Cours IFT6266: Algorithmes d'apprentissage

Yoshua Bengio

Concepts mathématiques préliminaires

La plupart de ces concepts sont vus dans d'autres cours et forment une bonne base pour la matière étudiée dans ce cours. Assurez vous d'avoir au moins une compréhension de la sémantique de ces concepts (pas nécessairement des détails mathématiques et des preuves sous-jacentes).

- **algèbre vectorielle**
  - Espace vectoriel et vecteur. Espace généré par une famille de vecteurs. Famille libre ou liée (vecteurs linéairement indépendants).
  - Base, produit scalaire, norme. Angle entre deux vecteurs.
  - Matrice et opérateur linéaire. Transposée. Déterminant. Diagonale. Matrice identité. Trace.
  - Addition, multiplication et inversion de matrices. Non-commutativité de la multiplication.
  - Déterminant. Rang.
  - Algèbre linéaire sur des espaces de dimension infinie, espaces de fonctions.
  - Valeurs propres et vecteurs propres.
  - Base des vecteurs propres. Décomposition d'une matrice carrée en somme de produits de vecteurs propres.
  - Inverse d'une matrice. Matrice inversible. Matrice mal conditionnée.
  - Décomposition en valeurs singulières.
  - Systèmes d'équations linéaires. Pseudo-inverse.
- **analyse**
  - Suites.
  - Séries. Limite. Série convergente.
  - Série de Taylor.

- **probabilités**
  - Combinatoire de base (permutations, combinaisons, factorielle).
  - Événement aléatoire. Variable aléatoire. Axiomes des probabilités.
  - Variables aléatoires discrètes, continues, hybrides (points de masse).
  - Fonction de probabilité. Fonction de densité. Fonction de répartition.
  - Probabilité comme limite d'une fréquence vs probabilité comme croyance.
  - Espérance. Variance. Covariance. Moments. Moment centré.
  - Convergence en loi.
  - Loi de Bernouilli et binomiale.
  - Loi normale univariée et multivariée. Matrice de covariance.
  - Probabilité jointe. Probabilité marginale. Probabilité conditionnelle. Espérance conditionnelle.
  - Théorème de Bayes. Distribution a priori et distribution a posteriori.
  - Indépendance entre variables aléatoires. Indépendance conditionnelle.
  - Décomposition d'une probabilité jointe en produit de conditionnelles.
  - Théorème de Shannon. Entropie. Information mutuelle. Divergence de Kullback-Leibler.
  - Échantillonnage selon une loi. Estimation d'une intégrale par Monte-Carlo.
  - Chaîne de Markov. Chaîne homogène.
  - Marche aléatoire. Processus stochastique. Champs markovien.
- **Statistiques**
  - Hypothèse de données i.i.d.
  - Inférence statistique. Statistique. Estimateur.
  - Biais et variance d'un estimateur. Estimateur efficace.
  - Vraisemblance. Estimateur du maximum de vraisemblance.
  - Lois des grands nombres et théorème limite centrale.
  - Solutions analytiques du maximum de vraisemblance.
  - Convergence asymptotique de cet estimateur: non-biaisé et efficace asymptotiquement.

- Intervalle de confiance.
- Régression linéaire ordinaire.
- Test statistique. Hypothèse nulle. Seuil critique. p-valeur d'un test.
- Niveau et puissance d'un test.
- Statistiques suffisantes.
- Estimation de densité. Estimation de probabilité conditionnelle.
- Bootstrap.
- Modèle paramétrique versus non-paramétrique.

- **Calcul différentiel**
  - Dérivée comme une limite.
  - Dérivée partielle.
  - Dérivée totale, dérivée en chaîne.
  - Interprétation géométrique des premières et secondes dérivées: pente et courbure, en dimension 1 et plus. Gradient et Hessien.

- **Optimisation**
  - Points critiques d'une fonction. Existence d'un minimum.
  - Minimum local et minimum global.
  - Conditions du premier ordre pour les points critiques. Conditions du 2eme ordre pour identifier min, max, et point selle.
  - Géométrie des points critiques en dimension > 1.
  - Optimisation sous contrainte d'égalité via les multiplicateurs de Lagrange.
  - Fonction convexe. Ensemble convexe. Inégalité de Jensen.
  - Descente de gradient. Optimisation par la règle de Newton.

Lexique informel

- **Espace vectoriel:** un espace vectoriel sur $\mathbb{R}$ est un espace stable par combinaison linéaire. Soit x_1 et x_2 deux éléments d'un espace vectoriel E et λ_1 et λ_2 deux réels, alors $\lambda_1 x_1 + \lambda_2 x_2$ est également un élément de E. *Exemples d'espaces vectoriels sur $\mathbb{R}$: l'espace des réels (évident), l'espace des fonctions nulles en 0, l'espace des complexes, ...*

- **Vecteur:** un vecteur est un élément d'un espace vectoriel. *Exemple: un nombre réel dans l'espace des nombres réels, la fonction sinus dans l'espace des fonctions nulles en 0, ...*

- **Espace généré par une famille de vecteurs:** soit une famille de vecteurs $(x_1, \ldots, x_n)$ (non forcément distincts ni non nuls). Alors l'ensemble des vecteurs de la forme $x = \sum_{i=1}^{n} \lambda_i x_i$ avec $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ (combinaison linéaire des x_i) forme un espace vectoriel appelé **espace vectoriel généré par** $(x_1, \ldots, x_n)$ et noté $Vect((x_1, \ldots, x_n))$.

- **Famille libre, famille liée:** une famille $(x_1, \ldots, x_n)$ est dite **libre** si aucun des x_i ne peut s'écrire comme combinaison linéaire des x_j pour $j \neq i$. Si la famille n'est pas libre, on dit qu'elle est **liée**. *Exemple: la famille* $((1,0),(0,1))$ *est libre mais la famille* $((1,0),(0,1),(2,1))$ *est liée car* $(2,1) = 2 * (1,0) + 1 * (0,1)$.

- **Base:** soit un espace vectoriel E. Soit $(x_1, \ldots, x_n)$ une famille libre. Si $Vect((x_1, \ldots, x_n)) = E$, alors on dit que $(x_1, \ldots, x_n)$ est une **base** de E. Il existe une infinité de bases de E mais elles ont toutes le même nombre de vecteurs. Ce nombre s'appelle la **dimension** de E.

- **Produit scalaire:** un produit scalaire dans un espace vectoriel E sur $\mathbb{R}$ est une fonction de $E \times E$ dans $\mathbb{R}$ (dont les propriétés seront évoquées en cours). Le produit scalaire de deux vecteurs x et y est souvent noté $(x|y)$, $< x|y >$ ou encore $x \cdot y$. *Exemple: dans le plan euclidien, le produit scalaire de deux vecteurs* (x_1, x_2) *et* (y_1, y_2) *est* $x_1y_1 + x_2y_2$.

- **Norme:** une norme dans un espace vectoriel E sur $\mathbb{R}$ est une fonction de E dans $\mathbb{R}$ (dont les propriétés seront évoquées en cours). La norme d'un vecteur x est souvent notée $|x|$ ou $||x||$. Si E est muni d'un produit scalaire, on peut définir une norme associée à ce produit scalaire par $||x|| = \sqrt{< x|x >}$.

- **Angle entre deux vecteurs:** $\cos\theta = \frac{u \cdot v}{||u||||v||}$.

- **Matrice:** Une matrice de taille (m, n) est la représentation d'une application linéaire d'un espace vectoriel E de dimension n muni d'une base $(e_{11}, \ldots, e_{1n})$ vers un espace vectoriel F de dimension m muni d'une base $(e_{21}, \ldots, e_{2m})$ (E peut être identique à F et e_1 peut être identique à e_2). La j^e colonne d'une matrice est l'image de e_{1j} exprimée dans la base $(e_{21}, \ldots, e_{2m})$. On s'efforcera de garder constamment à l'esprit cette équivalence matrice - application linéaire. Note: si une matrice est de taille (n, n), on simplifiera la notation en disant qu'elle est de taille n.

- *addition:* on ne peut additionner deux matrices A et B que si elles ont la même taille. Dans ce cas, il suffit d'additionner terme à terme les deux matrices. Si f est l'application linéaire associée à A et g l'application linéaire associée à B, alors l'application linéaire associée à $A + B$ est $f + g$ (cela n'étant bien entendu vrai que si les deux espaces de départ sont les mêmes, les deux espaces d'arrivée sont les mêmes, les deux bases de départ sont les mêmes et les deux bases d'arrivée sont les mêmes).
- *multiplication:* si A est une matrice de taille (m, n) et B est une matrice de taille (n, p), alors on peut multiplier les deux matrices et la matrice AB sera de taille (m, p). L'application linéaire associée à la matrice AB est $f \circ g$. Donc, tout comme la composition de fonctions, la multiplication de matrices n'est pas commutative.

- **Transposée d'une matrice:** la transposée M' d'une matrice M de taille (m, n) est la matrice telle que pour tout $1 \leq i \leq m$ et pour tout $1 \leq j \leq n$, on a $M_{ij} = M'_{ji}$. On dira qu'une matrice A est symétrique si elle vérifie $A' = A$. Il est évident qu'une matrice symétrique est obligatoirement carrée.

- **Diagonale d'une matrice:** la diagonale d'une matrice carrée A de taille n est le vecteur de taille n contenant les éléments diagonaux de la matrice. Une **matrice diagonale** est une matrice dont les seuls éléments non nuls sont les éléments situés sur la diagonale. Une matrice diagonale est bien entendu symétrique. Une matrice diagonale particulière est la **matrice identité** qui ne possède que des 1 sur sa diagonale. Elle est notée I_n ou I et vérifie pour tout x appartenant à E: $Ix = x$. Cette matrice est associée à l'application identité id telle que pour tout x appartenant à E, on a $id(x) = x$. On a également pour toute matrice A, $AI = IA = A$ (pour toute fonction g $g(id(x)) = id(g(x)) = g(x)$).

- **Trace:** la trace d'une matrice carrée A est la somme des éléments diagonaux de A et est notée $tr(A)$. Si A est une matrice (m, n) et B est une matrice (n, m), alors on a $tr(AB) = tr(BA)$. *Exemple:* $tr(I_n) = n$.

- **Déterminant:** on oublie la définition formelle et la manière de le calculer. On l'écrit parfois $|A|$ ou $det(A)$. On retiendra juste quelques propriétés de base:
  - le déterminant d'une matrice inversible est non nul
  - le déterminant d'une matrice singulière (non inversible) est nul
  - le déterminant d'une matrice diagonale est égal au produit des éléments de sa diagonale

  - le déterminant d'une matrice triangulaire supérieure est égal au produit des éléments de sa diagonale
  - le déterminant d'une matrice diagonalisable est égal au produit de ses valeurs propres
  - le déterminant d'un produit de matrices est égal au produit des déterminants des matrices
  - le déterminant d'une matrice non carrée est nul
  - on ne peut rien dire sur le déterminant d'une somme de matrices
  - $det(I) = 1$, ce qui est logique puisque $det(AI) = det(IA) = det(A) * det(I)$, mais comme $AI = IA = A$, on a $det(A) * det(I) = det(A)$

- **Rang:** soit une application linéaire f de E dans F et $Im(f) \subset F$ l'ensemble de toutes les images des éléments de E par f. Alors $Im(f)$ est un espace vectoriel. Soit r sa dimension. r est appelé le **rang** de f. r est également le rang de la matrice associée à f. Le rang d'une matrice est le mininum du nombre de colonnes linéairement indépendantes et du nombre de lignes linéairement indépendantes.

- **Espaces vectoriels de dimension infinie:** la notion d'espace vectoriel s'étend également à la dimension infinie (dénombrable ou indénombrable). Il faut cependant se méfier car il n'est plus question d'utiliser des matrices (mais la notion d'application linéaire reste valide) et certains résultats valables en dimension finie ne s'étendent pas à la dimension infinie. *Exemple d'espace vectoriel de dimension infinie: l'espace vectoriel des fonctions de période* 2π *a une base infinie (dénombrable) constituée des fonctions* $x \rightarrow sin(px)$ *et* $x \rightarrow cos(px)$ *pour* $p \in \mathbb{N}$.

- **Valeurs propres, vecteurs propres:** une application linéaire agit comme une transformation géométrique de l'espace. Il arrive que l'image par une application linéaire f d'un vecteur v de l'espace soit un vecteur colinéaire à v. On a donc $f(v) = \lambda v$. On dit alors que v est un **vecteur propre** de f associé à la **valeur propre** λ. Si v est vecteur propre, alors pour tout $\alpha \in \mathbb{R}$, αv est aussi un vecteur propre associé à la même valeur propre: en pratique, lorsqu'on s'intéresse aux vecteurs propres d'une application, on considère ceux qui ont norme 1. Si l'on peut trouver une base intégralement constituée de vecteurs propres (même associés à des valeurs propres différentes), on dira que l'application linéaire f est **diagonalisable**. Il faut noter que l'on utilise exactement la même terminologie (vecteur propre, valeur propre, diagonalisable) lorsque l'on parle de la matrice associée à f. On pourra donc

rencontrer des applications linéaires diagonalisables ou des matrices diagonalisables. *Exemples: dans le plan euclidien, la rotation de centre O et d'angle $\frac{\pi}{4}$ ne possède aucun vecteur propre (il suffit de faire un petit dessin). En revanche, la projection sur l'axe des abscisses (que l'on nommera p_{Ox}) est diagonalisable car elle possède une base de vecteurs propres. En effet, le vecteur $(1,0)$ ne bouge pas (on a $p_{Ox}(1,0) = (1,0)$). Il est donc vecteur propre pour la valeur propre 1. Au contraire, le vecteur $(0,1)$ devient le vecteur nul (on a $p_{Ox}(0,1) = (0,0)$). Il est donc vecteur propre pour la valeur propre 0. Ces deux vecteurs formant une base du plan, l'application p_{Ox} est diagonalisable.*

- **Décomposition d'une matrice symétrique en somme de produits de vecteurs propres:** toute matrice symétrique A (c'est-à-dire telle que $A' = A$) est diagonalisable dans une base orthonormale de vecteurs propres. Si elle a comme vecteurs propres $(v_1, \ldots, v_n)$ associés respectivement aux valeurs propres $(\lambda_1, \ldots, \lambda_n)$ (non forcément toutes distinctes), alors A peut s'écrire sous la forme $A = \sum_{i=1}^{n} \lambda_i v_i v_i'$.

- **Matrice inversible:** une matrice inversible A est une matrice dont le déterminant est non nul. Elle correspond à une application linéaire bijective f.

- **Inverse d'une matrice:** si A est une matrice inversible, alors on peut définir son inverse A^{-1} par $A^{-1}A = AA^{-1} = I$. Si A est associée à l'application linéaire bijective f, alors A^{-1} est associée à f^{-1}. Il faut bien faire attention à l'inversion des espaces vectoriels de départ et d'arrivée.

- **Matrice mal conditionnée:** une matrice mal conditionnée est une matrice donc le déterminant est très proche de 0. Cela peut mener à des erreurs de calcul lors de son inversion. L'un des moyens pour reconditionner une matrice est d'ajouter une certaine valeur ϵ (arbitraire) à tous les éléments de la diagonale. Cela présente l'avantage de ne pas modifier les vecteurs propres de la matrice, car $(A + \epsilon I)v = Av + \epsilon v = (\lambda + \epsilon)v$ si v est vecteur propre de valeur propre λ.

- **Décomposition en valeurs singulières:** toute matrice réelle A $(m \times n)$ peut se décomposer ainsi: $A = UDV'$ où $U'U = I$, $V'V = I$ et D est diagonale. Les éléments diagonaux de D sont appelés valeurs singulières. Le nombre de valeurs singulières non-nulles est le **rang** de A.

- **Système d'équations linéaires:** soit un système

$$\left\{\begin{array}{rcl} a_{11}x_1 + \ldots + a_{1j}x_j + \ldots + a_{1p}x_p & = & y_1 \\ \ldots & & \\ a_{i1}x_1 + \ldots + a_{ij}x_j + \ldots + a_{ip}x_p & = & y_i \\ \ldots & & \\ a_{n1}x_1 + \ldots + a_{nj}x_j + \ldots + a_{np}x_p & = & y_n \end{array}\right.$$

  dont les inconnues sont $x_1, \ldots, x_p$. Alors ce système est équivalent à l'équation $Ax = y$ avec $A_{ij} = a_{ij}$, $x = (x_1, \ldots, x_p)$ et $y = (y_1, \ldots, y_n)$. Ce système peut posséder une solution, une infinité de solutions ou pas de solutions selon le rang de la matrice A.

- **Pseudo-inverse d'une matrice:** alors que l'inverse d'une matrice n'est définie que pour une matrice inversible, on peut définir la pseudo-inverse d'une matrice singulière. La pseudo-inverse présente l'avantage d'être égale à l'inverse pour une matrice inversible. La pseudo-inverse B d'une matrice A vérifie $BAB = B$ et $ABA = A$. Il existe plusieurs pseudo-inverses d'une matrice A. L'une d'elles est $(A'A)^{-1}A'$. Cette pseudo-inverse correspond à une solution approximative du système linéaire $Ax = b$ dans laquelle on résout plutôt $Ax = \hat{b}$, où $\hat{b}$ est la projection de b sur l'espace généré par les vecteurs colonnes de A (ou les vecteurs singuliers gauches de la SVD). C'est donc la solution qui minimise l'erreur quadratique $||b - \hat{b}||^2$.

- **Suite:** une suite est un ensemble d'éléments indexés par $\mathbb{N}$.

- **Série:** une série (S_n) est définie à partir d'une suite (U_n) par $S_k = \sum_{i=0}^{k} U_i$.

- **Limite:** on dit qu'une suite (U_n) (resp. une série (S_n)) admet une limite l si $\forall \epsilon > 0$, $\exists N / \forall n > N$, $||U_n - l|| < \epsilon$ (resp. $||S_n - l|| < \epsilon$). Si $l = +\infty$ (resp. $l = -\infty$), ça devient $\forall A$, $\exists N / \forall n > N$, $U_n > A$ (resp. $U_n < A$) (dans le cas des suites réelles). Il est très important de noter que l'existence d'une limite dépend du choix de la norme. En revanche, si une suite possède une limite pour deux normes différentes, alors cette limite est la même.

- **Série convergente:** une série convergente est une série qui admet une limite finie.

- **Développement de Taylor d'une fonction:** soit f une fonction de $\mathbb{R}$ dans un espace vectoriel normé (espace vectoriel muni d'une norme) au moins n fois dérivable en a. On définit le développement de Taylor de f à l'ordre n en a qui est une fonction de $\mathbb{R}$ dans notre espace vectoriel normé par

$P_{f,n,a}(x) = f(a)+\sum_{k=1}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$. Les fonctions infiniment différentiables (analytiques) peuvent s'écrire exactement dans cette expansion (avec $n \to \infty$). Pour une fonction "lisse", les termes d'ordre supérieur recoivent un coefficient très petit. Ceci est d'autant plus vrai pour a proche de x, on peut donc obtenir une bonne approximation locale d'une fonction lisse en ignorant les termes d'ordre supérieur.

- **Factorielle:** $n!$ (à prononcer "Factorielle n") est égal à $1 * 2 * 3 * \ldots * n$.
- **Permutation:** une permutation de l'ensemble $(1, \ldots, n)$ est une fonction bijective de $(1, \ldots, n)$ dans $(1, \ldots, n)$
- **Combinaison:** $C(n, p)$ est égal au nombre de possibilités de sélectionner p éléments parmi n, l'ordre de sélection étant indifférent (c'est-à-dire que $(1, 2)$ et $(2, 1)$ représentent le même ensemble). On a $C(n, p) = \frac{n!}{p!(n-p)!}$
- **Variable aléatoire:** une variable aléatoire X est associée à un ensemble de valeurs Ω et à une **loi** (ou distribution, ou fonction de probabilité) P_X (ou P_x ou $P(X \in \ldots)$). Cette fonction fait correspondre à chaque sous-ensemble de Ω un nombre réel entre 0 et 1.
- **Axiomes des probabilités:** Soit $E \subset \Omega$ l'occurrence d'un événement.
  1. $0 \leq P(E) \leq 1$
  2. $P(\Omega) = 1$ et $P(\emptyset) = 0$.
  3. $P(E_1 \; ou \; E_2) = P(E_1) + P(E_2) - P(E_1 \; et E_2)$
- **Variable aléatoire discrète:** l'ensemble de ses valeurs est énumérable (donc est fini ou en bijection avec $\mathbb{N}$). On peut associer à chaque valeur i une probabilité à travers une **fonction de probabilité** $P(X = i) = P(X \in \{i\})$.
- **Variable aléatoire continue:** l'ensemble Ω de ses valeurs est continu (non-énumérable) et l'on peut associer à chaque valeur possible x dans Ω une valeur réelle à travers la **fonction de densité** $P_X(x)$. La probabilité d'un sous-ensemble A est obtenue en intégrant $P_X(x)$ dans ce sous-ensemble. Pour x un vecteur euclidien, $P(X \in A) = \int_A P_X(x)dx = \int P_X(x)1_{x \in A}dx$. *Abus courant de notation, $P(X = x)$ pour représenter $P_X(x)$, alors que formellement $P(X = x) = 0$.*
- **Variable aléatoire hybride:** l'ensemble des valeurs qu'elle peut prendre est indénombrable mais certaines valeurs ont une probabilité non-nulle. e.g. X est généré une fois sur deux en prenant l'uniforme sur $[0, 1]$ et une fois sur

deux en prenant la valeur 0.5 (qui est un **point de masse**). Dans ce cas la fonction de densité n'existe pas vraiment, à moins d'admette une valeur spéciale (le delta de Dirac $\delta(x)$) qui vaut ∞ mais dont l'intégrale dans un intervalle autour de x est 1.

- **Fonction de répartition:** c'est la **probabilité cumulative**, donc celle d'un événement qui couvre un intervalle de moins l'infini à x, parfois notée $F_X(x) = P(X < x) = \int_{-\infty}^{x} P_X(y)dy$ (dans le cas continu). Dans le cas où x est un vecteur on interprète l'événement $X < x$ comme l'intersection des événements $X_i < x_i$. On peut parler de la fonction de répartition dans le cas hybride soit en utilisant l'interprétation que $P_X(x = m) = \delta(m)P(X \in \{m\})$ où en utilisant la notation $P(X < x) = \int_{-\infty}^{x} dF_X(y)$ (donc $dF_X(y)$ est $\mu(dy)$, utilisée comme **mesure**).

- **Probabilité comme limite d'une fréquence:** si on mesure l'occurrence d'un événement A n fois, la fréquence de cette événement converge vers la probabilité $P(A)$ quand $n \to \infty$. C'est une manière (**fréquentiste**) de définir les probabilités.

- **Probabilité comme croyance:** une fonction de probabilité peut aussi être utilisée pour représenter une croyance concernant les résultats possibles d'une expérience (c'est l'approche **Bayesienne**).

- **Espérance:** $E_X[f(X)] = \int P_X(x)f(x)dx$ ce qui dans le cas discret donne $E_X[f(X)] = \sum_{x\in\Omega} P(X = x)f(x)$. Quand on écrit $E[...]$ on fait référence à l'intégrale sur toutes les variables aléatoires entre crochets.

- **Variance:** Pour X scalaire, $Var[X] = E[(X - E[X])^2]$. Pour X un vecteur, c'est la matrice $Var[X] = E[(X - E[X])'(X - E[X])]$, aussi appelée matrice de covariance ou matrice de variance-covariance (les variances sont dans la diagonale, les covariances hors-diagonales). La **covariance** entre X_i et X_j est $(Var[X])_{ij} = Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$.

- **Moments:** ce sont les espérances des puissances d'une variable aléatoire. Le p-ième moment de X est $E[X^p]$. Le p-ième **moment centré** est $E[(X - E[X])^p]$.

- **Convergence en loi:** plusieurs formes de convergence d'une suite de variables aléatoires $X_1, X_2, \ldots X_n$ existent, mais le principe est que la loi de X_n converge au fur et à mesure que $n \to \infty$.

- **Loi de Bernouilli:** on a une v.a. binaire ($X = 0$ ou $X = 1$), avec paramètre $p = P(X = 1)$ et $1 - p = P(X = 0)$.

- **Distribution binomiale:** on a N v.a. Bernouilli avec le même paramètre p et la somme S de ces v.a. suit la loi binomiale: $P(S = n) = C(N, n)p^n(1 - p)^{N-n}$.

- **Loi normale ou Gaussienne:** dans le cas où $X \in \mathbb{R}$ (**univarié**), on a simplement la densité $P_X(x) = \frac{e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}}{\sqrt{2\pi}\sigma}$ avec paramètres μ (moyenne) et σ^2 (variance). Dans le cas où $X \in \mathbb{R}^n$ (multivarié), on a $P_X(x) = \frac{e^{-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)}}{(2\pi)^{\frac{n}{2}} det(V)^{\frac{1}{2}}}$, avec paramètres $\mu \in \mathbb{R}^n$ (vecteur moyenne) et $V \in \mathbb{R}^{n\times n}$ (**matrice de covariance**).

- **Probabilité jointe:** soit deux v.a. X et Y. On peut définir une troisième v.a. $Z = (X, Y)$ qui représente l'événement joint "X prend une certaine valeur et Y prend une certaine valeur". Si $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ alors $Z \in \mathcal{X} \times \mathcal{Y}$. On écrit $P(Z = (x, y)) = P(X = x, Y = y) = P(X = x \; et \; Y = y)$.

- **Probabilité marginale:** $P(X)$ et $P(Y)$ sont des probabilités marginales par rapport à la loi jointe $P(X, Y)$. Notons que $P_X(x) = \int P_Z(x, y)dy$ ou dans le cas discret $P(X = x) = \sum_y P(X = x, Y = y)$.

- **Probabilité conditionnelle:** on définit $P(Y = y|X = x) = \frac{P(X=x,Y=y)}{P(X=x)}$. C'est la loi de Y quand on se restreint au sous-ensemble des occurrences possibles de Z pour lesquelles $X = x$. Pour une v.a. continue on a aussi $P_{Y|X=x}(y) = \frac{P_Z(x,y)}{P_X(x)}$.

- **Espérance conditionnelle:** simplement l'espérance sous la loi conditionnelle, notée $E[Y|X = x]$ et égale à $\int yP_{Y|X=x}(y)dy$ ou $\sum_y yP(Y = y|X = x)$.

- **Théorème de Bayes:** une conséquence de la définition de probabilité conditionelle, $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$. Dans cette équation on appelle $P(Y)$ la **loi a priori**, et $P(Y|X)$ la **loi a posteriori**.

- **Indépendance entre variables aléatoires:** notée $A \perp B$, ssi $P(A, B) = P(A)P(B)$. Pour les variables continues c'est également ssi $p_{AB}(a, b) = p_A(a)p_B(b)$.

- **Indépendance conditionnelle:** on dit que A et B sont indépendantes conditionnellement à C ssi $P(A, B|C) = P(A|C)P(B|C)$ (et aussi pour les densités), i.e. A et B sont indépendants dans le contexte où C est donné (cela pourrait être vrai en général ou pour une valeur particulière de C). On note $(A \perp B)|C$.

- **Notation pour une sous-séquence:** $X_i^j = (X_i, X_{i+1}, \ldots, X_j)$. Par convention X_i^{i-1} est la séquence vide.

- **Probabilité jointe = produit de conditionnelles:** en appliquant deux fois la définition de probabilité conditionnelle, on obtient que $P(A, B, C) = P(A)P(B|A)P(C|A, B)$ (OK pour tous les ordres), et plus généralement $P(X_1^n) = \prod_{i=1}^n P(X_i|X_1^{i-1})$ pour n'importe quel ordre des v.a.

- **Théorème de Shannon:** pour compresser des données tirées d'une v.a. discrète X le mieux qu'on puisse faire est d'utiliser en moyenne $-\log_2 P(X = x)$ bits pour encoder la valeur x.

- **Entropie:** l'entropie d'une v.a. X est $H(X) = -\int P_X(x) \log P_X(x) dx$ ou $H(X) = E[-\log P(X)] = -\sum_x P(X = x) \log P(X = x)$. N.B. les unités diffèrent selon la base du logarithme: ce sont des **bits** en base 2, des **nats** en base e. C'est donc le nombre de bits moyens pour encoder des données provenant de la v.a. discrète X. L'entropie mesure l'uniformité d'une v.a.: à un extrême on a une loi uniforme ($P(X = i) = P(X = j)$) et à l'autre on a une v.a. qui ne peut prendre qu'une valeur ($P(X = i) = 1_{i=j}$).

- **Information mutuelle:** mesure le nombre de bits qu'une v.a. contient à propos d'une autre. $I(X, Y) = E[\log \frac{P(X,Y)}{P(X)P(Y)}]$. Notons que $I(X, Y) = I(Y, X) \geq 0$ et $I(X, Y) = H(X) + H(Y) - H(X, Y)$.

- **Divergence de Kullback-Leibler:** une mesure *asymétrique* de l'éloignement entre deux distributions (sur le même espace). L'une d'elle (P ci-bas) est considérée comme la *référence* (la "vraie"), et l'on compte le nombre de bits SUPPLÉMENTAIRES pour encoder des données provenant de P, en utilisant un code optimisé pour encoder des données provenant de Q. On note $KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$. On peut aussi la définir pour des v.a. continues: $KL(P||Q) = \int P_X(x) \log \frac{P_X(x)}{Q_X(x)} dx$. On a que $KL(P||Q) \geq 0$ avec égalité ssi $P = Q$. Aussi on obtient ∞ si $\exists x$ t.q. $P(x) > 0$ et $Q(x) = 0$.

- **Échantillonnage selon une loi:** il s'agit de tirer des valeurs aléatoires x selon la loi d'une v.a. X. Pour certaines lois on sait facilement le faire par ordinateur, alors que pour d'autres il faut recourir à des techniques itératives (Monte-Carlo Markov Chain). Il est facile de tirer d'une loi uniforme et de là d'une loi normale. Notons que les tirages par ordinateurs ne sont jamais "parfaits" (i.e. les tirages successifs sont difficilements complètement indépendants).

- **Estimation d'une intégrale ou une somme par Monte-Carlo:** on peut bien estimer une *espérance* par une *moyenne* de v.a. i.i.d., comme l'indique le théorème limite centrale. Donc $\int f(x)P_X(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$ où les x_i sont tirés selon P_X. L'erreur quadratique espérée de cette approximation est $\frac{\sigma^2}{n}$ où $\sigma^2 = Var[f(X)]$. C'est la meilleure méthode quand X est de dimension élevée (plus que 5 ou 10). Pour de faibles dimensions on peut obtenir des résultats plus efficaces en choisissant les x_i pour quadriller l'espace de manière intelligente.

- **Marche aléatoire:** une suite de v.a. obtenue selon un certain processus aléatoire. Un **processus stochastique** est une famille *indexée* de v.a. (par exemple indexée par le temps t, dans une marche aléatoire, ou par la position (x, y) pour un **champ markovien**).

- **Chaîne de Markov d'ordre k:** une suite de variables aléatoires $\ldots X_t, X_{t+1}, \ldots$ t.q. $\forall t, \forall i < t-k+1,\ X_i \perp X_{t+1} | (X_{t-k+1}, \ldots, X_t)$, i.e. $P(X_{t+1}|X_1^t) = P(X_{t+1}|X_{t-k+1}^t)$. On dit que la chaîne est **homogène** ssi cette distribution de transition ne dépend pas de t.

- **Hypothèse de données i.i.d.:** Soit $D = \{X_1, \ldots X_n\}$ un ensemble de données. L'hypothèse i.i.d. (indépendantes et identiquement distribuées) est que $P(X_i) = P(X_j)$ et $\forall i \neq j, X_i \perp X_j$. Cette hypothèse est essentielle pour pouvoir prouver qu'un algorithme entraîné sur certaines données pourra bien généraliser sur d'autres tirées de la même distribution.

- **Inférence statistique:** estimer une propriété de la loi P dont proviennent les données D, à partir des données D.

- **Statistique:** une variable aléatoire obtenue de manière déterministe à partir de l'ensemble de données.

- **Estimateur:** une inférence statistique sur la valeur d'une quantité (habituellement un paramètre) qui intervient dans la définition de la loi génératrice P. On note habituellement $\hat{\theta}$ ou $\hat{\theta}(D)$ un estimateur de θ. C'est une statistique (donc une fonction) des données D.

- **Biais d'un estimateur:** $E_D[\hat{\theta}(D)] - \theta$. Si le biais est 0 alors évidemment $E_D[\hat{\theta}(D)] = \theta$.

- **Variance d'un estimateur:** $Var[\hat{\theta}(D)] = E_D[(E_D[\hat{\theta}(D)] - \hat{\theta}(D))^2]$. Dans le cas où θ est un vecteur on voudra parler de la trace de la matrice de covariance, donc de la somme des espérances des carrés.

- **Estimateur efficace:** de variance minimale tout en étant non-biaisé (= de biais nul).

- **Asymptotiquement XXX:** une propriété qui est vérifiée à la limite quand le nombre d'exemples tend vers ∞. Si un estimateur est asymptotiquement non-biaisé on dira aussi qu'il est consistant.

- **Vraisemblance:** la probabilité des données D selon une loi $\hat{P}$ vue comme une fonction de $\hat{P}$ ou de ses paramètres. Donc c'est la fonction dont la valeur est $\hat{P}(D)$ (pour un D fixe).

- **Estimateur du maximum de vraisemblance:** si θ est paramètre de $\hat{P}$ alors c'est $\text{argmax}_\theta \hat{P}(D)$. S'il existe une valeur du paramètre qui correspond au véritable processus générateur, alors cet estimateur est **asymptotiquement non-biaisé et asymptotiquement efficace** (donc aussi consistant).

- **Lois des grands nombres:** si $X_1, \ldots, X_n$ sont tirés de la même loi P alors la moyenne des X_i converge vers l'espérance $E[X_i] = E[X]$ sous cette loi quand $n \to \infty$. Ceci n'est pas un énoncé formel car on parle de la convergence d'une série de variables aléatoires (les moyennes de n observations).

- **Théorème limite centrale:** soit $X_1, \ldots, X_n$ indépendants d'espérance μ_i et de variance σ_i^2 respectivement, alors la loi de $\frac{(\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2}}$ converge vers la normale (de moyenne 0 et écart-type 1). En particulier si les X_i ont la même espérance μ et variance σ^2, alors la loi de leur moyenne s'approche d'une normale de moyenne μ et de variance σ^2/n. C'est ce qui rend la loi normale si spéciale car beaucoup de phénomènes résultent d'une accumulation d'un grand nombre d'aléas indépendants.

- **Solutions analytiques du maximum de vraisemblance:** (voir la section optimisation). On cherche les valeurs de θ telles que $\frac{\partial \log \hat{P}(D)}{\partial \theta} = 0$, et on vérifie (ou on garde) les solutions pour lesquelles la dérivée seconde est négative. Exemple de l'estimateur d'espérance par la moyenne sous la loi normale: $\log \hat{P}(D) = -0.5 \sum_i (X_i - \theta)^2/\sigma^2 - 0.5 \log(2\pi\sigma)$. Sa dérivée en θ est $\sum_{i=1}^n (X_i - \theta)/\sigma^2$, ce qui donne la solution $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$.

- **Intervalle de confiance:** soit $\hat{\theta}(D)$ un estimateur de θ, donc cette statistique suit une loi $P(\hat{\theta}(D))$. Un intervalle de confiance I à $p\%$ est tel que $P(\hat{\theta}(D) \in I) = p/100$. Comme on ne connaît généralement pas P on construit un estimateur de cette intervalle de confiance, souvent centré en $\hat{\theta}(D)$.

Cela indique une bande dans laquelle on s'attend trouver le vrai θ autour de $\hat{\theta}$.

- **Régression linéaire ordinaire:** soit $D = \{(X_i, Y_i)\}$ et on suppose $E[Y|X] = b + w \cdot X = \tilde{w} \cdot \tilde{X} = (b, w) \cdot (1, X)$. L'estimateur de $\tilde{w}$ pour la régression linéaire ordinaire est obtenu en minimisant la somme des carrés des erreurs, $\sum_i (Y_i - \tilde{w}.\tilde{X}_i)^2$, ce qui donne $(\mathbf{X'X})^{-1}\mathbf{X'Y}$, où $\mathbf{X}$ est la matrice dont les rangées sont les vecteurs $\tilde{X}_i$ et $\mathbf{Y}$ est le vecteur colonne dont les éléments sont les Y_i.

- **Test statistique:** une procédure pour vérifier une hypothèse concernant la loi des données. Habituellement on calcule une statistique $T(D)$ et on vérifie si elle est au-delà d'un certain seuil critique pour prendre cette décision.

- **Hypothèse nulle:** c'est l'hypothèse H_0 qu'on veut tester. Pour pouvoir faire le test il faut idéalement connaître la distribution de la statistique $T(D)$ sous l'hypothèse H_0. Quand on observe des valeurs de T qui ne sont pas plausibles sous cette distribution, on peut présumer que H_0 est fausse. Une hypothèse courante est celle qu'une variable aléatoires X (dont on connait des échantillons $X_1 \dots X_n$) a l'espérance 0. Dans ce cas une bonne statistique de test est la moyenne des X_i, divisée par un estimateur de l'écart-type de cette moyenne (par exemple l'écart-type empirique des X_i divisé par $\sqrt{n}$). Si les X_i sont normaux, on sait que cette statistique suit une loi de Student. Si les X_i sont $i.i.d$ et $n \to \infty$ alors cette statistique devient normale.

- **p-valeur d'un test:** si on va rejeter H_0 quand $T(D)$ est trop grand, la p-valeur du test est $P(T(D) > t|H_0)$ (test d'un seul côté), ou bien $P(|T(D)| > t|H_0)$ (pour un test des deux côtés). En pratique on va parfois devoir utiliser une estimation de la p-valeur car on ne connaît pas toujours P.

- **Seuil critique:** c'est la p-valeur en dessous de laquelle on va considérer que l'hypothèse nulle devrait être rejetée.

- **Niveau d'un test:** si notre calcul de p-valeur est correct et qu'on rejette au seuil correspondant à une p-valeur $= p$, quand les données proviennent réellement de H_0 on devrait rejeter avec une fréquence qui tend vers p (c'est donc le taux d'erreur de rejet quand on ne devrait pas rejeter). Pour évaluer un test on veut comparer cette fréquence empirique avec le niveau idéal p, pour différentes valeurs de p.

- **Puissance d'un test:** même si un test a le bon niveau, il peut être inutile s'il ne permet pas de séparer H_0 d'autres hypothèses pertinentes (les **hypothèses alternatives**). On mesure la puissance d'un test en générant des

données sous une hypothèse alternative et en comptant le nombre de fois que H_0 est (correctement) rejetée.

- **Statistiques suffisantes:** des statistiques T d'un échantillon D provenant d'une famille paramétrique de lois P indexées par un paramètre θ, telles que toute l'information disponible concernant θ dans D est résumée dans T. Par ailleurs, deux échantillons D_1 et D_2 tels que $T(D_1) = T(D_2)$ sont tels que $P(D_1|\theta) = P(D_2|\theta)$ pour tout θ (même vraisemblance). *Exemple: si les éléments de D sont tirés selon une loi normale dont les paramètres sont la moyenne et la variance, des statistiques suffisantes sont la somme des échantillons et la somme de leur carré.*

- **Estimation de densité:** on estime la véritable densité des données par une fonction f. En général on exige que $f > 0$ et $\int f(x)dx = 1$ (mais pas toujours). Si ces contraintes sont satisfaites, on peut mesurer la qualité de la solution par la log-vraisemblance moyenne sur des données non utilisées pour choisir f, i.e. $\sum_i \log f(x_i)$.

- **Bootstrap:** si on ne connait pas la loi d'une statistique $T(D)$ mais qu'on connaît la loi de D on peut approximer la loi de T par simulations (Monte-Carlo). Par exemple on peut estimer la variance de T par la variance empirique de $T(D')$ sous les différents D' tirés par Monte-Carlo sur P. Mais si on ne connaît pas la loi de D on peut approximer cela avec le Bootstrap en tirant d'une autre loi qui approxime celle de D, et qui permet de tirer des versions perturbées de D. Par exemple si on sait que les X_i dans D sont i.i.d. on peut utiliser le bootstrap non-paramétrique: on tire un nouvel ensemble d'apprentissage D' en choisissant pour le i-ème élément de D' un élément de D pris au hasard (tirage avec remplacement).

- **Modèle paramétrique:** on approxime P (ou une caractéristique de P, comme $E[Y|X]$) avec un estimateur $\hat{P}$ qui est une fonction définie par un nombre fini et fixe de paramètres θ (ne dépend pas de n). En statistique on va même jusqu'à supposer que cette classe de fonctions inclut P.

- **Modèle non-paramétrique:** tout ce qui n'est pas paramétrique. En général notre modèle va avoir un nombre de degrés de libertés proportionnel au nombre d'exemples.

- **Dérivée comme une limite:** $\frac{df(x)}{dx} = \lim_{\epsilon \to 0} \frac{f(x+\epsilon)-f(x)}{\epsilon}$. On peut aussi estimer une dérivée par cette formule, mais un estimateur qui converge beaucoup plus vite est $\frac{f(x+\epsilon)-f(x-\epsilon)}{2\epsilon}$.

- **Dérivée partielle:** si on a une variable y qui dépend de plusieurs variables $x_1, \dots x_n$, la dérivée de y par rapport à x_i en gardant les autres fixes s'écrit $\frac{\partial y}{\partial x_i}$.

- **Dérivée en chaîne et dérivée totale:** si y dépend de plusieurs variables x_i et chacune d'elle dépend d'une variable z, alors une variation dz de z entraîne une variation totale de y (dérivée totale) $dy = \sum_i \frac{\partial y}{\partial x_i}\frac{\partial x_i}{\partial z}dz$. La règle de dérivée en chaîne s'écrit alors $\frac{\partial y}{\partial z} = \sum_i \frac{\partial y}{\partial x_i}\frac{\partial x_i}{\partial z}$. Notons par exemple que si $y = f(x, g(x), z)$ et z ne dépend pas de x alors $\frac{\partial y}{\partial x} = \frac{\partial f(x,g,z)}{\partial x} + \frac{\partial f(x,g,z)}{\partial g}\frac{\partial g(x)}{\partial x}$, où $\frac{\partial f(x,g,z)}{\partial x}$ dénote la dérivée de f par rapport à son premier argument, en gardant les autres fixes, au point $(x, g(x), z)$.

- **Interprétation des premières et secondes dérivées:** $\frac{\partial f(x)}{\partial x}$ est la **pente** de la fonction $f(\cdot)$ au point x; une valeur positive indique qu'une augmentation de x entraînerait une augmentation de $f(x)$. La dérivée seconde $\frac{\partial^2 f(x)}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial f(x)}{\partial x}$ est la variation de pente au point x, correspondant à la courbure (courbure 0 $\rightarrow$ courbe affine = plate $\rightarrow$ pente constante). Quand x est un vecteur de dimension > 1, la dérivée $\frac{\partial f(x)}{\partial x}$ est le **vecteur de gradient**: il indique la direction en x dans laquelle une variation positive de x entraînerait localement la plus grande augmentation de $f(x)$. Le produit scalaire du gradient avec le vecteur v indique la pente dans la direction v. La matrice de dérivées secondes (**Hessien**) $H = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ indique comment est la courbure dans toutes les directions autour de x: le produit matrice vecteur Hv donne la variation du gradient dans la direction v, et la **forme quadratique** $v'Hv$ donne la courbure dans cette direction.

- **Points critiques d'une fonction:** ils sont donnés par la **condition du 1er ordre** $\{x : \frac{\partial f(x)}{\partial x} = 0\}$. Ce sont soit des minima, des maxima, ou des points selle. Une fonction n'a pas toujours de minimum ou de maximum (p.e. la fonction augmente toujours quand son argument tend vers ∞).

- **Conditions du 2ème ordre:** pour distinguer les points critiques (min, max, point selle) il suffit de considérer la dérivée seconde. Dans $\mathbf{R}$, le signe de $\frac{\partial^2 f}{\partial x^2}$ donne la réponse: positif $\rightarrow$ minimum, 0 $\rightarrow$ point selle, négatif $\rightarrow$ maximum.

- **Géométrie des points critiques en dimension > 1:** En dimension plus élevée on peut obtenir des signes différents selon la direction considérée. Pour caractériser le point critique il faut considérer les valeurs propres λ_i

de la matrice Hessienne: leur signe indique la courbure dans la direction du vecteur propre v_i correspondant.

- **Minimum local et global:** il peut y avoir plusieurs minima (maxima) locaux x_i. Certains d'entre eux sont appelés minima (maxima) globaux si $f(x_i) \leq f(x) \forall x$.

- **Optimisation:** la recherche d'optima (minima ou maxima) locaux ou globaux d'une fonction.

- **Optimisation sous contrainte:** on cherche une ou l'ensemble des valeurs de $x \in C$ qui minimisent $f(x)$, où C est un ensemble de valeurs qui indique des contraintes. On considère principalement les contraintes d'égalité et les contraintes d'inégalités.

- **Multiplicateurs de Lagrange:** si on cherche un extrême de $f(x)$ sous la contrainte $g(x) = 0$ alors on peut considérer l'optimisation de $L(x, \lambda) = f(x) - \lambda \cdot g(x)$ en x et λ. Ça fonctionne pour x ou g des vecteurs (auquel cas λ aussi en est un, de la même dimension que g). Pour les contraintes d'inégalité $g(x) \geq 0$ on obtient en plus les conditions de Kuhn-Tucker, soit que $\lambda_i \geq 0$ et $\lambda_i g_i(x) = 0$ (i.e. soit la contrainte g_i est atteinte ou $\lambda_i = 0$).

- **Fonction convexe:** si sa courbure est positive (dans toutes les directions). Concave: si sa courbure est négative dans toutes les directions (ou si son opposé est convexe). Il existe donc des fonctions qui ne sont ni concave ni convexe ou qui peuventt être convexe seulement dans certaines régions.

- **Ensemble convexe:** ssi tous les points sur le segment joignant deux points sont aussi dans l'ensemble, i.e. $\forall x, y \in C, \forall 0 \leq a \leq 1$, $ax + (1 - a)y \in C$.

- **Inégalité de Jensen:** pour f convexe, $E[f(X)] \geq f(E[X])$. Illustration simple et graphique: $\frac{1}{2}(f(x_1) + f(x_2)) \geq f(\frac{1}{2}(x_1 + x_2))$. Le sens de l'inégalité est inversé pour les fonctions concaves.

- **Descente de gradient:** dans le but de minimiser (ou tout au moins réduire) $f(x)$, on peut faire un pas dans la direction opposée au gradient $\frac{\partial f(x)}{\partial x}$. Si le gradient est non-nul, il existe forcément un ϵ t.q. $f(x - \epsilon \frac{\partial f(x)}{\partial x}) < f(x)$.

- **Optimisation de Newton:** le gradient est la direction de descente la plus rapide mais si on fait un pas non-infinitésimal et que la fonction est localement convexe il vaut mieux faire un pas dans la direction $H^{-1} \frac{\partial f(x)}{\partial x}$ où H est la matrice Hessienne de f au point x. Si $f(x)$ est quadratique ce pas nous amène

directement au minimum. Si $f(x)$ est convexe ce pas nous garantit une descente très rapide. Cependant en pratique on utilise rarement cette méthode car (a) si H est mal conditionnée ça ne fonctionne pas numériquement, (b) le calcul de H est coûteux si x est de dimension élevée, (c) si f n'est pas strictement convexe on peut partir dans les choux.