Probabilité et Statistiques

Tronc commun aux Master 1 Big Data et Intelligence Artificielle de l'Université Virtuelle du Sénégal

Jean-Michel Amath Sarr

Table des matières

Chapitre Un

Introduction à l'estimation paramétrique

Sommaire

1.1 Prérequis

1.1.1 Définition de base

Définition : *espace probabilisé*, espace probabilisable.
Un espace probabilisé est un triplet $(\Omega, \mathcal{F}, \mathbb{P})$. Ω est l'univers, ou l'ensemble des événements possibles, $\mathcal{F}$ est une σ-algèbre et $\mathbb{P}$ est une mesure de probabilité. De même on appelle espace probabilisable un couple $(\Omega, \mathcal{F})$

Définition : $\sigma-$ *Algèbre.*
Soit Ω un ensemble, une σ-algèbre $\mathcal{F}$ est un ensemble de partie de Ω vérifiant trois règles :

- $\Omega \in \mathcal{F}$
- $\forall A \in \mathcal{F}, A^c \in \mathcal{F}$
- $\forall A_1, \cdots, A_n, \cdots \in \Omega, \bigcup_{i>1} A_i \in \mathcal{F}$

Exemples

- La tribu pleine : $\mathcal{P}(\Omega)$
- la tribu triviale $\{\emptyset, \Omega\}$
- la tribue borélienne $\mathcal{B}(\mathbb{R}) = \{]a, b[,]a, b[\subset \mathbb{R}\}$ formée des ouverts de $\mathbb{R}$

Exercice :

- Soit un jeu de pile ou face effectué deux fois, quelle est l'univers et la tribu résultante
- Soit un lancé de dé effectué deux fois, quelle est l'univers et la tribu résultante

Définition : *probabilité.*
On appelle loi de probabilité, ou tout simplement probabilité une mesure $\mathbb{P} : \mathcal{F} \to [0, 1]$ vérifiant :

- $\mathbb{P}(\Omega) = 1$ la probabilité de l'événement certain est 1

— $\mathbb{P}(\bigcup_{i>1} A_i) = \sum_i \mathbb{P}(A_i)$ pour des événements deux à deux disjoints.

Remarque : Si A est un événement de Ω alors $\mathbb{P}(A) = \frac{Cardinal(A)}{Cardinal(\Omega)}$.

Définition : *indépendance de deux événements.*
Deux événements A, B sont indépendants si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Définition : *probabilité conditionnelle.*
Soit $A, B \in \Omega$ deux événements. On appelle probabilité conditionelle de A sachant B et on note $\mathbb{P}(A|B)$ la quantité suivante : $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

1.1.2 Variable Aléatoires

Définition : *variable aléatoire.*
Soit un espace probabilisable $(\Omega, \mathcal{F})$ et un espace mesurable $(E, \mathcal{E})$. Une variable aléatoire, abbréviée *v.a*, est une fonction mesurable X de $(\Omega, \mathcal{F})$ à valeur dans $(E, \mathcal{E})$. Autrement dit c'est une fonction qui vérifie la propriété suivante :

$$\forall A \in \mathcal{E}, X^{-1}(A) \in \mathcal{F}$$

Lorsque $E = \mathbb{N}$ on parle de variable aléatoire discrète, si $E = \mathbb{R}$, on parle de variable aléatoire continue. Lorsque E est un espace de plusieurs dimension, on parle de vecteur aléatoire, ou de variable aléatoire vectorielle. Une variable aléatoire X défini une loi de probabilité sur $(\Omega, \mathcal{F})$ souvent noté P_X. Ainsi $\forall B \in \mathcal{E}\ P_X(E) = \mathbb{P}(X^{-1}(B))$. En fait variable aléatoire est simplement une fonction pour representer des évènements aléatoires sous forme numérique.

Exemple(s) : Dans le cadre d'un jeu de pile ou face l'univers $\Omega = \{P, F\}$, typiquement, une v.a va associer chaque évènement à un nombre par exemple $\{0, 1\}$. On peut alors attribuer une valeur à chaque évènement à l'aide d'une v.a. Ainsi $\mathbb{P}(X = 0) = \mathbb{P}(P)$ et $\mathbb{P}(X = 1) = \mathbb{P}(F)$. Une loi permet de modéliser cet expérience : il s'agit de la loi de Bernoulli.

Exercice : Donnez la loi associé au lancé d'un dé équilibré.
Définition : *fonction de répartition.*
Soit X une v.a définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. On appelle fonction caractéristique associée à X la fonction $F_X(x) := \mathbb{P}(X < x)$.

Définition : *espérance d'une v.a.*
Soit X une v.a définie sur $(\Omega, \mathcal{F}, \mathbb{P})$, on appelle espérance de X la quantité suivante lors-

qu'elle est définie :

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) \text{ si X une v.a discrète} \tag{1.1}$$

$$\mathbb{E}[X] = \int_{\omega} X(\omega) d\mathbb{P}(\omega) \text{ si X une v.a continue} \tag{1.2}$$

Définition : *variance d'une v.a.*
On appelle variance d'une v.a X la quantité suivante :

$$Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Définition : *indépendance de variables aléatoires*
Deux variables aléatoires X, Y sont indépendantes si

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \text{ dans le cas continu}$$
$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \text{ dans le cas discret}$$

Exercice

- Soit deux variables aléatoires de loi jointe $f_{X,Y}(x, y) = 4xy\mathbb{1}_{[0,1]}(x)\mathbb{1}_{[0,1]}(y)$, vérifier que c'est bien une densité de probabilité.
- Montrer que les v.a X et Y sont indépendantes.

Le théorème suivant permet de s'affranchir du calcul des lois marginales pour montrer que deux variables sont indépendantes.

Théorème 1.1. *Deux variables aléatoires X, Y sont indépendants si et seulement si il existe deux fonctions g, h telles que :*

$$f_{X,Y}(x, y) = g(x)h(y)$$

1.1.3 Convergence

Dans la suite on se place dans un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. Soit (X_n) une suite de v.a et X une v.a à définies dans l'espace probabilisé. Nous allons énoncer plusieurs types de convergence utiles par la suite.

Définition : *convergence presque sûre.*
(X_n) converge presque sûrement vers X et on note $X_n \xrightarrow{p.s} X$, lorsque la convergence est vrai avec une probabilité égale à 1.

$$\mathbb{P}(\lim_{n \to \infty} X_n = X) = 1$$

.
Autrement dit, la convergence est presque sûre si les événements ω pour lesquels la suite (X_n) ne converge pas ont une probabilité nulle.

$$\mathbb{P}(\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)) = 1$$

Définition : *convergence en moyenne quadratique.*
(X_n) converge en moyenne quadratique vers X et on note $X_n \xrightarrow{L^2} X$ lorsque :

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0$$

Définition : *convergence en probabilité.*
(X_n) converge en probabilité vers X et on note $X_n \xrightarrow{\mathcal{P}} X$ lorsque :

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

Définition : *convergence en loi.*
Soit $(F_{n)}$ la suite de fonction de répartition associé à la suite de v.a (X_n) et F la fonction de répartition associé à X. On dit que X_n converge en loi vers X et on note $X_n \xrightarrow{\mathcal{L}} X$ lorsque pour tout point x ou F est continue, $F_n(x)$ converge vers $F(x)$

1.1.4 Fonction génératrice de moments et application

DéfinitionSoit X une v.a, on appelle fonction génératrice de moments et on note ψ_X ou M_X la fonction suivante :

$$\psi_X(t) = \mathbb{E}[e^{tX}] = \begin{cases} \int_{S_X} e^{tx} f_X(x)dx \text{ dans le cas continu} \\ \sum_{S_X} e^{tx} p_X(x) \text{ dans le cas discret} \end{cases}$$

Avec S_X le support de X.
Propriété
- $\psi_{aX+b}(t) = e^{bt}\phi_X(at)$
- Si $X_1, \cdots, X_n$ sont indépendantes, alors $\psi_{\sum_i X_i}(t) = \prod_i \psi_{X_i}(t)$

La preuve est donnée en cours. Cette fonction peut être utilisée pour calculer une somme de v.a, on peut montrer ainsi qu'une somme de loi Gamma, est encore une loi Gamma, idem pour une somme de Bernoulli, de Binomiale, ou de loi de Poisson.

Théorème 1.2. *Soit deux v.a X, Y et f_X, f_Y leur densités associés, ψ_X, ψ_Y leur fonction génératrice de moments. Alors*

$$f_X(t) = f_Y(t)\forall x \Leftrightarrow \psi_X(t) = \psi_Y(t)\forall t$$

La preuve sera donné en cours pour le cas discret. Ce que ce théorème nous dit, c'est qu'on peut caractériser une variable aléatoire par sa fonction génératrice de moments au même titre qu'on peut la caractériser par sa densité.

1.1.5 Transformation de variables aléatoires

Dans cette sous section on va donner un théorème important pour calculer une transformée de variable aléatoire. L'idée de base revient à utiliser le rapport entre la fonction de répartition et la densité. En effet on sait d'après le théorème fondamental de l'analyse que $\frac{dF_X(x)}{dx} = f_X(x)$. Donc si on a une v.a X qui est transformée par une fonction u et donne $u(X) = Y$, on peut calculer la densité de Y avec le théorème suivant :

Théorème 1.3. *Soit X une v.a continue sur $(\Omega, \mathcal{F}, \mathbb{P})$ avec comme densité $f_X \cdot \mathbb{1}_S$ avec S le support de X. Si u est strictement monotone et v est son inverse, alors la v.a $Y = u(X)$ a pour densité :*

$$f_Y(y) = f_X(v(y)) \left| \frac{dv(y)}{dy} \right| \mathbb{1}_{u(S)}(y)$$

Nous donnons deux preuves dans le cours.

Exemple(s)

— Soit X une v.a ayant la densité $f_X(x) = 2x \cdot \mathbb{1}_{[0,1]}(x)$, calculez la densité de $Y = \sqrt{X}$.

— $X \sim \mathcal{U}(-1, 3)$, calculer la densité de $Y = X^2$

Exercice

1. $X \sim \mathcal{U}(-1, 1)$, calculer la densité de $Y = e^X$
2. $X \sim \mathcal{U}(0, 1)$ calculez la densité de $Y = aX + b$ avec $a > 0$.
3. $X \sim \mathcal{N}(\mu, \sigma^2)$, montrez que $Y = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
4. X une v.a de densité $f_X(x) = \frac{x^2}{3} \cdot \mathbb{1}_{[-1,2]}(x)$, calculez la densité de $Y = X^2$.

1.2 Modèlisation statistique

1.2.1 Modèle paramétrique

Un modèle statistique est un objet mathématique associé à l'observation des données provenant d'un phénomène aléatoire.

Définition : *modèle statistique.*
Un modèle statistique est la donne d'un espace probabilisable $(\Omega, \mathcal{F})$ et d'une famille de loi de probabilité $\mathcal{L}$. Lorsque cette famille de loi peut s'écrire sous la forme $\mathbb{P}_\Theta := \{\mathbb{P}_\theta, \theta \in \Theta\}$ (Θ est l'espace des paramètres) on parle de modèle statistique paramétrique ou plus simplement de modèles paramétriques. Sinon on parle de modèles non-paramétriques.

Une expérience statistique consiste à observer des données souvent représentées par une suite $x_1, \cdots, x_n$ et à les interpréter avec un modèle statistique. Cela revient à voir les données x_i comme des réalisations de v.a X_i. Dans ce cours nous nous concentrerons exclusivement sur les modèles paramétriques.

Définition : *expérience statistique.*
Une expérience statistique est la donne d'un modèle paramétrique $(\Omega, \mathcal{F}, \mathbb{P}_\Theta)$ et d'une variable aléatoire X à définie dans $(\Omega, \mathcal{F})$.

Définition : *identifiabilité d'un modèle paramétrique*
Un modèle paramétrique est dit identifiable si $\forall \theta, \theta' \in \Theta$

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \implies \theta = \theta'$$

Exercice : Montrer que le modèle $\mathcal{P} = \{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$ est identifiable.

Deux types d'estimation existent : l'estimation ponctuelle et l'estimation par intervalle de confiance. Nous allons nous concentrer dans les premiers chapitres sur l'estimation ponctuelle. Vers la fin du cours nous ferons une ouverture sur l'estimation par intervalle de confiance.

1.3 Estimateurs

1.3.1 Généralités

Dans l'interprétation fréquentiste des phénomènes aléatoires on considère que la loi est connue, mais que le paramètre est inconnu mais fixe. Un estimateur est donc une fonction permettant d'estimer le paramètre inconnu à partir des données.

Définition : *statistique.*
Soit un espace mesurable $(E, \mathcal{E})$, une statistique est une fonction mesurable T quelquonque défini sur $(E, \mathcal{E})$.

Exemple(s) : Soit $x_1, \cdots, x_n \in E$
- $T(x_1, \cdots, x_n) = \max(x_1, \cdots, x_n)$
- $T(x_1, \cdots, x_n) = \sum_{i=1}^{n} x_i$

Remarque : Les définitions de variable aléatoire et de statistique sont pratiquement les même. En pratique cependant, une statistique est définie sur l'espace de sortie des variables aléatoires.

Définition : *estimateur.*
Soit une expérience statistique associée à un modèle paramétrique $(\Omega, \mathcal{F}, \mathbb{P}_\Theta)$ et $X_1, \cdots, X_n$ des v.a identiquement et indépendemment distribuées. Un estimateur T_n du paramètre θ est une statistique définie sur $\Omega, \mathcal{F}$ à valeur dans l'espace des paramètres possibles Θ.

$$T_n : \Omega^n \to \Theta \tag{1.3}$$
$$(x_1, \cdots, x_n) \mapsto t_n \tag{1.4}$$

Une estimation de θ est une réalisation t_n de l'estimateur T_n. Il faut bien distinguer un estimateur qui est une v.a et une estimation qui est une valeur déterministique. En effet un estimateur est une v.a car c'est une fonction de v.a, donc ses réalisations varient en fonction des valeurs d'entrée. Les méthodes classiques pour construire les estimateurs sont la méthodes des moments et la méthode du maximum de vraisemblance. Nous nous intéresseront à cetter dernière méthode dans ce cours. Dans la suite on considère que les données sont identiquement et indépendemment distribuées i.i.d. C'est à dire que les réalisations sont de même loi et indépendantes.

Exercice : Exprimiez la différence entre estimateur et estimation en utilisant l'exemple discuté dans le TP : 'Notions de base de modélisation statistiques'.

Exemple(s)
Soit $X_1, \cdots, X_n$ une suite de v.a

— L'estimateur de la moyenne souvent noté $\bar{X}_n$ est donné par la formule suivante :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

— L'estimateur de la variance empirique est la moyenne des carrés des écarts à la moyenne empirique :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

— L'estimateur de la variance empirique corrigé :

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

1.3.2 Construction d'un estimateur

Définition : *fonction de vraisemblance.*
Soit $(x_1, \cdots, x_n)$ un échantillon issu d'une expérience statistique. On appelle fonction de vraisemblance ou tout simplement vraisemblance la fonction suivante

$$\mathcal{L}(x_1, \cdots, x_n; \theta) = \mathbb{P}_\theta(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i)$$

Ainsi la vraisemblance est la probabilité d'observer les données en fixant le paramètre θ.

Définition : *estimateur du maximum de vraisemblance.*
L'estimateur du maximum de vraisemblance $\hat{\theta}_n$ est un estimateur permettant de trouver la valeur de θ rendant la vraisemblance maximale.

$$\hat{\theta}_n = \arg\max_\theta \mathcal{L}(x_1, \cdots, x_n; \theta) = \arg\max_\theta \log \mathcal{L}(x_1, \cdots, x_n; \theta)$$

En pratique on utilise maximize log-vraisemblance, en effet le logarithme est une fonction croissante.

Définition : *score.*
On appelle score le gradient de la log-vraisemblance.

$$s(\theta) = \frac{\partial \log \mathcal{L}(x_1, \cdots, x_n; \theta)}{\partial \theta}$$

En pratique on utilise le score pour trouver le maximum de vraisemblance en cherchant la valeur de θ annulant le score.

Exercice : Soit $x \sim \mathcal{N}(\mu, \sigma^2)$.

1. Donnez l'expression du score .
2. Montrer que l'espérance du score est nul de manière générale.
3. On considère un jeu de données $x_1, \cdots, x_n$ i.i.d suivant $\mathcal{N}(\mu, \sigma^2)$. Exprimez la vraisemblance de cette loi et donnez une estimation du maximum de vraisemblance pour μ et σ^2.
4. Refaire la question 3 pour une loi de Bernoulli.

Chapitre Deux

Performances

Sommaire

Dans ce chapitre nous allons étudier comment comparer des estimateurs. Un estimateur T_n de θ sera un bon estimateur s'il s'approche suffisamment du paramètre dans un sens que nous allons préciser.

2.1 Qualité d'un estimateur

2.1.1 Risque quadratique, biais et variance

Définition : *risque quadratique.*
Le risque quadratique ou erreur quadratique moyenne d'un estimateur T_n est donné par l'expression suivante :

$$EQM(T_n) = \mathbb{E}[(T_n - \theta)^2]$$

Le risque quadratique permet de comparer deux estimateurs, c'est une mesure assez brute de la qualité d'un estimateur qui peut être raffiné avec les notions de biais et de variance.

Définition : *bias d'un estimateur*
On appelle biais de T_n pour θ la valeur $b_\theta(T_n) := \mathbb{E}(T_n) - \theta$
Par conséquent on dira d'un estimateur T_n de θ qu'il est sans biais si et seulement si $\mathbb{E}(T_n) = \theta$, dans le cas contraire, on dira que T_n est un estimateur biaisé. En d'autres termes un estimateur est sans biais si son espérance est égale au paramètre recherché.

Exercice : Montrer que $EQM(T_n) = Var(T_n) + b_\theta(T_n)^2$

Le risque quadratique d'un estimateur est égal à sa variance plus le carré de son biais. Lorsqu'on évalue plusieurs estimateurs le meilleur sera sans biais et de variance minimale. Cependant, le biais et la variance ne sont pas les seuls propriété caractérisant la qualité

d'un estimateur. En effet, augmenter la taille de l'échantillon est aussi à prendre en considération quand on compare des estimateurs. D'ailleurs le succès du big data revient à utiliser d'énormes échantillons, ce qui conduit à bénéficier des performances asymptotiques des estimateurs.

Exercice : On considère une variable aléatoire X telle que $\mathbb{E}[X] = \mu$ et $Var(X) = \sigma^2$.

- Montrer que les estimateurs de la moyenne et de la variance d'une loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ que vous avez trouvé à l'aide du maximum de vraisemblance sont sans biais.
- Montrer que l'estimateur de la variance empirique $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ est biaisé. Pour cela commencez par développer et simplifier cette somme.
- Déduisez en un estimateur de la variance non biaisé et retrouvez l'estimateur de la variance empirique corrigé présenté plus haut $\hat{S}_n^2$.

2.1.2 Convergence des estimateurs

Définition : *convergence d'un estimateur.*
Un estimateur T_n de θ est dit convergeant s'il converge en probabilité vers θ lorsque n tend vers l'infini. On notera $T_n \xrightarrow{\mathcal{P}} \theta$

Définition : *convergence en moyenne quadratique.*
L'estimateur T_n converge en moyenne quadratique vers θ si et seulement si son erreur quadratique moyenne tend vers 0 quand n tend vers l'infini :

$$T_n \xrightarrow{L^2} \theta \Leftrightarrow \lim_{n\to\infty} \mathbb{E}[(T_n - \theta)^2] = 0$$

Remarque : Lorsqu'un estimateur est convergent, son biais converge vers 0. On dit aussi qu'il est asymptotiquement sans biais.

Théorème *loi faible des grands nombres*
Soit $X_1, \cdots, X_n$ un ensemble de variables aléatoires $i.i.d$ d'espérance finie $\mathbb{E}[X_i] = \mu$, alors l'estimateur de la moyenne empirique converge en probabilité vers la moyenne μ

$$\bar{X}_n \xrightarrow{\mathcal{P}} \mu$$

Théorème *loi forte des grands nombres*
Sous les mêmes hypothèses que le théorème précédent

$$\bar{X}_n \xrightarrow{p.s} \mu$$

Remarque :

- Les deux lois des grands nombres justifient l'interprétation fréquentiste des probabilités. C'est l'intuition selon laquelle, l'observation d'un grand nombre de données issues d'un phénomène aléatoire permet de déterminer les paramètres de la loi sous jacente. L'exemple le plus direct est le lancé d'une pièce équilibré pour obtenir pile

ou face. À mesure que vous augmentez le nombre de lancés, vous obtiendrez une probabilité convergent vers 0.5. Ces théorèmes justifient théoriquement les big data pour l'estimation paramétrique.

— Vous noterez cependant que ces théorèmes concernent l'estimateur de la moyenne empirique. Cela laisse suggérer que si vous construisez votre propre estimateur à l'avenir, il faudra peut-être prouver que celui ci converge bien. Enfin, vous observerez que la seule différence entre ces deux théorèmes et la nature de la convergence sous jacente.

— Bien que la loi faible des grands nombres nous indique que la distribution de $\bar{X}_n$ se concentre autour de $\mathbb{E}[X]$, elle ne nous dit rien sur la loi que suit cet estimateur. Le théorème de la limite centrale (TCL) dit que $\bar{X}_n$ a une distribution qui est approximativement normale avec une moyenne $\mathbb{E}[X]$ et une variance $Var(X)$. Étonnamment, rien n'est supposé au sujet de la loi de X, sauf l'existence de la moyenne et de la variance.

Théorème *théorème central limite*
Soit (X_i) une suite de v.a i.i.d et $\bar{X}_n$ l'estimateur de la moyenne empirique, alors on a le résultat suivant :

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Var(X))$$

Le point clé du théorème central limite est de donner la loi de l'estimateur
Exercice :

— Montrez que les estimateurs de la moyenne empirique et de la variance empirique sont convergent en moyenne quadratique.

— Montrez l'équivalence entre les deux formulations suivantes du TCL :

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Var(X)) \Leftrightarrow \sqrt{n}(\frac{\bar{X}_n - \mathbb{E}(X)}{\sigma}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

— Donnez un diagramme d'implication comparant les 4 types de convergence suivants : convergence en loi, convergence en probabilité, convergence presque sûre, et convergence quadratique. Le but de cet exercice est de vous donner un aperçu de haut niveau des propriétés de convergence.

2.1.3 Efficacité d'un estimateur

Définition : *information de Fisher.*
On appelle quantité d'information de Fisher sur θ apporté par l'échantillon $(x_1, \cdots, x_n)$ la matrice de covariance du score :

$$I_n(\theta) = Var[s(\theta)] = Var\left[\frac{\partial \log \mathcal{L}(x_1, \cdots, x_n; \theta)}{\partial \theta}\right]$$

Proposition 2.1. *Inégalité de Cramer-Rao.*
Si T_n est un estimateur sans biais alors :

$$Var(T_n) \geq \frac{1}{I_n}$$

Ainsi une propriété de la quantité d'information de Fisher est qu'elle procure une borne inférieure à tout estimateur sans biais.

Définition : *borne de Cramer-Rao.*
La quantité $\frac{1}{I_n(\theta)}$ est appelée la borne de Cramer-Rao.
La borne de Cramer-Rao nous informe sur la variance minimale que l'on peut obtenir d'un estimateur sans biais.

Définition : *efficacité.*
On appelle efficacité d'un estimateur T_n sans biais de θ la quantité suivante :

$$Eff(T_n) = \frac{1}{I_n(\theta)Var(T_n)}$$

L'efficacité est une grandeur comprise entre 0 et 1, observez que cette quantité se rapproche de 0 lorsque la variance est très grande. Lorsque la variance est égale à la borne de Cramer-Rao, on est en présence d'un estimateur efficace. Autrement dit si $\hat{\theta}_1$ et $\hat{\theta}_2$ sont des estimateurs tels que $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$ alors $\hat{\theta}_1$ est plus efficace que $\hat{\theta}_2$. Lorsqu'un estimateur atteint l'efficacité quand n tend vers l'infini, on dit que c'est un estimateur asymptotiquement efficace. L'orsqu'un estimateur est sans biais et efficace, on dit aussi qu'il est sans biais et de variance minimale. En pratique pour comparer deux estimateurs avec un échantillon de taille n fixé on choisit celui qui à l'erreur quadratique moyenne la plus faible, i.e sans biais et efficace. Si on a accès à une infinité de données, on peut comparer leurs propriétés de convergence. Exemple, l'estimateur est-il asymptotiquement sans-biais, efficace ?

Exercice : Considérons $X_1, \cdots, X_n$ des variables aléatoires i.i.d.

1. Montrer que $I_n(\theta) = nI_1(\theta)$
2. Montrer que $I_1(\theta) = \mathbb{E}[(\nabla \log f(x;\theta))^2]$
3. Montrez la relation $\nabla^2_\theta \log(f(x;\theta)) = \frac{\nabla^2_\theta f(x;\theta)}{f(x;\theta)} - (\frac{\nabla_\theta f(x;\theta)}{f(x;\theta)})^2$
4. Déduisez-en cette écriture alternative $I_1(\theta) = -\mathbb{E}[\nabla^2_\theta \log \mathcal{L}(X_1;\theta)]$
5. En pratique on utilise souvent cette dernière expression pour calculer l'information de Fisher. Montrez qu'une variable aléatoire de Bernoulli de paramètre p à pour information de Fisher l'expression : $\frac{1}{p(1-p)}$.

2.1.4 Exhaustivité

On rappelle qu'un estimateur est une statistique, i.e une fonction des données, par exemple $\bar{X}_n, S^2_n$. Mais aussi $T : (X_1, \cdots, X_n) \to 1$ est une statistique.

Évidemment, il y a beaucoup de fonctions de $X_1, \cdots, X_n$ et donc beaucoup de statistiques. Lorsque nous recherchons un bon estimateur, devons-nous vraiment les considérer tous, ou existe-t-il un ensemble de statistiques beaucoup plus restreint que nous pourrions envisager ? Une autre façon de poser la question est de savoir s'il existe quelques fonctions clés de l'échantillon aléatoire qui contiennent eux-mêmes toutes les informations contenues

dans l'échantillon. Par exemple, supposons que nous connaissions la moyenne de l'échantillon et la variance de l'échantillon. L'échantillon aléatoire contient-il plus d'informations sur la population que cela ? La réponse cette question dépendra de la famille de loi de probabilité que nous supposons décrire la population. Commençons par une définition heuristique d'une statistique suffisante.

Définition : *statistique exhaustive.*
Soit un modèle statistique $(\Omega, \mathcal{F}, \mathbb{P}_\Theta)$, un échantillon $X_1, \cdots, X_n$ issu de $(\Omega, \mathcal{F})$ et une statistique T. On dit que T est exhaustive si la probabilité conditionnelle de l'échantillon sachant T ne dépend pas du paramètre θ quelque soit la valeur prise par T. Autrement dit $p(X_1, \cdots, X_n | T = t)$ ne dépend pas de θ.

Le statisticien qui connaît la valeur de T peut faire un aussi bon travail d'estimation du paramètre inconnu θ que le statisticien qui connaît l'ensemble de l'échantillon aléatoire. Montrons cela avec un exercice.

Exercice : *Exemple loi Bernoulli*
Soit $X_1, \cdots, X_n$ des v.a de Bernoulli de paramètre p i.e $\mathcal{B}(p)$, montrez que $T : (X_1, \cdots, X_n) \to \sum_{i=1}^n X_i$ est une statistique exhaustive pour le paramètre p.
Indices
- Commencez par exprimer $p(X_1, \cdots, X_n, T(X_1, \cdots, X_n) = t)$.
- Exprimez aussi $p(T(X_1, \cdots, X_n) = t)$.

Cependant en pratique, on ne vérifie pas directement ce critère, on utilise plutôt, le théorème de factorisation que nous présentons ci après :

Théorème 2.1. *Pour qu'une statistique T soit exhaustive pour θ, il faut et il suffit qu'il existe deux fonctions g et h mesurables telles que*

$$\forall x \in \Omega, \forall \theta \in \Theta,\ \mathcal{L}(x; \theta) = g(T(x), \theta) h(x)$$

Exercice : Trouver une statistique exhaustive pour
- une loi normale $\mathcal{N}(\mu, \sigma^2)$.
- une loi de Poisson $P(\lambda)$.
- La loi à densité suivante : $f(x|\theta) = \begin{cases} \theta x^{\theta-1} \text{ pour } x \in]0,1[\\ 0 \text{ sinon} \end{cases}$

2.2 Propriétés de l'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance n'est pas forcément unique, ni sans biais, ni efficace. Cependant, il possède d'excellentes propriétés asymptotiques.

Propriété Soit (X_i) une suite de v.a i.i.d suivant une loi P_θ, cette loi vérifiant certaines conditions de régularité, et T_n l'estimateur du maximum de vraisemblance associé à un n-échantillon. Alors on a :

— $T_n \xrightarrow{p.s} \theta$, T_n converge presque sûrement vers θ
— $\sqrt{I_n(\theta)}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, autrement dit quand n tend vers l'infini T_n est approximativement de loi $\mathcal{N}(\theta, \frac{1}{I_n(\theta)})$. On dit que T_n est asymptotiquement gaussien, sans biais et efficace. Cette propriété peut aussi s'écrire :

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{I_1(\theta)})$$

2.3 Estimation par intervalle de confiance

Dans la section précédente, on se proposait d'estimer uniquement le paramètre θ. On veut maintenant proposer un ensemble $I \subset \Theta$ aussi petit que possible, tel que θ appartienne souvent à I. Dans la suite, on suppose que X est une v.a et que $x_1, \cdots, x_n$ est un échantillon i.i.d tiré suivant la loi de X, on se donne également un estimateur T_n.

Définition : *intervalle de confiance.*
Soit $\alpha \in]0,1[$. S'il existe des v.a.r. $\theta_{min}(X_1, \cdots, X_n)$ et $\theta_{max}(X_1, \cdots, X_n)$ telles que :

$$P(\theta \in [\theta_{min}(X_1, \cdots, X_n), \theta_{max}(X_1, \cdots, X_n)] = 1 - \alpha$$

on dit alors que $[\theta_{min}(X_1, \cdots, X_n), \theta_{max}(X_1, \cdots, X_n)]$ est un intervalle de confiance pour θ, avec coefficient de sécurité $1 - \alpha$. On le note $IC_{1-\alpha}(\theta)$.

En pratique on choisit souvent $\alpha = 5\%$ ce qui emmène $1 - \alpha = 95\%$. Pour construire un intervalle de confiance du paramètre θ, on utilise les inégalités de concentration Markov, Hoeffding, Tchebychev. Ou si on connaît la loi de T_n on peut aussi directement l'utiliser. Sinon on peut utiliser une convergence en loi quand n tend vers l'infini pour construire un intervalle de confiance asymptotique.

Deux résultats utile pour calculer ces intervalles asymptotiques sont le théorème de Slutsky et la proposition suivante
Proposition
Soit $(X_n)_{n>0}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$ qui converge presque sûrement, respectivement en probabilité, vers une variable aléatoire X et soit f une fonction continue de $\mathbb{R}^d$ dans $\mathbb{R}^m$. Alors la suite $f(X_n)$ converge presque sûrement, respectivement en probabilités, vers $f(X)$.

Théorème *Slutsky*
Soient $(X_n)_{n>0}$ et $(Y_n)_{n>0}$ deux suites de variables aléatoires définies sur un même espace de probabilités, à valeurs dans $\mathbb{R}$, telles que (X_n) converge en loi vers une variable aléatoire X et (Y_n) converge en loi vers une variable aléatoire a constante. Alors le couple (X_n, Y_n) converge en loi vers le couple (X, a). On en déduit en particulier que $(X_n + Y_n)$ converge en loi vers $X + a$ et que (X_n, Y_n) converge en loi vers aX.

Exemple(s) : intervalle de confiance d'une moyenne lorsque la variance est connue, lorsqu'elle est inconnue. Intervalle de confiance de la variance. Intervalle de confiance d'une proportion.

Chapitre Trois

Estimation en contexte classique

Sommaire

Les modèles de régression linéaire permettent de rendre compte d'un phénomène aléatoire comme combinaison linéaire ou affine de variables explicatives appelées aussi régresseurs. Par exemple, on peut modéliser la taille d'une personne par une combinaison linéaire de la taille de ses deux parents. Lorsqu'il n'y a qu'une seule variable explicative, on parle de régression univariée, lorsqu'il y a plusieurs variables explicatives on parle de régression multivariée. Dans ce chapitre, nous verrons trois méthode de régression selon l'hypothèse que l'on fair sur les données. Ces méthodes sont :

— La régression linéaire
— La régression linéaire pondéré
— La régression non-linéaire

Lorsque l'on utilise la régression linéaire simple on fait deux hypothèses sur les données : l'hypothèse de linéarité des données, et aussi l'hypothèse d'homoscédasticité. La régression pondérée conserve l'hypothèse de linéarité des données, mais relaxe l'hypothèse d'homoscédasticité, par une hypothèse plus faible : l'hypothèse d'hétéroscédasticité. Enfin la régression non-linéaire est envisagée lorsque l'on relaxe l'hypothèse de linéarité des données. En pratique on vérifie ces hypothèses à l'aide de tests adapté à l'expérience générant les données. Cependant cela sort du cadre de ce cours.

Définition : *homoscédasticité.*
Homoscédasticité signifie "qui a une dispersion identique" Cette notion provient du grec et est composée du préfixe homós ("semblable, pareil") et de skedasê ("dissipation"). On suppose que les erreurs ε_i sont identiquement distribuées et de même loi $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Définition : *hétéroscédasticité.*
Hétéroscédasticitésignifie "qui a une dispersion différente" Cette notion provient du grec et est composée du préfixe préfixe hétéro ("autre"), et de skedasê ("dissipation"). On suppose que les erreurs ne sont pas de même loi $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. En d'autres termes chaque variable

explicative à sa propre erreur.

3.0.1 Notations

Nous avons à traiter des données S modélisées comme suit :
— $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$
— $S = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(m)}, y^{(m)})\} \subset \mathcal{X} \times \mathcal{Y}$

Notons par ailleurs :

$$X = \begin{bmatrix} 1 & \cdots & 1 \\ x_0^{(1)} & \cdots & x_0^{(m)} \\ | & | & | \\ x_j^{(1)} & \cdots & x_j^{(m)} \\ | & | & | \end{bmatrix} \in \mathbb{R}^{d+1 \times m} \qquad Y = [y^{(1)} \cdots y^{(m)}] \in \mathbb{R}^{1 \times m}$$

$$W = \begin{bmatrix} | \\ w_i \\ | \end{bmatrix} \in \mathbb{R}^{d+1 \times 1}$$

Proposition
Quelques résultats de calcul différentiel matriciel utile dans la suite :
Soit $W \in \mathbb{R}^{n \times 1}$ et $A \in \mathbb{R}^{n \times n}$ alors :
— $\frac{d}{dW} AW = A$
— $\frac{d}{dW} W^T A = A$
— A est symmétrique $\implies \frac{d}{dW} W^T AW = 2AW$
Exercice : prouver ces propositions.

3.1 Régression linéaire

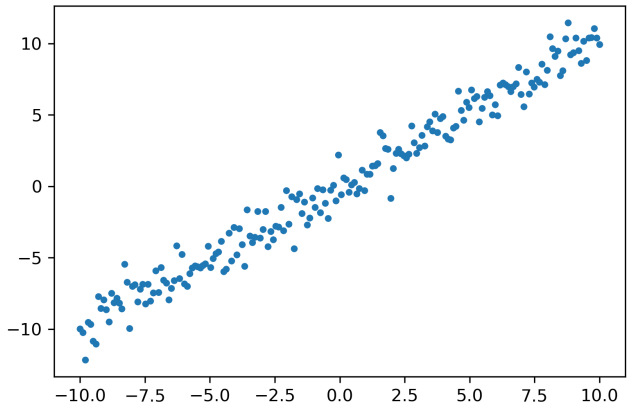
Lorsqu'on cherche à déterminer la variable à expliquer $Y \in \mathcal{Y}$ en fonction des régresseurs $X \in \mathcal{X}$ (chaque dimension constitue une variable explicative indépendante i.e X est de rang $d+1$), l'hypothèse la plus simple à effectuer est l'hypothèse de linéarité :

$$Y = W^T X + \varepsilon$$

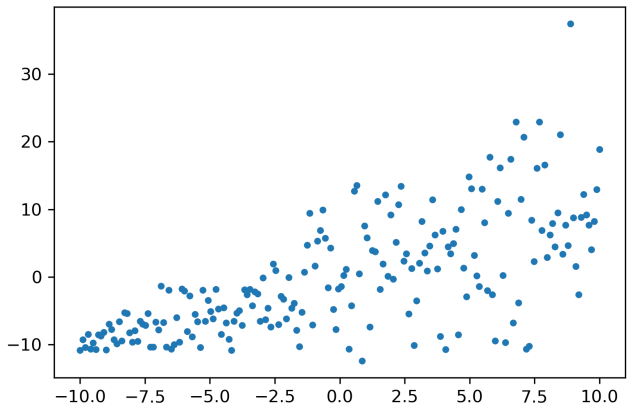
Ou les $\varepsilon \in \mathbb{R}^m$ est le vecteur aléatoire correspondants aux erreurs effectuées sur chaque donnée. Par ailleurs l'hypothèse d'homoscédasticité stipule que les erreurs associées sur chaque éléments sont identiquement distribuées $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ Ainsi on suppose que $\varepsilon = Y - W^T X \sim \mathcal{N}(0, \sigma^2)$. En d'autres termes ε est une estimation de l'erreur commise pour une valeur de W. Pour trouver la meilleure le modèle estimant au mieux les données on cherchera donc à minimiser cette erreur. On procède alors en minimisant la norme de l'erreur $||\varepsilon||^2 = \varepsilon\varepsilon^T$. En d'autres termes on cherche :

$$\hat{W} = \arg\min_W \varepsilon\varepsilon^T$$

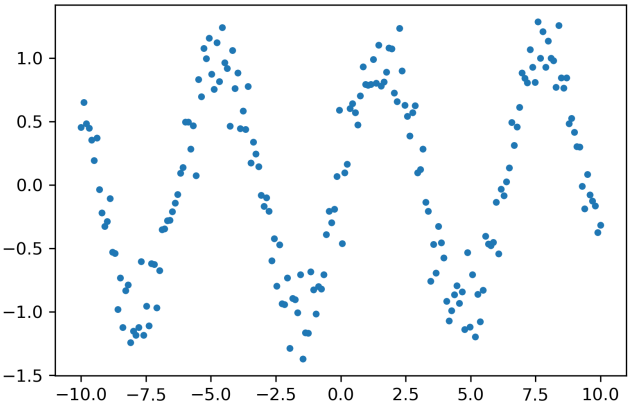
Pour cela on peut utiliser la méthode de l'équation normale et trouver une solution analytique au problème, ou utiliser l'algorithme du gradient.



(a) Homoscédasticité



(b) Hétéroscédasticité



(c) Non linéarité

FIGURE 3.1: Exemple de données représentants les différentes hypothèses que nous pouvons faire sur les données.

3.1.1 Résolution à l'aide de l'équation normale

$$
\begin{aligned}
||Y - W^T X||^2 &= \varepsilon\varepsilon^T && (3.1)\\
&= (Y - W^T X)(Y - W^T X)^T && (3.2)\\
&= (Y - W^T X)(Y^T - X^T W) && (3.3)\\
&= (YY^T - YX^T W - W^T XY^T + W^T XX^T W) && (3.4)\\
&= (YY^T - 2W^T XY^T + W^T XX^T W) && (3.5)
\end{aligned}
$$

Exercice : Vérifier que :
— $(YX^T W)^T = W^T XY^T$
— $YX^T W$ et $W^T XY^T$ sont des matrices de dimension 1×1.

Nous souhaitons minimiser le risque quadratique, pour cela on peut faire varier le paramètre W, ainsi le minimum est atteint lorsque $\nabla_W[\varepsilon\varepsilon^T] = 0$, pour cela on

$$
\begin{aligned}
\nabla_W[\varepsilon\varepsilon^T] &= \nabla(\frac{1}{n}YY^T - 2W^T XY^T + W^T XX^T W) && (3.6)\\
&= -2XY^T + 2XX^T W && (3.7)
\end{aligned}
$$

Ainsi on trouve la valeur de W minimisant l'erreur quadratique.

$$
\begin{aligned}
\nabla_W[\varepsilon\varepsilon^T] = 0 &\iff XY^T = XX^T W && (3.8)\\
&\iff (XX^T)^{-1}XY^T = W && (3.9)
\end{aligned}
$$

Cette dernière expression s'appelle l'équation normale pour la régression linéaire. Ainsi si on note $\hat{W} = (XX^T)^{-1}XY^T$, alors l'expression $\hat{W}^T X$ aura l'erreur quadratique la plus faible pour prédire les données Y.

Remarque : On aurait pu choisir d'additionner les matrices de dimension 1×1 dans l'autre sens. On a choisi ce sens pour avoir l'expression de gauche dans l'équation normale de même dimension que W. Par ailleurs XX^T est une matrice symétrique définie positive, ce qui explique son invertibilité. Cependant lorsque certaines lignes sont presque corrélée des problèmes numériques peuvent survenir. On appelle ce probleme le problème de multicolinearité. C'est pour cela qu'il faut s'assurer que les variables sont lineairement independantes. Pour tester cela on peut calculer le determinant de XX^T et verifier s'il est proche de 0. Si c'est le cas il faut procéder a une sélection de variables. Une autre technique est de regulariser les poids en introduisant une contrainte, par exemple exiger que $||W|| < a$ ou a est un entier a déterminer. Cela change la donne du probleme d'optimisation, en effet, en introduisant une contrainte la methode change. Dans ce cas particulier, on peut utiliser la methode des multiplicateurs de Lagrange, ou plus généralement la méthode de Karush-Kuhn-Tucker.

3.1.2 Résolution à l'aide de la descente du gradient

Bien que nous ayons défini les équations normales pour trouver les paramètres expliquant au mieux les données, cette approche ne convient pas lorsque les données sont très importante (> 100000 exemples). Dans ce cas une alternative revient à utiliser la méthode de la descente du gradient. Une autre écriture de l'estimateur de l'erreur est la suivante :

$$\varepsilon\varepsilon^T = \sum_{i=1}^{n}(y^{(i)} - W^T x^{(i)})^2 \tag{3.10}$$

$$= \sum_{i=1}^{n}\sum_{j=0}^{d}(y^{(i)} - w_j x_j^{(i)})^2 \tag{3.11}$$

En interprétant que cette estimateur est une fonction de W on peut alors résoudre l'équation $\nabla_W[\varepsilon\varepsilon^T]$ avec la méthode du gradient (décrite plus bas).

3.2 Régression linéaire pondérée

On conserve l'hypothèse de linéarité, à savoir $Y = W^T X + \varepsilon$. Cependant lorsque le bruit dans les données n'est pas distribué de manière homogène on dit que l'on suppose l'hypothèse d'hétéroscédasticité. Les erreurs associé à chaque variable suit sa propre loi $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. La régression pondéré consiste alors à proposer l'estimateur suivant : $\varepsilon A \varepsilon^T$ ou A est une matrice diagonale composée des valeurs $\frac{1}{\sigma_i^2}$.

$$\varepsilon A \varepsilon^T = (Y - W^T X)A(Y - W^T X)^T \tag{3.12}$$

$$= (Y - W^T X)(AY^T - AX^T W) \tag{3.13}$$

$$= YAY^T - YAX^T W - W^T XAY^T + W^T XAX^T W \tag{3.14}$$

$$= YAY^T - 2W^T XAY^T + W^T XAX^T W \tag{3.15}$$

Comme précedemment pour minimiser le risque quadratique, on résoud $\nabla_W[\varepsilon A \varepsilon^T] = 0$ pour obtenir l'équation normale pour la régression linéaire pondérée.

$$\nabla_W[\varepsilon A \varepsilon^T] = -2XAY^T + 2XAX^T W \tag{3.16}$$

$$\nabla_W[\varepsilon A \varepsilon^T] = 0 \implies XAY^T = XAX^T W \tag{3.17}$$

$$\implies (XAX^T)^{-1} XAY^T = W \tag{3.18}$$

3.2.1 Résolution à l'aide de la descente du gradient

De la même manière que précédemment lorsque les données sont importante, il vaut mieux utiliser une méthode itérative pour trouver le minimum de l'estimateur $[\varepsilon A \varepsilon^T]$ En utilisant par exemple la méthode de la descente du gradient pour minimiser l'expression suivante :

$$[\varepsilon A \varepsilon^T] = \sum_{i=1}^{n} a_i (y^{(i)} - W^T x^{(i)})^2$$

Où les a_i sont les termes de la matrice A supposée connue.

3.3 Optimisation : méthode du gradient

La méthode du gradient ou de la plus forte pente est un algorithme d'optimisation différentiable utilisée très fréquemment en apprentissage machine. En général on cherche à minimiser une fonction à entrée multiple $f : \mathbb{R}^n \to \mathbb{R}$, pour cela on utilise les dérivées partielles $\frac{\partial f}{\partial x}$ qui mesurent le taux de variation de f dans les n directions. Le gradient de f en x s'exprimant $\nabla f(x) = (\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n})$ est le vecteur des taux de variations de la fonction f dans chacune des n directions.
En calcul différentiel les points critiques d'une fonction sont ceux annulant le gradient, ainsi on cherche l'ensemble suivant :

$$\{x \in \mathbb{R}^n : \|\nabla f(x)\| = 0\} = \{x \in \mathbb{R}^n : \frac{\partial f(x)}{\partial xi} = 0, i = 1, \cdots, n\}$$

Dans le cas ou la fonction à minimiser est convexe, il n'y a pas de points selles et le minimum est global. Alors minimiser f revient à chercher la direction dans laquelle f décroit le plus rapidement. La méthode de la plus forte descente est une méthode itérative consistant à choisir une valeur initiale et à prendre la direction opposée au gradient. On peut le résumer par l'algorithme suivant :

Algorithme 1 : Méthode de la plus forte pente

Data : $f : \mathbb{R}^n \to \mathbb{R}, x \in \mathbb{R}^n, \lambda, \alpha \in \mathbb{R}$;
initialization x = 0;
while $\|\nabla f(x)\| > \lambda$ **do**
  $x \coloneqq x - \alpha \nabla f(x)$;

Ici λ contrôle la précision attendue, et α est le coefficient d'apprentissage, c'est la taille du pas que l'on fait vers le minimum, en apprentissage automatique ce type de coefficient est appelé un hyper-paramètre défini par le modélisateur.

La méthode que nous venons de décrire permet de résoudre de manière itérative des problèmes d'optimisation convexe sans se poser de question. Dans le cas ou la fonction à minimiser n'est pas convexe, il n'y a pas de garantie de trouver un minimum global, à ce moment la démarche devient plus complexe, et d'autres techniques entrent en jeu : optimisation sous contrainte, méthodes d'optimisation du second ordre, etc.

3.4 Régression non linéaire

Dans le cas de la régression non linéaire on suppose que $Y = f(W^T X) + \varepsilon$ ou f est supposée différentiable. On va directement proposer l'algorithme de la descente du gradient pour trouver la valeur de W permettant au modèle d'expliquer au mieux les données. Notons $\mathcal{L}$ une fonction coût à minimiser

$$\mathcal{L}(W) = \mathcal{L}(w_0, \cdots, w_d) = [\varepsilon \varepsilon^T] \tag{3.19}$$

$$= (Y - f(W^T X))(Y - f(W^T X))^T \tag{3.20}$$

$$= \sum_{i=1}^{n} \sum_{j=0}^{d} (y^{(i)} - f(w_j x^{(i)}))^2 \tag{3.21}$$

Alors l'algorithme de la descente du gradient s'écrit

Algorithme 2 : Écriture synthétique

Data : $\mathcal{L} : \mathbb{R}^{d+1} \to \mathbb{R}, W \in \mathbb{R}^{d+1}, \lambda, \alpha \in \mathbb{R}$;
`// On initialize W arbitrairement`
initialization W = 0;
while $\|\nabla \mathcal{L}\| > \lambda$ **do**
  $W \coloneqq W - \alpha \nabla \mathcal{L}$;

Exercice :

— Calculer $\frac{\partial \mathcal{L}}{\partial w_k}$.

— Donner l'équation de mise à jour des paramètres w_k.

— Répéter les deux questions précédentes pour le cas linéaire et le cas linéaire pondéré (optionnel) .

On utilise la régression non linéaire notamment dans le cadre de la classification binaire et multi-classe. Par exemple classer des images de chien et chat. Dans ce cas précis on utilise une fonction non-linéaire appelé sigmoïde. En somme la fonction non linéaire que vous utilisez dépend du problème en main.

Remarque : La descente du gradient n'est pas la seule méthode d'optimisation pour trouver les bon paramètres. On peut citer la méthode de Gauss-Newton et de Levenberg-Marquardt.

Exercice : *Vraisemblance et régression linéaire*
On suppose que $y|x \sim \mathcal{N}(W^T x, \sigma^2)$. On suppose que les données sont conditionnellement indépendantes, i.e $\mathbb{P}(y^{(i)}, y^{(j)}|x^{(i)}, x^{(j)}) = \mathbb{P}(y^{(i)}|x^{(i)})\mathbb{P}(y^{(j)}|x^{(j)})$ Montrer que

$$\arg\max_W \mathcal{L}(y^{(1)}, \cdots, y^{(n)}|x^{(1)}, \cdots, x^{(n)}, W) = \arg\min_W \frac{1}{n}||Y - W^T X||_2^2$$

Où $\mathcal{L}(y^{(1)}, \cdots, y^{(n)}|x^{(1)}, \cdots, x^{(n)})$ est la vraisemblance de la loi conditionnelle $y^{(1)}, \cdots, y^{(n)}|x^{(1)}, \cdots, x^{(n)}$, et $||.||_2$ la norme euclidienne.

Chapitre Quatre

Estimateurs Bayésiens

Sommaire

4.0.1 Rappels

Définition : *loi jointe, loi marginales.*
Si X, Y sont deux v.a, on dit que $\mathbb{P}(X, Y)$ est la loi jointe des v.a, on appelle loi marginale les lois individuelles $\mathbb{P}(X), \mathbb{P}(Y)$

Définition : *probabilité conditionnelle entre v.a.*
Soit X, Y deux v.a, on appelle probabilité conditionnelle de Y sachant X et on note $\mathbb{P}(Y|X)$ la quantité suivante

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

Définition : *marginalisation.*
Pour trouver la loi marginale X connaissant la loi jointe on marginalise suivant Y

$$\mathbb{P}(X) = \int_{-\infty}^{\infty} \mathbb{P}(X, Y) dY$$

4.1 Cadre Bayésien

Pour bien comprendre la différence de point de vue entre l'approche fréquentiste et l'approche Bayésienne de la modélisation statistique, rappelons l'interprétation fréquentiste. Un modèle paramétrique est la donne d'un triplet $(\Omega, \mathcal{F}, \mathbb{P}_\Theta)$. Dans le contexte fréquentiste, on suppose que le paramètre expliquant au mieux ldes données est fixe et inconnu. On le détermine avec le maximum de vraisemblance. Dans le cadre Bayésien, on suppose que le paramètre expliquant les données est aléatoire suivant une loi inconnue. Le théorème de

Bayes introduit la formule donnant la valeur de la loi du paramètre à rechercher.

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\mathbb{P}(X)}$$

On appelle
- $\mathbb{P}(\theta)$ l'à priori
- $\mathbb{P}(\theta|X)$ l'à posteriori
- $\mathbb{P}(X)$ l'évidence
- $\mathbb{P}(X|\theta)$ la vraisemblance comme vu précedemment.

Notez que la difficulté est dans l'évaluation de l'évidence $\mathbb{P}(X)$. En effet, on peut réecrire la loi de Bayes ainsi :

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\mathbb{P}(\theta)}{\int_\Theta \mathbb{P}(X,\theta)d\theta}$$

Cependant il existe une approche pour s'affranchir de calculer l'évidence, c'est l'approximation du maximum a posteriori

4.2 Estimation du maximum a posteriori (MAP)

Définition : *proportionalité.*
On dit que l'a posteriori est proportionel à la vraisemblance multiplié par l'a priori et on note ainsi $\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)\mathbb{P}(\theta)$
L'estimateur du maximum a posteriori part de la remarque suivante :

$$\arg\max_\theta \mathbb{P}(\theta|X) = \arg\max_\theta \mathbb{P}(X|\theta)\mathbb{P}(\theta) \tag{4.1}$$

Exercice : *MAP et régression linéaire.*
Soit $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ des données, on suppose que $y|x, w \sim \mathcal{N}(wx, \sigma^2)$ et que $w \sim \mathcal{N}(0, \gamma^2)$. Déterminer le maximum a posteriori de la régression linéaire.

Comme observé dans l'exercice, le maximum à posteriori ne nécessite que de supposer un a priori sur le modèle. Cependant, en opérant ainsi on n'est pas en mesure de connaître la loi a posteriori, c'est un inférence Bayésienne incomplète. Ne connaissant pas la loi, on perd la capacité à estimer l'incertitude de l'approche Bayésienne traditionnelle. Cependant l'inférence est quand même meilleure que la simple maximisation de la vraisemblance car on obtient un terme régularisateur. Ce qui permet d'éviter certain problèmes numériques comme discuté dans la remarque en section 3.1.1.

4.3 Loi conjuguées

Pour obtenir l'inférence Bayésienne complète il existe au moins 3 stratégies, utiliser des lois conjuguées, utiliser la méthode de Monte Carlo par les chaînes de Markov, ou enfin utiliser l'inférence variationnelle. Dans la suite de ce chapitre, on se concentrera sur les lois conjuguées en donnant plusieurs exemples et exercices.

Définition : *statistique exhaustive.*
On dit qu'une statistique T est exhaustive pour un échantillon $X_1, \cdots, X_n$ si $\mathbb{P}(\theta|T(X_1, \cdots, X_n)) = \mathbb{P}(X_1, \cdots, X_n)$.
En d'autre terme toute l'information contenu dans l'échantillon sur le paramètre θ est contenu dans la statistique appliqué à l'échantillon. Autrement dit, on ne perd pas d'information sur θ en considérant $T(X_1, \cdots, X_n)$ à la place de $X_1, \cdots, X_n$.

Définition : *conjugaison.*
On dit que l'a priori et la vraisemblance sont conjugués si l'a priori et l'a posteriori sont dans la même famille de lois de probabilités.

Quelques exemples de familles de loi de probabilité incluent : la famille exponentielle, les mixtures à densité et les loi variationnelles à champ moyen (mean field variational families). Dans la suite de ce cours, nous nous concentrerons uniquement sur la famille exponentielle, car elle englobe la plupart des lois usuelles.

Définition : *famille exponentielle.*
Une loi de probabilité de la famille exponentielle peut se mettre sous la forme :

$$\mathbb{P}(x|\eta) = h(x)e^{(\eta^T t(x)+a(\eta))}$$

ou

- $h(x)$ est la mesure sous jacente pour s'assurer que x est dans le bon espace
- η est le paramètre naturel
- $t(x)$ est la statistique exhaustive
- $a(\eta)$ est le log-normalisateur.

Parmi les membres de la famille exponentielle, on trouve la loi normale, exponentielle, gamma, chi-carré, bêta, Dirichlet, Bernoulli, Bernoulli multinomiale, Poisson, Wishart, Wishart inverse, etc. C'est une manière standardisé de représenter ces lois.

Exemple(s) : Exprimons la loi Gaussienne univariée dans le famille exponentielle.

$$\mathbb{P}(x|\mu, \sigma) = \mathcal{N}(x|\mu, \sigma) \tag{4.2}$$

$$= \frac{e^{\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{4.3}$$

$$= \frac{e^{\frac{-x^2+2x\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma}}{\sqrt{2\pi}} \tag{4.4}$$

Alors en posant

- $h(x) = \frac{1}{\sqrt{2\pi}}$
- $t(x) = (x^2, x)$
- $\eta = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$
- $a(\eta) = \frac{\mu^2}{2\sigma^2} - \log\sigma$

On a une paramétrisation de la loi Gaussienne univariée.

Exercice :
Exprimer la loi de Bernoulli $\mathcal{B}(p)$ de paramètre p sous forme exponentielle.

4.3.1 Exemples