Cours: MAPSI

Charles Vin

2022

Nouveau cours du 13/09

1 Introduction

— Exam final : 50%
— Partiel : 35%
— Participation : 15%
  — travail dans la séance
  — TME soumis en fin de séance omg

Deux grand type de modèle :
— Modèle paramétrique : connaissance sur la distribution stat des données. Puis on estime les paramètres de la loi.
— Modèle non paramétrique : l'inverse, on ne connait pas la loi. exemple : regression logistique

Echantillons :
— population
— ect

Définition 1.1. Vocabulaire :
— Voir diapo 9/51

Définition 1.2 (Mesure de proba)**.** Une fonction qui associe chaque événement à une valeur entre 0 et 1. Voir diapo 15, definition importante.

Définition 1.3.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Densité de proba

Retrouver la définition.

Fonction de répartition

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(x)dx.$$

Espérance :

$$E(X) = \sum x_k * p_k$$

$$E(X) = \int Xp(x)dx$$

$$E(aX + b) = aE(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$

Le Mode

$$p(Mo) = \max_k p(x_k) p(Mo) = \max_x p(x)$$

Variance :

$$\sigma^2 = \sum (x_k - E(X))^2$$

$$\sigma^2 = \int (x - E(X))^2 p(x) dx$$

$$V(aX + b) = a^2 V(X)$$

$$V(X) = E(X^2) - E(X)$$

Médiane et quantile

idk diapo

Définition 1.4 (Loi marginale)**.** La marginalisation consiste à projeter une loi jointe sur l'une des variables aléatoires. Par exemple extraire $P(A)$ à partir de $P(A, B)$.

$$P(A) = \sum_i P(A, B = pb_i).$$

C'est la somme de la ligne ou de la colonne du tableau.

Définition 1.5. Probabilités conditionnelles

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

$$\Leftrightarrow P(A \cup B) = P(A|B)P(B)$$

Proposition 1.1.
— *Réversibilité :* $P(A, B) = P(A|B)$
— *Théorème de Bayes :*

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

— *Intégration des probabilités totale*
— *DIAPO 39*

Définition 1.6 (Indépendance probabiliste)**.** Deux événements A et B sont indépendants si

$$P(A, B) = P(A) * P(B).$$

Corollaire : $P(A|B) = P(A)$

Définition 1.7. La covariance

$$cov(X, Y) = E[(X - E(X))(Y - E(Y))].$$

Définition 1.8 (Coefficient de corrélation linéaire)**.** Soit X,Y deux variables. Le coefficient de corrélation linéaire entre X et Y est :

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

CCL : VOIR DIAPO
— Probabilité
— Marginalisation
— Conditionnement
— Indépendance : Si X_1 et X_2 sont indépendantes : $P(X_1, X_2) = P(X_1)P(X_)$
Nouveau cours du 20/09

Définition 1.9 (Indépendance de deux variables discrète)**.** Discrète : Continue :

Définition 1.10 (Indépendance mutuelle de n variable)**.** Soient n variables aléatoires $(X_1, ..., X_n)$. Elle sont **mutuellement indépendantes** si tout événement lié à une partie d'entre elles est indépendant de tout événement lié à toute autre partie disjointe de la précédente. Propriété :

— Indépendance mutuelle → Indépendance deux à deux. **Attention :** réciproque fausse
→ Permet de réduire la taille du tableau des probabilité de chaque événement !

Définition 1.11 (Indépendance conditionnelles)**.** On reprend les formules de l'indépendances mais en sachant une variable, au final c'est dans un cas particulier.

$$X \perp Y|Z$$
$$\forall x, \forall y, \forall z P(X = x \cap Y = y|Z = z) = P(X = x|Z = z) * P(Y = y|Z = z)$$

→

$$\Rightarrow P(X, Y|Z) = P(X|Z) * P(Y|Z).$$

Définition 1.12. Loi normale

Proposition 1.2. *- Moyenne linéaire et variance comme bilinéaire*

$$X \sim \mathcal{N}(\mu, \sigma^2) \text{ alors } Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

- Centrer et réduire

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Définition 1.13 (Convergence en loi)**.**

$$\forall x, \lim_{n\to\infty} F_n(x) = F(x).$$

On ne sais pas comment ça converge

Définition 1.14 (Convergence en probabilité)**.** (X_n) **converge en probabilité** vers X si, pour tout $\epsilon > 0$ la probabilité que l'écart absolu entre X_n et X dépasse ϵ tend vers 0 quand $n \to \infty$

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

Définition 1.15 (convergence presque sur)**.** (X_n) **converge presque surement**vers X s'il y a yne proba 1 que la suite des réalisation des X_n tende vers X

Définition 1.16 (Loi faible des grands nombre)**.** Soit $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires :
— De même loi
— D'éspérance m
— Possédant une variance σ^2
— **Deux à deux** indépendante

Alors

$$\bar{X_n} = \frac{\sum_{k=1}^{n} X_k}{n} \to_{\mathbb{P}} m.$$

Rappel :

$$E(\bar{X_n}) = m$$
$$V(\bar{X_n}) = \frac{\sigma^2}{n}$$

Définition 1.17 (Loi forte des grands nombres)**.** Soit $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires :
— De même loi
— D'éspérance m
— Possédant une variance σ^2
— **mutuellement** indépendante

Alors

$$\bar{X_n} = \frac{\sum_{k=1}^{n} X_k}{n} \to_{p.s} m.$$

Définition 1.18 (Théorème centrale limite)**.** Soit $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires :
— De même loi
— D'éspérance μ
— Possédant une variance σ^2
— **mutuellement** indépendantes

Alors

$$\frac{\bar{X_n} - \mu}{\sigma/\sqrt{n}} \to_{loi} \mathcal{N}(0, 1).$$

Nouveau cours du 27/09

2 Maximum Vraissemblance

Définition 2.1 (Vraissemblance d'un échantillon)**.** Soit $x = (x_1, \ldots, x_n)$ réalisation de $(X_1, \ldots, X_n)$ **iid = Mutuellement indépendant** Alors on défini la vraisemblance dans le cas discret comem étant la proba d'obtenir **cet** échantillon sachant la loi P

$$L(x) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i).$$

Dans le cas continue :

Exemple 2.1 (avec des pièces de monnaies)**.** DIAPO 5

Exemple 2.2 (innondation)**.** 3 type de parcelles : Inondables (PI), partiellement inondables (PPI), non inondable (NI) On a deux loi caractérisant le niveau de gris par rapport à la catégorie d'inondation.

$$P(n|PI) = \mathcal{N}(\mu_1, \sigma_1^2), P(n|PPI) = \mathcal{N}(\mu_2, \sigma_2^2).$$

Avec n le niveau de gris.
Soit une image Z avec un niveau de gris $n = 80$; Deux hypothèses
— θ_1 Z PI
— θ_2 Z PPI
On vas calculer le max de vraisemblance d'obtenir la zone Z sous $\theta 1$ ou $\theta 2$

2.1 Maximum de vraissemblance

Exemple 2.3 (Pièce de monnaie)**.** On vas faire la même chose mais cette fois ci, on prend des paramètre $\theta_1 et \theta_2$

Définition 2.2 (Vraissemblance d'un échantillon)**.** On cherche à estimer un paramètre Θ Soit $x = (x_1, \ldots, x_n)$ réalisation de $(X_1, \ldots, X_n)$ **iid = Mutuellement indépendant** Alors on défini la vraisemblance dans le cas discret comme étant la proba d'obtenir **cet** échantillon sachant la loi P

$$L(x) = P(X_1 = x_1, \ldots, X_n = x_n | \Theta = \theta) = \prod_{i=1}^{n} P(X_i = x_i | \Theta = \theta).$$

On peut utiliser la fonction de densité.

Définition 2.3 (Maximum vraissemblance)**.** On cherche le maximum de la fonction $L(x, \theta), \forall \theta$. Donc on vas la dériver ! et utiliser le $\log$

Exemple 2.4. Plein d'exemple dans le diapo

Exemple 2.5 (problème d'ajustement)**.** On a des points un peu random, réparti comme un sinusoide qu'on vas approximé par un polynome. Problème : on a une erreur Normale.
il vas vite. Mais ça resemble à une regression.

2.2 Estimation par maximum a posteriori

Exemple 2.6 (Pièce de monnaie)**.** Imaginons qu'on a un tirage de 3 piles. Le maximum de vraisemblance vaut 1. Mais ça va à l'encontre du bon sens. $\rightarrow$ Solution : Maximum a posteriori. (A voir pourquoi ça fait 1)

Définition 2.4 (Maximum a posteriori)**.** On se base dans un modèle bayésien avec
— $\mathcal{X}$ = l'espace des observations x de taille n
— Θ

Formile de la vraissemblance diapo 32.

Estimateur du maximum a posteriori toujours égal à l'argmax de la vraissemblance

$$x \mapsto t = Argmax_{\theta \in \Theta} \pi(\theta|x).$$

Exemple 2.7 (pièce de monnaie)**.** En faite la grosse différence c'est la dernière ligne du diapo 34. On pose, on invente l'information a priori de la proba de chaque paramètre qu'on vas tester. Cette information vas permettre d'être utiliser dans le modèle bayésien. On l'a choisi en fonction d'une loi normale. Fin du cours sans qu'il ait fini.

Nouveau cours du 04/10

3 Principes d'apprentissage avec donnée manquantes

- Solution ez : Supprimer les lignes avec des données manquantes. Problème, ça peut changer les probabilités qu'on peut apprendre.
- Remplacer les valeurs manquante par les plus probable : C'est équivalent à l'algo des K-Means. Mais pareil ça change les probas. Mais la valeur la plus probable peut quand même valoir 15% si notre variables aléatoire peut prendre beaucoup de valeur.
- Replacer les valeurs manquante par toutes les valeurs possible : marche pas non plus.
- Tenir compte de la distribution des valeur : mais ça change encore X
- **Solution algo EM :**Replacer les valeurs manquante par toutes les valeurs possible ponderers par leur probabilité d'apparition. Cette fois-ci les proba sont équivalente entre les deux tableaux.

Idée de K-MEans et EM :

- Se donner un modèle initial (pas trop mauvais)
- Ce modèle $\rightarrow$ Donne des données complètés
- Apprendre un nouveau modèle avec ces données
- Boucle

Y'a-t-il convergence ?

3.1 L'algorithme EM

Notation :

- x^o données observées, x^h données manquantes, $x = x^o \cup x^h$
- $M_{ij} = P(r_i^j \in x^h)$ proba de position des données manquantes

Plusieurs cas sur les proba de missing data :

- Missing Completely at Random (MCAR) : $P(M|x) = P(M)$ Aucune relation entre le fait qu'une donnée soit manquante ou observée
- Missing at Random (MAR) : $P(M|x) = P(M|xo)$ données manquantes en relation avec les données observées mais pas avec les autres données manquantes
- Not Missing At Random (NMAR) : $P(M|x)$ données manquantes en relation avec toutes les données

On vas regarder que **MCAR**.

On calcule une log vraissemblance sur les données observées

$$\log L(x^o, \Theta) = \sum_{i=1}^{n} \log P(x_i^o|\Theta) = \sum_{i=1}^{n} \log(\sum_{x_i^h \in x^h} P(x_i^o, x_i^h|\Theta))$$

On fait apparaitre les x^h avec la formule de somme loi marginale. Rappel : on leur a donnée des probas manuellement.

Soit $Q_i(x_i^h)$ une loi de proba **quelconque** alors on peut l'insérer dans l'équation pour ensuite utiliser l'équation de Jensen des fonction convexe/concave. Comme ça on vas pouvoir sortir la somme du log.

Bref, pour plus de détail sur les math voir le diapo, finalement on arrive diapo 16.

Algo EM :

1. Choisir valeur initale
2. **Pour chaque ligne**, je fais le calcul diapo 17
3. Maximisation
4. Boucle tant que pas de convergence

En faite on converge vers un optimum **local** qui dépend du point initial. Donc du coup on en test plein

Exemple 3.1. DIAPO 18

1. On est pas obligé de faire cette méthode d'initialisation. C'est juste une sorte d'indication mais au final on défini encore en random après
2.

Pourquoi ça converge ? Grace à l'inégalité de Jensen et la concavité du log. Voir diapo 26 et 27.

4 Mixure de gaussiennes

Les données suivent plusieurs gaussienne différente ?

Exemple 4.1 (application : apprentissage de prix fonciers)**.** On a un échantillon de prix de logement avec leur caractéristique et le quartier. Le prix est sur une gaussienne de paramètre variant en fonction du quartier. (Je suis pas sur de si on a l'info sur le quartier, je crois on cherche à la prédire avec de l'apprentissage non-supervisé)

On peut tenter de calculer directement ça je crois mais bref ça marche pas, pas de solution analytiquement. Solution $\rightarrow$ EM

On créer une colonne vide, pleine de valeur manquante pour le quartier et on vas faire EM dessus. C'est assez drôle on créer des données à partir de rien.

Exemple 4.2 (Classification d'image)**.** Deux dernière diapo

Nouveau cours du 11/10

5 Tests d'hypothèses

tout comme l'année dernière
Nouveau cours du 18/10

6 Chaine de Markov

Cette fois-ci on vas pas utiliser des données sous forme matriciel iid. Ici on vas s'intéresser à des modèles de séquences, qui ont une dynamique temporelle. Application :

— Musique/reconnaissance de paroles
— Reconnaissance de mouvement
— Diffusion dans les graphes

Problème : Les méthodes standard de classification = données de tailles fixes $\rightarrow$ transition difficile vers des données de taille variable

Diapo 6/45

1. $\Leftrightarrow$ vraissemblance
2. .
3. proba à posteriori

Faire de l'apprentissage d'un modèle de séquence $\Leftrightarrow$ apprendre une fonction de densité. On support toujours les θ_k iid (HP forte).

Diapo 9 :

Au final on modèlise la dépendance par la chaine de markov. On vas essayer d'associer à une séquence, un label pour prédire la classe. Paramètre du modèle $\{\Pi, A\}$. Permet de faire des prévision dans les espaces discrets

Diapo 10 :

Hypothèse Markovienne : La proba de l'état suivant ne dépend que de k état d'avant. On prend en général $k = 1$: l'état prochain ne dépend uniquement de l'était présent. Si on prend plus ça rajoute beaucoup beaucoup de paramètre.

Diapo 11 : Définition des paramètres
On a une matrice de transition $A = [a_{ij} = p(x_{t+1} = q_j | x_t = q_i)]$ avec la somme des ligne égal à 1 (matrice stochastique). Et $\Pi =$ proba d'état initiale

Diapo 14 : Représentation matriciel
On peut avoir la proba $p(x_{t+1} = q_j)$ en un calcul matriciel

$$p_{t+1} = p_t * A.$$

Diapo 21 : Stationarité

Définition 6.1. Existe-t-il un état stationnaire μ

$$\mu = \mu A.$$

C'est à dire qu'on

— Si A irreductible $\rightarrow \mu$ est unique

Définition 6.2 (Irréductible finie)**.** Si on part d'un état donnée, la probabilité d'y revenir est n on nulle. en un nombre d'étape fini $\Leftrightarrow$ graphe fortement connexe, pas d'état final/absorbant.

Définition 6.3 (Périodicité)**.** Etat périodique de période k si on peut y revenir en un nombre d'étape multiple de k.

Période d'une chaine de markov = PGCD de la période de tout ces états. Diapo 23

Théorème 6.1 (Ergodique)**.** *C'est la loi forte des grands nombres pour les chaines de markov. On a convergence vers la moyenne.*

Les chaines irréductible et apériodique sont ***ergodique***

7 Apprentissage des paramètres

Comment apprendre une CM à partir d'exemple? Comment faire de la classification de séquences avec des CM?

On vas maximiser la vraissemblance, mais cette fois on a des contraintes sur les distributions de proba égal à 1 (comme avec la multinomiale). $\rightarrow$ Langrandien

Il s'est arrêté diapo 34. et rush le reste. Au final c'est comme d'habitude la moyenne d'apparition