_X(x) \geq 0$,
$\int_{-\infty}^{\infty} f_X(x)dx = 1$.
L'aire sous une courbe de densité doit être égale à 1.

2.10 Espérance

Si X est une variable aléatoire **discrète** :

$$\mathbb{E}[g(X)] := \sum_{x \in \text{Val}(X)} g(x)p_X(x)$$

Si X est une variable aléatoire **continue** :

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

avec g une fonction à valeur réelle arbitraire.

2.10.1 Propriétés

Pour toute valeur constante $a \in \mathbb{R}$ et une fonction arbitraire f :
$\mathbb{E}[a] = a$
$\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$

Linéarité de l'espérance :
Soient n fonctions à valeurs réelles $f_1(X), \ldots, f_n(X)$,

$$\mathbb{E}\left[\sum_{i=1}^{n} f_i(X)\right] = \sum_{i=1}^{n} \mathbb{E}\left[f_i(X)\right]$$

2.11 Variance

La variance d'une variable aléatoire X mesure la concentration de la distribution de X autour de sa moyenne.

$$\begin{aligned} \mathrm{Var}(X) : &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

2.11.1 Propriétés

Sois a une constante.

$$\begin{aligned} &\mathrm{Var}[a] = 0 \\ &\mathrm{Var}[\mathrm{af}(X)] = a^2 \,\mathrm{Var}[f(X)] \end{aligned}$$

2.12 Distributions marginales et jointes

Fonction de masse jointe pour variables aléatoires discrètes X, Y :

$$p_{XY}(x, y) = P(X = x, Y = y)$$

Notons que $\sum_{x \in \mathrm{Val}(X)} \sum_{y \in \mathrm{Val}(Y)} p_{XY}(x, y) = 1$

Fonction de masse marginale de X, en donnant la fonction de masse jointe de X, Y :

$$p_X(x) = \sum_y p_{XY}(x, y)$$

Variables aléatoire de densité jointe pour X, Y continus :

$$f_{XY}(x, y) = \frac{\delta^2 F_{XY}(x, y)}{\delta x \delta y}$$

Notons que $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$

Variable aléatoire à densité marginale de X, en donnant la jointe de X, Y :

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

2.13 Distributions marginales et jointes pour plusieurs variables aléatoires

Fonction de masse jointe pour variables aléatoires discrètes $X_1, \ldots, X_n$:
$p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$
Notons que $\sum_{x_1} \sum_{x_2} \ldots \sum_{x_n} p(x_1, \ldots, x_n) = 1$

Fonction de masse marginale de X, en donnant la fonction de masse jointe de $X_1, \ldots, X_n$:

$$p_{X_1}(x_1) = \sum_{x_2} \ldots \sum_{x_n} p(x_1, \ldots, x_n)$$

Variables aléatoire de densité jointe pour $X_1, \ldots, X_n$ continus :

$$f(x_1, \ldots, x_n) = \frac{\delta^n F(x_1, \ldots x_n)}{\delta x_1 \delta x_2 \ldots \delta x_n}$$

Notons que $\int_{x_1} \int_{x_2} \cdots \int_{x_n} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1$

Variable aléatoire à densité marginale de X_1, en donnant la jointe de $X_1, \ldots, X_n$:
$f_{X_1}(x_1) = \int_{x_2} \ldots \int_{x_n} f(x_1, \ldots, x_n)\, dx_2 \ldots dx_n$

2.14 Propriétés sur les variables aléatoires

2.14.1 Distribution conditionnelles

Marche de la même manière avec les variables aléatoires qu'avec les évènements :

Pour des variables discrètes X, Y :

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

Pour des variables continues X, Y :

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

En général, pour des variables continues $X_1, \ldots, X_n$:

$$f_{X_1|X_2,\ldots,X_n}(x_1 \mid x_2, \ldots, x_n) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}{f_{X_2,\ldots,X_n}(x_2, \ldots, x_n)}$$

2.14.2 Théorème de Bayes

Fonctionne aussi de la même manière :
Pour des variables discrètes X, Y :

$$p_{Y|X}(y \mid x) = \frac{p_{X|Y}(x \mid y) p_Y(y)}{\sum_{y' \in \mathrm{Val}(Y)} p_{X|Y}(x \mid y')\, p_Y(y')}$$

Pour des variables continues X, Y :

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x \mid y')\, f_Y(y')\, dy'}$$

2.14.3 Chain rule

De la même manière :

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_1)\, f(x_2 \mid x_1) \ldots f(x_n \mid x_1, x_2, \ldots, x_{n-1}) \\ &= f(x_1) \prod_{i=2}^{n} f(x_i \mid x_1, \ldots, x_{i-1}) \end{aligned}$$

2.14.4 Indépendance

Pour $X \perp Y$, il faut que $F_X Y(x, y) = F_X(x) F_Y(y)$ pour **toutes les valeurs** de x,y
Tant que $f_{Y|X}(y|x) = f_Y(y)$, si $X \perp Y$ alors la chain rule pour les variables indépendantes $X_1 ... X_n$ est :

$$f(x_1, ..., x_n) = f(x_1) f(x_2) ... f(x_n) = \prod_i f(x_i)$$