

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 connaît 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, définition 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 :

$$\sigma^2 = \sum (x_k - E(X))^2$$
$$\sigma^2 = \int (x - E(X))^2 p(x) dx$$
$$V(aX + b) = a^2 V(X)$$
$$V(X) = E(X^2) - E(X)^2$$

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

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

Proposition 1.1. — Réversibilité : $P(A, B) = P(A|B)P(B)$

— 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

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

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

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

CCL : VOIR DIAPO

— Probabilité

— Marginalisation

— Conditionnement

— Indépendance : Si X_1 et X_2 sont indépendantes : $P(X_1, X_2) = P(X_1)P(X_2)$

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, \dots, 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 \rightarrow Indépendance deux à deux. **Attention** : réciproque fausse
- \rightarrow Permet de réduire la taille du tableau des probabilités de chaque événement !

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

$$X \perp Y | Z$$

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

$$\rightarrow$$

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

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).

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

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 $n \rightarrow \infty$

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

Définition 1.15 (convergence presque sur). (X_n) **converge presque sûrement** vers X s'il y a une proba 1 que la suite des réalisations 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'espérance m
- Possédant une variance σ^2
- **Deux à deux** indépendante

Alors

$$\bar{X}_n = \frac{\sum_{k=1}^n X_k}{n} \rightarrow_{\mathbb{P}} m.$$

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'espérance m
- Possédant une variance σ^2
- **mutuellement** indépendante

Alors

$$\bar{X}_n = \frac{\sum_{k=1}^n X_k}{n} \rightarrow_{p.s} m.$$

Définition 1.18 (Théorème centrale limite). Soit $(X_n)_{n \in \mathbb{N}}$ est une suite de variables aléatoires :

- De même loi
- D'espérance μ
- Possédant une variance σ^2
- **mutuellement** indépendantes

Alors

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

Nouveau cours du 27/09

2 Maximum Vraisemblance

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

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

Dans le cas continu :

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

Exemple 2.2 (inondation). 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 va calculer le max de vraisemblance d'obtenir la zone Z sous θ_1 ou θ_2

2.1 Maximum de vraisemblance

Exemple 2.3 (Pièce de monnaie). On va faire la même chose mais cette fois ci, on prend des paramètres θ_1 et θ_2

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

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

On peut utiliser la fonction de densité.

Définition 2.3 (Maximum vraisemblance). On cherche le maximum de la fonction $L(x, \theta), \forall \theta$. Donc on va 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 sinusoides qu'on va approximer par un polynome. Problème : on a une erreur Normale.

il va vite. Mais ça ressemble 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. → 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

Formule de la vraisemblance diapo 32.

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

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

Exemple 2.7 (pièce de monnaie). En fait 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 → Donne des données complétés
- 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) : $P(M|x) = P(M|x^o)$ 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 vraisemblance sur les données observées

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

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

En fait 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 définit encore en random après
- 2.

Pourquoi ça converge? Grâce 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 sûr 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 crée une colonne vide, pleine de valeur manquante pour le quartier et on va faire EM dessus. C'est assez drôle on crée 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 Chaîne de Markov

Cette fois-ci on va pas utiliser des données sous forme matricielle iid. Ici on va 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 vraisemblance
2. .
3. proba à posteriori

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

Diapo 9 :

Au final on modélise la dépendance par la chaîne de Markov. On va essayer d'associer à une séquence, un label pour prédire la classe. Paramètre du modèle $\{\Pi, A\}$. Permet de faire des prévisions 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'état présent. Si on prend plus ça rajoute 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 Π = 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 irréductible $\rightarrow \mu$ est unique

Définition 6.2 (Irréductible finie). Si on part d'un état donnée, la probabilité d'y revenir est non 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 chaîne 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 chaînes de Markov. On a convergence vers la moyenne.*

*Les chaînes 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 vraisemblance, mais cette fois on a des contraintes sur les distributions de proba égal à 1 (comme avec la multinomiale). \rightarrow Langrandon

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

Nouveau cours du 25/10

8 Procédure d'évaluation

Méthode d'évaluation :

- beaucoup de donnée : On sépare train/Test
- Le plus souvent : Cross validation
- Leave one out : Sur un set de N , on prend $N - 1$ pour apprendre, et 1 pour évaluer puis on fait toute les permutations.

Diapo 4 : Le risque de overfitting dépend de :

- Du nombre de données
- Du nombre de paramètre aussi comme dans le TME.

9 Chaîne de Markov Cachée

- Séparation des observation et des états : une séquence d'observation, avec chaque observation générée par un état.
- Même formalisme que pour la chaîne de Markov
- Pour les observation, on modélise une matrice d'observation des données B qui pour chaque état associe une proba d'observer les données. Proba d'observer w_i dans l'état s_i .
- Diapo 10 : en plus de ce qu'on avait avant, Pour une observation X_i une proba d'être dans l'état 1, 2, 3 (comme en RL)

- Les Hypothèse importante :
 - On a toujours l'hypothèse de Markov
 - Les observation successives sont indépendantes conditionnellement aux états. Une observation = dépend d'un état seulement.
- Notation $s_1^t = s_1, \dots, s_t$

Trois problèmes pour les POMDP :

- Evaluation : calculer une proba pour une séquence d'observation
- Décodage : Séquence d'observation, quelle séquence d'états a généré les observation
- Apprentissage : Trouver les paramètre de mon modèle à partir d'une série d'observation.

Calcul de la proba des observations? $P(x_1^T | \lambda)$

- Diapo 22 : On peut pas utiliser une proba totale en sommant pour tous les états
- → On utilise l'HP de Markov, super on peut le faire par récursion, c'est assez cohérent avec le fait qu'on travail sur une chaîne.
- Voir diapo pour le détail de la récursion j'ai pas écouté

Problème du décodage :

- Même problème que tout à l'heure
- Donc on utilise l'HP de Markov
- Et on peut retrouver un truc récursif : l'algo de Viterbi

Problème d'apprentissage : diapo

- Gros lien avec EM : Donnée manquante → C'est un cas particulier de EM
- Algo diapo 32
- Il faut comprendre les deux problèmes précédents
- Dans la suite des diapo il détaille le même algo mais en plus complexe qui ne donne pas une assignation dur

Convergence de la vraisemblance : diapo 44

- Le problème c'est qu'on optimise sur une fonction qui n'est pas convexe → On risque de tomber sur des max locaux → L'algo est sensible à l'initialisation

En continue :

- On fait pareil pour estimer les paramètres de la loi

Exemple d'application :

- NLP dans l'étiquetage morpho-syntaxique (verbe, sujet, ect). Observation : le corpus de mot, Etat : les propriété morpho-syntaxique (sujet, verbe, ect)
- Reconnaissance de la paroles : signal audio = donnée séquentiel
- Reconnaissance de l'écriture
- Segmentation d'image : Mais on passe en 2D : o → ça devient super dur à calculer. Idée de chaîne n'est plus existante : plus de récursion. Du coup on fait des méthode approximé ou on transforme l'image en un arbre

Nouveau cours du 22/11

Je suis pas aller en cours pour réviser et parce que j'ouvre jamais ce document il sere à rien

Nouveau cours du 29/11

- Regression = prédiction d'une variable continue
- X entrées, Y sortie continue, ϵ = résidu = erreur = bruit
- Modélisation : $Y = \alpha + \beta X + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$
- $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma)$
- On peut trouver une solution analytique à β, α
- Par changement de variable on peut passer sur du non linéaire

Modélisation par max vraisemblance

Par fonction de cout :

L'indicateur R^2

- Si proche de 1 : la variance des données est entièrement expliqué par les données, la variance des erreur e_i vaut 0
- A l'inverse si petit alors variance des erreurs fortes
- On peut aussi tester le modèle sur un set de test

Passage aux données multi-dimensionnelles en X

- On veut estimer $E(Y|X_1, \dots, X_n)$
- on transforme tout en matrice
- $W = (X^T X)^{-1} X^T Y$

Passage au non linéaire :

- On réécrit nos données comme $Xe = [1, X, X * X]$
- Puis solution standard

Exemple de fonction de cout :

- Cout charnière : voir la forme sur internet, lorsque bonne prédiction, x_i et y_i du même signe \rightarrow égal à zéros comme ça on pénalise pas le modèle dans la fonction de coût (proche du perceptron)

Optimisation général

- On peut soit faire une résolution analytique
- ou une descente de gradient, souvent moins coûteuse qu'une inversion de matrice
- Quand on a beaucoup d'exemple, calculer les moyennes ect ça peut être long. Donc on peut calculer le gradient sur un sous ensemble de $X \rightarrow$ plus d'update du gradient mais on a un gradient moins précis.