Prérequis Mathématiques

Jian Tang
HEC Montréal
Institut IA Mila-Québec
Courriel : jian.tang@hec.ca

Mathématiques

- Algèbre Linéaire
- Probabilités et Statistiques
- Fondements de l’apprentissage automatique
- Optimisation

Algèbre Linéaire et Probabilités

Scalaires, Vecteurs et Matrices

- **Scalaires**: une seule valeur, ex: $x = 1.5 \in R$
- **Vecteurs**: Un tableau (liste) de valeurs. Un vecteur $\boldsymbol{x}$ à n dimensions:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n$$

- **Matrices**: Une matrice est un tableau en 2-D de valeurs, alors chaque élément est identifié par deux indices au lieu d'un

$$\boldsymbol{A} = \begin{bmatrix} A_{11}, A_{12} \\ A_{21}\ , A_{22} \end{bmatrix} \in R^{2 \times 2}$$

Transposer des Vecteurs et Matrices

- Transposer un vecteur $\boldsymbol{x}$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n \qquad \boldsymbol{x}^{\boldsymbol{T}} = (x_1, x_2, \dots, x_n)$$

- Transposer une matrice $\boldsymbol{A}$: $\left(\boldsymbol{A}^{\boldsymbol{T}}\right)_{\boldsymbol{ij}} = \boldsymbol{A}_{\boldsymbol{ji}}$

$$\boldsymbol{A} = \begin{bmatrix} A_{11}, A_{12} \\ A_{21}\ , A_{22} \end{bmatrix} \qquad \boldsymbol{A}^{\boldsymbol{T}} = \begin{bmatrix} A_{11}, A_{21} \\ A_{12}\ , A_{22} \end{bmatrix}$$

Opérations

- Pour deux vecteurs:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \in R^n$$

- On a

$$\boldsymbol{x} + \boldsymbol{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \dots \\ x_n + y_n \end{pmatrix} \qquad \boldsymbol{x} - \boldsymbol{y} = \begin{pmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \dots \\ x_n - y_n \end{pmatrix}$$

- Produit Interne (Dot Product)

$$\boldsymbol{x} \cdot \boldsymbol{y} = x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{k=1}^{n} x_k y_k$$

Opérations

- Multiplier un scalaire et un vecteur

$$a \in R \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n \qquad a\boldsymbol{x} = \begin{pmatrix} ax_1 \\ ax_2 \\ \dots \\ ax_n \end{pmatrix} \in R^n$$

- Multiplier deux matrices : **C** = **AB**

$$\boldsymbol{C}_{ij} = \sum_k \boldsymbol{A}_{ik} \boldsymbol{B}_{kj}$$

- Notons que le nombre de colonnes dans **A** doit être égal au nombre de lignes dans **B**

Normes

- Norme L^p d'un vecteur $\boldsymbol{x}$

$$||\boldsymbol{x}||_{\boldsymbol{p}} = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

- Une norme courante est la norme L^2

$$||\boldsymbol{x}||_{\boldsymbol{2}} = \sqrt{\sum_i x_i^2}$$

Probabilités

- Plusieurs évènements de la vraie vie sont incertains. Les probabilités sont utilisées pour capturer cette incertitude.
- Exemple:
  - Quel serait le résultat de rouler un dé?
  - Comment sera la météo la semaine prochaine?

	M	T	W	TH	F	S	S
Chance of rainfall	70%	80%	90%	80%	60%	20%	0%

Variables Aléatoires & Fonction de Probabilité

- Une **variable aléatoire** est une variable qui peut prendre plusieurs états aléatoirement.
- Par exemple
  - X1 représente le résultat de rouler un dé $X1 \in \{1,2,3,4,5,6\}$
  - X2 représente la température de demain
- Une fonction de probabilité décrit à quel point il est plausible qu'une variable aléatoire $p(X)$ ou un ensemble de variables puisse prendre chacun des états disponibles $p(X1, X2, \ldots)$

Variables Aléatoires Discrètes et Fonction de masse

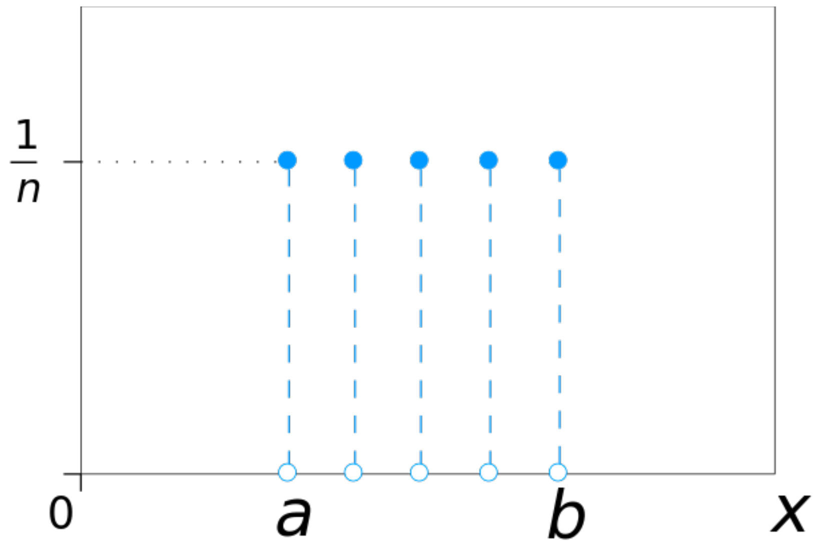
- Une variable aléatoire discrète prend un nombre fini de valeurs
- Une fonction de distribution pour une variable aléatoire discrète peut être décrite comme une fonction de masse (PMF): $p\ (X)$

$$p(X = x_i) \geq 0, \forall i$$

$$\sum_i p(X = x_i) = 1$$

- Ex: Distribution uniforme discrète

$$p(X = x_i) = \frac{1}{n}, \forall i$$

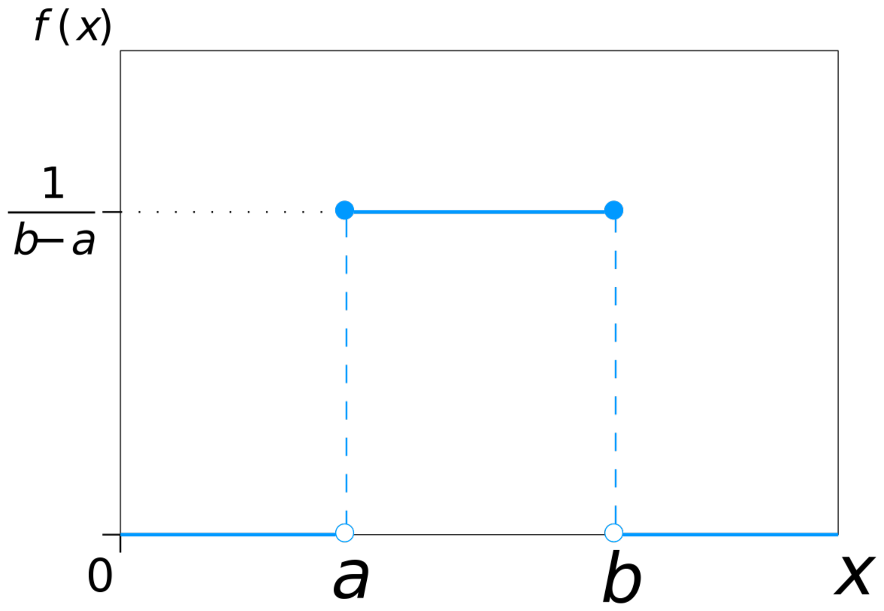


Variables Aléatoires Continues et Fonction de Densité

- Les variables aléatoires continues sont décrites par une fonction de densité $f(x)$:

$$f(x) \geq 0, \forall x \in X$$

$$\int f(x)dx = 1$$



- Ex: distribution uniforme continue

$$f(x) = \frac{1}{b-a}, \forall a \leq x \leq b$$

Propriétés des Distributions de Probabilités

- Formule des probabilités totales: $p(x) = \sum_y p(x, y)$

- Formule des produits: $p(x, y) = p(x|y)p(y)$

- Théorème de Bayes: $p(y|x) \propto p(x|y)p(y)$

Valeur Espérée & Variance

- **Valeur Espérée**: La valeur moyenne de X provenant de $p(X)$

$$E[X] = \sum_i p(X = x_i)x_i$$

- **Variance**: une mesure de combien varie x lorsqu'on échantillonne différentes valeurs de X provenant de la distribution $p\,(X)$

$$Var[X] = E\left[\left(X - E(X)\right)^2\right]$$

Variable Binaire

- Une variable Binaire $X \in \{0, 1\}$, ex. jeter une pièce. X = 1 représente face et X = 0 représente pile.
- Définissons la probabilité d'obtenir face comme:

$$p(X = 1) = \mu$$
$$p(X = 0) = 1 - \mu$$

$$E[X] = \mu \qquad\qquad Var[X] = \mu(1 - \mu)$$

Variables Multinomiales

- Considérons une variable aléatoire pouvant prendre une valeur parmi K et que ces valeurs sont toutes des états mutuellement exclusifs (ex: rouler un dé).
- Nous utiliserons un schéma d'encodage de 1-parmi-K.
- Si une variable aléatoire peut prendre K=6 états et qu'une observation particulière de la variable correspond à l'état x_3=1, alors $\boldsymbol{x}$ sera présenté ainsi:

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$$

- Si on note la probabilité de x_k=1 par le paramètre μ_k, alors la distribution pour $\boldsymbol{x}$ est définie par:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \forall k : \mu_k \geqslant 0 \quad \text{et} \quad \sum_{k=1}^{K} \mu_k = 1$$

Estimation par Maximum de Vraisemblance

- Supposons que nous observons un jeu de données $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- On peut construire la fonction de vraisemblance, qui est une fonction de µ.

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$$

- Notons que la fonction de vraisemblance dépend seulement des N points de données par les K quantités suivantes:

$$m_k = \sum_n x_{nk}, \quad k = 1, \ldots, K.$$

- Ceci représente le nombre d'observations de $x_k = 1$.
- Ces statistiques sont dites exhaustives à cette distribution.

Estimation par Maximum de Vraisemblance

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k}$$

- Pour trouver une solution par maximum de vraisemblance à μ,nous devons maximiser le maximum de vraisemblance en prenant en compte la contrainte: $\sum_k \mu_k = 1$

- On utilise l'algorithme du lagrangien: $\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K}\mu_k - 1\right)$

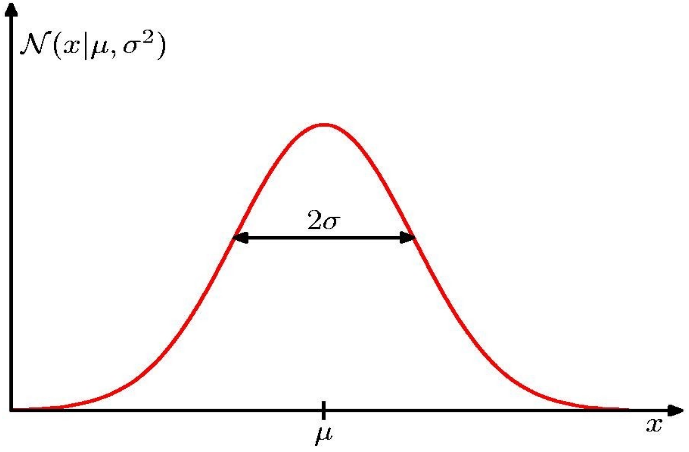
$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N} \qquad \lambda = -N$$

Ce qui est la fraction d'observations pour lesquelles x_k=1.

Distribution Gaussienne Univariée

- Dans le cas d'une variable individuelle x, la distribution gaussienne prend la forme:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



Et est gouvernée par deux paramètres:

- μ (moyenne)
- σ^2(variance)

- La distribution Gaussienne satisfait:

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

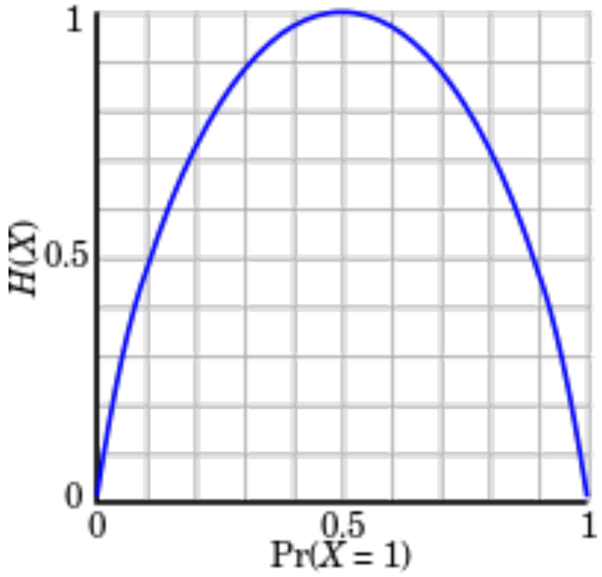
$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

Entropie de Shannon

- L'entropie $H(X)$ d'une distribution $P(X)$ caractérise la quantité d'incertitude d'une variable aléatoire X.

$$H(X) = -\sum P(x) \log P(x) = -\mathbb{E}_{x \sim P} \log P(x)$$

- Ex: X est une variable binaire



La Divergence de Kullback-Leibler (KL)

- KL-divergence: mesurer la distance entre deux distributions de probabilités $P(x)$ et $Q(x)$

$$D_{KL}(P||Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)]$$

- Note:
  - $D_{KL}(P||Q) \geq 0$
  - $D_{KL}(P||Q) = 0$ si et seulement si $P = Q$
  - $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

Entropie Croisée H(P, Q)

- Une autre fonction pour mesurer la distance entre deux fonctions P(x) et Q(x)

$$CE(P,Q) = -\mathbb{E}_{x \sim P} \log Q(x)$$

- On peut trouver que

$$CE(P,Q) = H(P) + D_{KL}(P||Q)$$

- Minimiser l’entropie croisée pour Q est équivalent à minimiser la KL-divergence.

Merci!

jian.tang@hec.ca