EJN 2022
Outils mathématiques pour l'apprentissage statistique

Jean-Marc Martinez

Université Paris-Saclay, CEA
Département de Modélisation des Systèmes et Structures
91191 Gif-sur-Yvette, France

CEA Digitéo Saclay 6-8 avril 2022

Variable aléatoire

- Notation : X variable aléatoire à valeurs dans $\mathcal{X}$.

Variable aléatoire discrète à valeurs dans $\mathcal{X} = \{x_1, x_2, \ldots\}$ (dénombrable fini ou non)

- Loi de probabilité $p(x) = \mathbb{P}(X = x)$ avec $\sum_{x \in \mathcal{X}} p(x) = 1$
  - $\mathcal{X}$ ensemble de labels : {labrador, caniche, teckel,...} ou de nombres : {0,1}
  - $\mathcal{X} = \{0, 1\}$, loi de Bernoulli $\mathcal{B}(1, \theta)$ $\quad p(x|\theta) = \theta^x (1-\theta)^{1-x}$
  - $\mathcal{X} = \{0, 1, \ldots, n\}$, loi binomiale $\mathcal{B}(n, \theta) := \sum_{i=1}^{n} \mathcal{B}(1, \theta)$ $\quad p(x|\theta) = C_n^x \theta^x (1-\theta)^{n-x}$

Variable aléatoire à densité dans $\mathcal{X} \subset \mathbb{R}$

- Loi de probabilité $\mathbb{P}(X \leq x) = F_X(x)$ fonction de répartition, monotone, croissante
- Densité de probabilité $p(x)$:

$$p(x) = \frac{\partial F_X(x)}{\partial x} \geq 0 \quad \texttt{(existence)}$$

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b p(x) dx$$

- Loi normale $\mathcal{N}(\mu, \sigma^2)$, $\quad \mathcal{X} = \mathbb{R}$, $\quad p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Couple de variables aléatoires

Discrètes	à Densité
(X, Y) dans $\mathcal{X} \times \mathcal{Y}$ dénombrables	(X, Y) dans $\mathcal{X} \times \mathcal{Y} \in \mathbb{R}^2$
Loi jointe : $p(x, y) = \mathbb{P}(X = x, Y = y)$	Loi jointe : $\mathbb{P}(X \leq x, Y \leq y)$ = Fonction de répartition $F_{(X,Y)}(x, y)$ Densité $p(x, y) = \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x, y)$
Lois marginales : $\mathbb{P}(X = x) = p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$ $\mathbb{P}(Y = y) = p(y) = \sum_{x \in \mathcal{X}} p(x, y)$	Lois marginales : $F_X(x) = \mathbb{P}(X \leq x, Y \leq \infty)$ $F_Y(y) = \mathbb{P}(X \leq \infty, Y \leq y)$ Densités marginales : $p(x) = \frac{\partial}{\partial x} F_X(x) = \int p(x, y) dy$ $p(y) = \frac{\partial}{\partial y} F_Y(y) = \int p(x, y) dx$
Lois conditionnelles (Bayes) : $\mathbb{P}(Y = y \vert X = x) = p(y \vert x) = \frac{p(x,y)}{p(x)}$ $\mathbb{P}(X = x \vert Y = y) = p(x \vert y) = \frac{p(x,y)}{p(y)}$	Densités conditionnelles (Bayes) : $p(y \vert x) = \frac{p(x,y)}{p(x)}$ $p(x \vert y) = \frac{p(x,y)}{p(y)}$
Indépendance : $p(x, y) = p(x)p(y)$	Indépendance : $p(x, y) = p(x)p(y)$ $F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$

- Pour simplifier : notations dans les 2 cas (variables discretes ou à densité) similaires
  - toutes les lois et densités sont notées par $p()$
  - distinction (par variable, jointe, marginale et conditionnelle) se fait en fonction des arguments de $p()$

Echantillon

Un n-échantillon $(x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ obtenu par tirages aléatoires (i.i.d.) d'une loi X inconnue

- tirages i.i.d. : indépendants, identiquement distribués (issus de la même loi)

Inférence statistique

Estimer certaines caractéristiques de la loi X (moyenne, variance, quantiles) ou les paramètres θ inconnus de sa loi/densité $p(x|\theta)$ (familles de lois supposée/connue)

- notation conditionnelle $x|\theta$ fait apparaître la dépendance de la loi en fonction de θ.

Statistique

Une statistique S_n est une fonction des n valeurs du n-échantillon $\rightarrow s_n(x^{(1)}, x^{(2)}, \ldots, x^{(n)})$

- n-échantillon aléatoire $\rightarrow S_n = (X^{(1)}, X^{(2)}, \ldots, X^{(n)})$ est une variable aléatoire

Estimateur $\rightarrow$ estimation

Estimateur = statistique estimant les caractéristiques de X ou les paramètres θ de sa loi/densité

- estimateur de $\mathbb{E}(X)$ par la moyenne empirique $S_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$
- estimation = une réalisation aléatoire de l'estimateur $s_n = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$

Vraisemblance

$D_n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ un n-échantillon (i.i.d.) de $X \sim p(x|\theta)$ où θ inconnu. La vraisemblance de θ conditionnellement au n-échantillon est définie par :

$$L(\theta|D_n) = \prod_{i=1}^{n} p(x^{(i)}|\theta)$$

- La vraisemblance quantifie la probabilité de la réalisation du n-échantillon en fonction de θ.

Estimateur par maximum de vraisemblance

L'estimateur de θ par maximum de vraisemblance est la valeur qui maximise la vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} L(\theta|D_n) \quad \texttt{conditions nécessaires :} \quad \frac{\partial}{\partial\theta} L(\theta|D) = 0 \quad \texttt{et} \quad \frac{\partial^2}{\partial\theta^2} L(\theta|D) < 0 \quad \texttt{en} \quad \hat{\theta}$$

Propriéte du maximum de vraisemblance

Lorsque l'estimateur existe, sa loi asymptotique est normale, sans biais et il est efficace (variance minimale atteinte (borne de Cramér-Rao))

- exemple pour la loi normale $X \sim \mathcal{N}(\mu, \sigma^2)$. Dans ce cas $\theta = (\mu, \sigma^2)$:

$$\lim_{n\to\infty} \begin{pmatrix} \hat{\mu} \\ \widehat{\sigma^2} \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \right)$$

Estimateur par maximum de vraisemblance pour une loi de Bernoulli

Soit $D_n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ un n-échantillon (i.i.d) d'une loi de Bernoulli $X \sim \mathcal{B}(1, \theta)$.
Les $x^{(k)}$ sont dans $\{0, 1\}$, $\mathbb{P}(X = 1) = \theta$ et $\mathbb{P}(X = 0) = 1 - \theta$. On a donc :

$$\mathbb{P}(X = x) = p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

La vraisemblance de θ conditionnellement à l'échantillon D est :

$$L(\theta|D_n) = \prod_{k=1}^{n} \theta^{x^{(k)}} (1 - \theta)^{1-x^{(k)}}$$

Considérons la fonction log-vraisemblance (plus facile à manipuler pour le calcul du maximum) :

$$\ell(\theta|D_n) = \ln L(\theta|D_n) = \sum_k [x^{(k)} \log \theta + (1 - x^{(k)}) \log(1 - \theta)]$$

$$\frac{\partial}{\partial \theta} \ell(\theta|D_n) = \sum_{k=1}^{n} [\frac{x^{(k)}}{\theta} - \frac{1 - x^{(k)}}{1 - \theta}] = \frac{1}{\theta(1 - \theta)} \sum_{k=1}^{n} (x^{(k)} - \theta)$$

$$\frac{\partial}{\partial \theta} \ell(\hat{\theta}|D_n) = 0 \Rightarrow \hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} x^{(k)}$$

On vérifie la condition du maximum $\frac{\partial^2}{\partial \theta^2} \ell(\hat{\theta}|D_n)) < 0$

On retrouve l'estimateur *classique* de la moyenne empirique.

Estimateur par maximum de vraisemblance pour une loi normale

Soit $D_n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ un n-échantillon (i.i.d) d'une loi normale $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

La vraisemblance de (μ, σ^2) conditionnellement à l'échantillon D_n est :

$$L(\mu, \sigma^2|D_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x^{(k)}-\mu)^2}{2\sigma^2}$$

Considérons la moins log-vraisemblance à minimiser par rapport aux variables (μ, σ^2) :

$$\ell(\theta|D_n) = \sum_k \frac{1}{2} \ln(\sigma^2) + \frac{(x^{(k)}-\mu)^2}{2\sigma^2} + \texttt{Constante}$$

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma^2|D_n) \propto - \sum_{k=1}^{n} (x^{(k)} - \mu), \quad \frac{\partial}{\partial \mu} \ell(\hat{\mu}, \sigma^2|D_n) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x^{(k)}$$

$$\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2|D_n) \propto \sum_{k=1}^{n} [\frac{1}{\sigma^2} - \frac{1}{\sigma^4}(x^{(k)} - \mu)^2], \quad \frac{\partial}{\partial \sigma^2} \ell(\hat{\mu}, \widehat{\sigma^2}|D_n) = 0 \Rightarrow \widehat{\sigma^2} = \frac{1}{n} \sum_{k=1}^{n} (x^{(k)} - \hat{\mu})^2$$

On vérifie les conditions du minimum $\frac{\partial^2}{\partial \mu^2} \ell(\hat{\mu}, \sigma^2|D_n) > 0$ et $\frac{\partial^2}{\partial \sigma^4} \ell(\hat{\mu}, \widehat{\sigma^2}|D_n) > 0$

On retrouve les estimateurs *classiques* de la moyenne et la variance empirique.

Loi catégorielle/multinoulli à p catégories/valeurs

Généralisation de la loi de Bernoulli → exemple d'un lancer d'un dé à 6 faces
X loi multinoulli dans $\mathcal{X} = \{x_1, x_2, \ldots, x_p\}$ avec $\mathbb{P}(X = x_i) = \theta_i$ et $\sum_{i=1}^{p} \theta_i = 1$.

Définition : codage binaire m parmi p

Un code m parmi p est un code à p bits dont m bits sont à 1, les autres à 0.

Application à la loi multinoulli : Codage binaire 1 parmi p (encodage one-hot)

Toute catégorie $x \in \mathcal{X}$ sera codée par le vecteur y à p composantes toutes nulles sauf une composante à 1 codant le rang de x dans $\mathcal{X} \rightarrow y \in \{0,1\}^p$ et $||y||_1 = 1$.
Exemple du dé à 6 faces : "3" associé à la $3^{\text{ème}}$ catégorie est codé par $y = (0,0,1,0,0,0)$.

Probabilité d'une loi multinoulli à p catégories de paramètres $\theta_{[1:p]}$

$$\mathbb{P}(X = x_i) = \theta_i \text{ pour } i = 1, 2, \ldots, p$$

$$\forall x \in \mathcal{X} \quad \mathbb{P}(X = x) = \mathbb{P}(Y = y(x)) = \prod_{i=1}^{p} \theta_i^{y(x)_i}$$

Vraisemblance d'une loi multinoulli

Soit $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ un n-échantillon (i.i.d) d'une loi mulitinoulli de paramètres $\theta_{[1:p]}$.
Codage 1 parmi $p \rightarrow$ n-échantillon de $Y : D_n = \{y^{(1)}, y^{(2)}, \ldots, y^{(n)}\}$ où $y^{(k)} \in \{0,1\}^p$.
La vraisemblance des paramètres $\theta_{[1:p]}$ conditionnellement à l'échantillon D_n s'exprime par :

$$L(\theta_1, \theta_2, \ldots, \theta_p | D_n) = \prod_{k=1}^{n} \prod_{i=1}^{p} \theta_i^{y_i^{(k)}}$$

D'où la log-vraisemblance :

$$\ell(\theta_1, \theta_2, \ldots, \theta_p | D_n) = \sum_{k=1}^{n} \sum_{i=1}^{p} y_i^{(k)} \ln \theta_i$$

Et en notant n_i le nombre d'occurences de la catégorie x_i parmi n :

$$n_i = \sum_{k=1}^{n} y_i^{(k)} \Rightarrow \ell(\theta_1, \theta_2, \ldots, \theta_p | D_n) = \sum_{i=1}^{p} n_i \ln \theta_i$$

Estimateur par maximum de vraisemblance des paramètres d'une loi multinoulli

Maximisation de la log-vraisemblance sous la contrainte $\sum_{i=1}^{p} \theta_i = 1 \rightarrow$ Lagrangien

$$\begin{aligned} \mathcal{L}(\theta_1, \theta_2, \ldots, \theta_p) &= \ell(\theta_1, \theta_2, \ldots, \theta_p | D_n) + \lambda(\sum_{i=1}^{p} \theta_i - 1) \\ &= \sum_{i=1}^{p} n_i \ln \theta_i + \lambda(\sum_{i=1}^{p} \theta_i - 1) \end{aligned}$$

Calcul des gradients du lagrangien :

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\theta_1, \theta_2, \ldots, \theta_p) = \frac{n_i}{\theta_i} + \lambda, \quad \texttt{pour tout} \quad i = 1, 2, \ldots, p$$

Des conditions d'optimalité, on déduit :

$$\forall i, \frac{n_i}{\hat{\theta}_i} + \lambda = 0 \Rightarrow \frac{n_1}{\hat{\theta}_1} = \frac{n_2}{\hat{\theta}_2} = \ldots = \frac{n_p}{\hat{\theta}_p} = \frac{\sum_{i=1}^{p} n_i}{\sum_{i=1}^{p} \hat{\theta}_i} = n$$

D'où

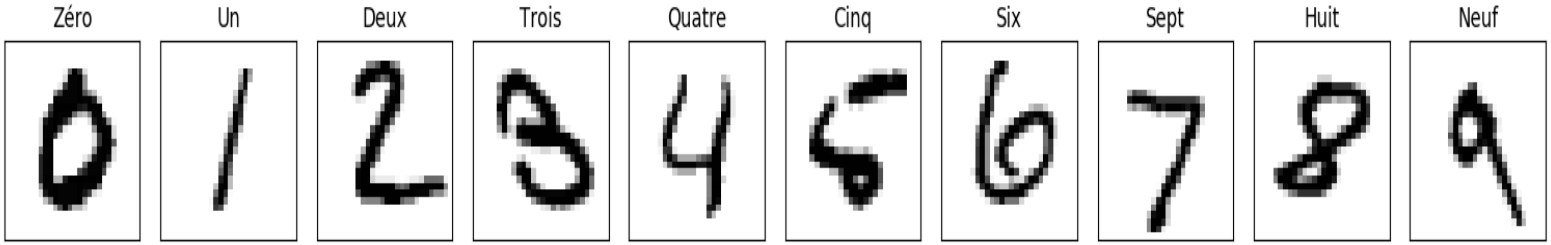
$$\hat{\theta}_i = \frac{n_i}{n}, \quad \texttt{pour} \quad i = 1, 2, \ldots, p$$

On retrouve l'estimateur *classique* de la moyenne empirique.

Bases de données

Un ensemble de données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$ représentant une *corrélation/liaison fonctionnelle* entre les composantes des couples $(x^{(k)}, y^{(k)})$.

Base MNIST
(drosophile de l'IA)
x := images
y := chiffres



Interprétation probabiliste

Chaque $(x^{(k)}, y^{(k)})$ est interprété comme réalisation d'un couple de variables aléatoires (X, Y) et vu comme exemple de liaison entre les variables X et Y.

Base d'exemples D_n vu comme un n-échantillon (i.i.d.) issu de la loi jointe (X, Y).

On note $p(x, y), p(x), p(y|x)$ respectivement la loi/densité jointe de (X, Y), la marginale de X et la conditionnelle de $Y|X = x$. Rappel de la règle de Bayes :

$$p(y|x) = \frac{p(x, y)}{p(x)}$$

Vraisemblance d'un modèle d'une loi conditionnelle

La loi/densité conditionnelle $p(y|x)$ inconnue.
Approximation par un modèle/fonction $p(y|x,w)$, paramètres w à estimer.
Vraisemblance des paramètres w conditionnellement au choix du modèle et aux données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$:

$$L(w|D_n) = \prod_{k=1}^{n} p(y^{(k)}|x^{(k)}, w)$$

Apprentissage supervisé du modèle

L'apprentissage supervisé du modèle de la probabilité conditionnelle vise à rechercher les valeurs w qui maximisent la vraisemblance.
Equivalence à la minimisation de la moins log-vraisemblance appelée *loss function*[a] :

$$\begin{aligned} \ell(w|D_n) &= -\sum_{k=1}^{n} \ln p(y^{(k)}|x^{(k)}, w) \\ \hat{w} &= \arg\min_{w} \ell(w|D_n) \end{aligned}$$

Dans le cas général, la recherche de $\hat{w}$ s'effectue par des algorithmes d'optimisation de type descente de gradients de la *loss function* ($\delta w \propto -\nabla_w \ell(w|D_n)$).

a. On reviendra sur cette hypothèse pour pallier les risques de surapprentissage.

Modèle de la loi de Bernoulli $\mathcal{B}(1, \theta(x))$

Probabilité conditionnelle $p(y = 1|x) = \theta(x)$ fonction de x à valeurs dans $[0, 1]$ inconnue :

$$p(y|x) = \theta(x)^y (1 - \theta(x))^{(1-y)}$$

Approximation de la probabilité conditionnelle $\theta(x)$ par $f(x, w) \in [0, 1]$ pour tout w

$$p(y|x, w) = f(x, w)^y (1 - f(w, x))^{(1-y)}$$

$$\texttt{Vraisemblance } L(w|D_n) = \prod_{k=1}^{n} f(x^{(k)}, w)^{y^{(k)}} (1 - f(x^{(k)}, w))^{1-y^{(k)}}$$

$$\texttt{loss function } \ell(w|D_n) = -\ln L(w|D_n) = -\sum_{k=1}^{n} y^{(k)} \ln f(x^{(k)}, w) + (1 - y^{(k)}) \ln(1 - f(x^{(k)}, w))$$

$$\texttt{estimateur } \hat{w} = \arg\min_w \ell(w|D_n)$$

Modèle de loi normale $\mathcal{N}(\mu(x), \sigma^2)$ (hypothèse d'homoscédasticité)

Données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$ issues du modèle probabiliste :

$$y(x) = \underbrace{\mu(x)}_{\texttt{déterministe}} + \underbrace{\epsilon}_{\texttt{aléatoire } \mathcal{N}(0,\sigma^2)}$$

$$p(y|x,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y-\mu(x))^2}{2\sigma^2}$$

Approximation de la fonction moyenne $\mu(x)$ par $f(x,w) \in \mathbb{R}$

$$p(y|x,w,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y-f(x,w))^2}{2\sigma^2}$$

$$\texttt{Vraisemblance } L(w,\sigma^2|D_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y^{(k)}-f(x^{(k)},w))^2}{2\sigma^2}$$

$$\texttt{loss function } \ell(w,\sigma^2|D_n) = -\ln L(w,\sigma^2|D_n) \propto n \ln \sigma^2 + \frac{1}{\sigma^2} \sum_{k=1}^{n} (y^{(k)}-f(x^{(k)},w))^2$$

$$\texttt{estimateur } (\hat{w}, \widehat{\sigma^2}) = \arg\min_{w,\sigma^2} \ell(w,\sigma^2|D_n)$$

$$\Rightarrow \hat{w} = \arg\min_{w} \sum_{k=1}^{n} (y^{(k)}-f(x^{(k)},w))^2 \ : \texttt{moindres carrés !!}$$

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{k=1}^{n} (y^{(k)}-f(x^{(k)},\hat{w}))^2$$

Modèle de loi normale $\mathcal{N}(\mu(x), \sigma^2(x))$

Données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$ issues du modèle probabiliste :

$$y(x) = \underbrace{\mu(x)}_{\texttt{déterministe}} + \underbrace{\epsilon(x)}_{\texttt{aléatoire } \mathcal{N}(0,\sigma^2(x))}$$

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2(x)}} \exp - \frac{(y-\mu(x))^2}{2\sigma^2(x)}$$

Approximation de la moyenne $\mu(x)$ par $f_1(x, w_1) \in \mathbb{R}$ et de la variance $\sigma^2(x)$ par $f_2(x, w_2) \in \mathbb{R}^+$

$$p(y|x, w_1, w_2) = \frac{1}{\sqrt{2\pi f_2(x, w_2)}} \exp - \frac{(y - f_1(x, w_1))^2}{2 f_2(x, w_2)}$$

$$\texttt{Vraisemblance } L(w_1, w_2|D_n) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi f_2(x^{(k)}, w_2)}} \exp - \frac{(y^{(k)} - f_1(x^{(k)}, w_1))^2}{2 f_2(x^{(k)}, w_2)}$$

$$\texttt{loss function } \ell(w_1, w_2|D_n) \propto \sum_{k=1}^{n} \ln f_2(x^{(k)}, w_2) + \frac{(y^{(k)} - f_1(x^{(k)}, w_1))^2}{f_2(x^{(k)}, w_2)}$$

$$\texttt{estimateur } (\hat{w}_1, \hat{w}_2) = \arg \min_{w_1, w_2} \ell(w_1, w_2|D_n)$$

Modèle de loi catégorielle/multinoulli avec probabilités $\theta_{[1:p]}(x)$

Données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$ avec codage 1 parmi p des $y^{(k)}$.
On note par e_i le codage des p classes (base canonique de $\mathbb{R}^p$) et $\theta_i(x)$ la probabilité conditionnelle de la classe codée par e_i :

$$p(y = e_i|x) = \theta_i(x) \in [0,1]^p$$

On a $\sum_{i=1}^{p} \theta_i(x) = 1$ et pour tout y, la probabilité conditionnelle s'exprime par :

$$p(y|x) = \prod_{i=1}^{p} \theta_i(x)^{y_i}$$

Approximation de $\theta_i(x)$ par $f_i(x,w) \in [0,1]$ avec $\sum_{i=1}^{p} f_i(x,w) = 1$.

$$p(y|x,w) = \prod_{i=1}^{p} f_i(x,w)^{y_i}$$

$$\texttt{Vraisemblance } L(w|D_n) = \prod_{k=1}^{n}\prod_{i=1}^{p} f_i(x^{(k)},w)^{y_i^{(k)}}$$

$$\texttt{loss function } \ell(w|D_n) = -\ln L(w|D_n) = -\sum_{k=1}^{n}\sum_{i=1}^{p} y_i^{(k)} \ln f_i(x^{(k)},w)$$

$$\texttt{estimateur } \hat{w} = \arg\min_{w} \ell(w|D_n)$$

Lois de Bernoulli indépendantes de probabilités $\theta_{[1:p]}(x)$

Exemple d'un lancer de p pièces où chacune des probabilités $\theta_i(x)$ est une fonction de x.

Données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$, $y^{(k)} \in \{0,1\}^p$ ($\neq$ codage 1 parmi p).

Compte tenu de l'indépendance des p lois de Bernoulli :

$$p(y|x) = \prod_{i=1}^{p} p(y_i|x) = \prod_{i=1}^{p} \theta_i(x)^{y_i} (1-\theta_i(x))^{1-y_i}$$

Approximation des $\theta_i(x)$ par $f_i(x, w_i) \in [0,1]$. Posons $w := (w_1, w_2, \ldots, w_p)$:

$$p(y|x,w) = \prod_{i=1}^{p} f_i(x,w)^{y_i} (1-f_i(x,w)^{1-y_i}$$

$$\texttt{Vraisemblance } L(w|D_n) = \prod_{k=1}^{n} \prod_{i=1}^{p} f_i(x^{(k)}, w)^{y_i^{(k)}} (1-f_i(x^{(k)}, w))^{1-y_i^{(k)}}$$

$$\texttt{loss function } \ell(w|D_n) = -\ln L(w|D_n) = -\sum_{k=1}^{n} \sum_{i=1}^{p} y_i^{(k)} \ln f_i(x^{(k)}, w) + (1-y_i^{(k)}) \ln(1-f_i(x^{(k)}, w))$$

$$\texttt{estimateur } \hat{w} = \arg\min_w \ell(w|D_n)$$

Exemple : classification supervisée à 2 classes (1/3)

Classer un élément $x \in \mathbb{R}^d$ dans une des 2 classes notées ω_1, ω_2.

Formule de Bayes définissant la probabilité *a posteriori* $p(\omega_1|x)$ en fonction de la densité de probabilité conditionnelle $p(x|\omega_1)$ et de la probabilité *a priori* $\mathbb{P}(\omega_1)$:

$$p(\omega_1|x) = \frac{p(x|\omega_1)\mathbb{P}(\omega_1)}{p(x|\omega_1)\mathbb{P}(\omega_1) + p(x|\omega_2)\mathbb{P}(\omega_2)}, \quad p(\omega_2|x) = 1 - p(\omega_1|x)$$

• Hypothèse : densités de probabilité conditionnelles $p(x|\omega_i)$ issues de lois normales à d composantes non corrélées de moyenne $\mu^{(i)} \in \mathbb{R}^d$ et de même variance σ^2 :

$$p(x|\omega_i) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{\|(x-\mu^{(i)})\|_2^2}{2\sigma^2}}, \quad i = 1, 2$$

On en déduit (exercice) la densité de probabilité *a posteriori* $p(\omega_1|x)$:

$$p(\omega_1|x) = \frac{1}{1 + e^{-(w^t x + w_0)}} := \texttt{fonction logistique de } (w^t x + w_0)$$

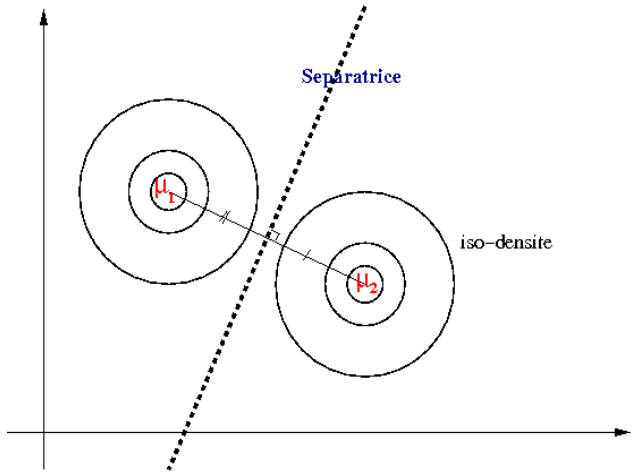
$$w = \frac{\mu^{(1)} - \mu^{(2)}}{\sigma^2} \texttt{ et } w_0 = \frac{\|\mu^{(2)}\|_2^2 - \|\|\mu^{(1)}\|_2^2}{2\sigma^2} + \ln \frac{\mathbb{P}(\omega_1)}{\mathbb{P}(\omega_2)}$$

Exemple : classification supervisée à 2 classes (2/3)

- séparatrice définie par les points x tels que

$$p(\omega_1|x) = p(\omega_2|x) = 0.5 \rightarrow w^t x + w_0 = 0$$

- cas $x \in \mathbb{R}^2$ et $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2)$
  - séparatrice = médiatrice du segment $\mu_1 - \mu_2$



Lien avec la régression logistique

Reformulation de la densité de probabilité *a posteriori*

$$p(\omega_1|x) = \frac{1}{1 + e^{-\ln \frac{p(x|\omega_1)\mathbb{P}(\omega_1)}{p(x|\omega_2)\mathbb{P}(\omega_2)}}}$$

La régression logistique postule le modèle suivant sur le rapport des probabilités *a posteriori* :

$$\ln \frac{p(x|\omega_1)\mathbb{P}(\omega_1)}{p(x|\omega_2)\mathbb{P}(\omega_2)} = w^t x + w_0 \Rightarrow p(\omega_1|x) = \frac{1}{1 + e^{-(w^t x + w_0)}}$$

La régression logistique = *le classifieur optimale de Bayes* lorsque les distributions statistiques des catégories/classes $\sim \mathcal{N}(\mu_i, \sigma^2 I_d)$ (homoscédasticité)

D'où l'intérêt de la fonction logistique pour coder les probabilités *a posteriori*.

Exemple : classification supervisée à 2 classes (3/3)

Paramètres $\mu^{(1)}, \mu^{(2)}, \sigma^2$ et probabilités *a priori* $\mathbb{P}(\omega_1), \mathbb{P}(\omega_2)$ inconnues.

Les paramètres du modèle de la régression logistique sont estimés par maximum de vraisemblance conditionnellement à l'échantillon $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$.

Pour simplifier, on redéfinit $x := (1, x)$ et $w := (w_0, w)$. On a :

$$\begin{aligned}
\texttt{modèle } p(y=1|x,w) &= f(x,w) = \frac{1}{1+e^{-w^t x}} \\
\texttt{loss function } \ell(w|D_n) &= -\sum_{k=1}^{n} y^{(k)} \ln f(x^{(k)}, w) + (1-y^{(k)}) \ln(1 - f(x^{(k)}, w)) \\
&= \sum_{k=1}^{n} y^{(k)} \ln(1+e^{-w^t x^{(k)}}) + (1-y^{(k)}) \ln(1+e^{w^t x^{(k)}}) \\
\texttt{(exercice) } \Rightarrow \nabla_w \ell(w|D_n) &= \sum_{k=1}^{n} [f(x^{(k)}, w) - y^{(k)}] x^{(k)} \quad \texttt{(colinarité)}
\end{aligned}$$

☺ Pour chaque exemple la colinéarité du gradient avec l'entrée $x^{(k)}$! ! ! (les non linéarités, celle du modèle et celle de la loss function, se compensent dans le calcul du gradient).

☺ Facile à coder (TP). En notant $X_{d\times n}$ la matrice de lignes $x^{(k)} \in \mathbb{R}^d$ et $(f-y) \in \mathbb{R}^n$ le vecteur de composantes $(f(x^{(k)}, w) - y^{(k)})$:

$$\nabla_w \ell(w|D_n) = X(f-y)$$

Exemple : classification supervisée à p classes

Classer un élément caractérisé par $x \in \mathbb{R}^d$ dans une des p classes notées $\omega_1, \omega_2, \ldots, \omega_p$.

Cas équivalent à une loi catégorielle/multinoulli (codage 1 parmi p) de probabilité conditionnelle $p(y = e_i|x)$ (où e_i codage des p catégories, base canonique de $\mathbb{R}^p$).

Généralisation du modèle de la régression logistique à p catégories/classes $\rightarrow$ fonction SoftMax.

$$
\begin{aligned}
w &:= (w^{(1)}, w^{(2)}, \ldots, w^{(p)}) \\
p(y = e_i|x, w) &= f_i(x, w) = \frac{e^{x^t w^{(i)}}}{\sum_{j=1}^{p} e^{x^t w^{(j)}}} \in [0, 1] \text{ et } \sum_{i=1}^{p} f_i(x, w) = 1 \\
\texttt{vraisemblance } L(w|D_n) &= \prod_{k=1}^{n} \prod_{i=1}^{p} f_i(x^{(k)}, w)^{y_i^{(k)}} \\
\texttt{loss function } \ell(w|D_n) &= -\ln L(w|D_n) = -\sum_{k=1}^{n} \sum_{i=1}^{p} y_i^{(k)} \ln f_i(x^{(k)}, w) \\
\texttt{(exercice)} \Rightarrow \nabla_{w^{(i)}} \ell(w|D_n) &= \sum_{k=1}^{n} [f_i(x^{(k)}, w) - y_i^{(k)}] x^{(k)}, \ i = 1, 2, \ldots, p
\end{aligned}
$$

☺ Pour chaque exemple colinéarité du gradient avec l'entrée $x^{(k)}$!!! (les non linéarités, celle des p modèles et celle de la loss function, se compensent dans le calcul des gradients).

☺ Facile à coder (TP). $X_{d \times n}$ la matrice de colonnes $x^{(k)} \in \mathbb{R}^d$, $(f - y) \in \mathbb{R}^n$ la matrice $n \times p$ d'éléments $(f(x^{(i)}, w)_j - y_j^{(i)})$, les gradients $\nabla_w \ell(w|D_n)$ (matrice $d \times p$) s'expriment par :

$$\nabla_w \ell(w|D_n) = X(f - y)$$

Modèle linéaire

Données $D_n = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$, où $x \in \mathbb{R}^d, y \in \mathbb{R}$.

Exemple de l'approximation d'une fonction de $\mathbb{R}^d \to \mathbb{R}$ par la fonction $f(x, w)$

On modélise l'écart $[f(x, w) - y]$ par une loi gaussienne $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Cas vu précédemment de l'approximation d'une loi normale $y \sim \mathcal{N}(\mu(x), \sigma^2)$

Approximation de la moyenne $\mu(x)$ par $f(x, w)$ forme linéaire de p régresseurs $h_{[1:p]}(x)$

$$\begin{aligned} f(x, w) &:= \sum_{i=1}^{p} w_i h_i(x) = w^t \mathbf{h}(x) \\ p(y|x, w) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(w^t \mathbf{h}(x) - y)^2}{2\sigma^2} \\ \texttt{vraisemblance } L(w|D_n) &= \prod_{k=1}^{n} p(y^{(k)}|x^{(k)}, w) \end{aligned}$$

Soit $\boldsymbol{H}$ la matrice $n \times p$ d'éléments $\boldsymbol{H}_{ij} = h_j(x^{(i)})$ (rang plein) et $\mathbf{y} \in \mathbb{R}^n$ de composantes $y^{(i)}$:

$$\begin{aligned} \texttt{loss function } \ell(w|D_n) &\propto ||\boldsymbol{H}w - \mathbf{y}||_2^2 \texttt{ moindres carrés} \\ \texttt{estimateurs : } \hat{w} &= (\boldsymbol{H}^t\boldsymbol{H})^{-1}\boldsymbol{H}^t\mathbf{y} \texttt{ et } \widehat{\sigma^2} = \frac{1}{n}||\boldsymbol{H}\hat{w} - \mathbf{y}||_2^2 \end{aligned}$$

Conseil : Premier régresseur $h_1(x) = 1 \to$ erreur moyenne nulle $\sum_{k=1}^{n}(\hat{w}^t\mathbf{h}(x^{(k)}) - y^{(k)}) = 0$

☺ Trivial à coder, formule explicite des estimateurs et cela quels que soient les régresseurs (fonctions de x).

Entropie

Entropie d'une variable aléatoire = mesure de son désordre dû à la variabilité de ses réalisations.

Entropie de Shannon (variables discrètes)

L'entropie d'une variable aléatoire discrète $X \in \mathcal{X}$ avec $\mathbb{P}(X = x) = p(x)$ est définie par :

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \ln(p(x)) = -\mathbb{E}_X[\ln(p)]$$

Propriétés : $0 \leq H(X) \leq \ln|\mathcal{X}|$ où $|\mathcal{X}|$ est le cardinal de $\mathcal{X}$

Entropie différentielle (variables à densité)

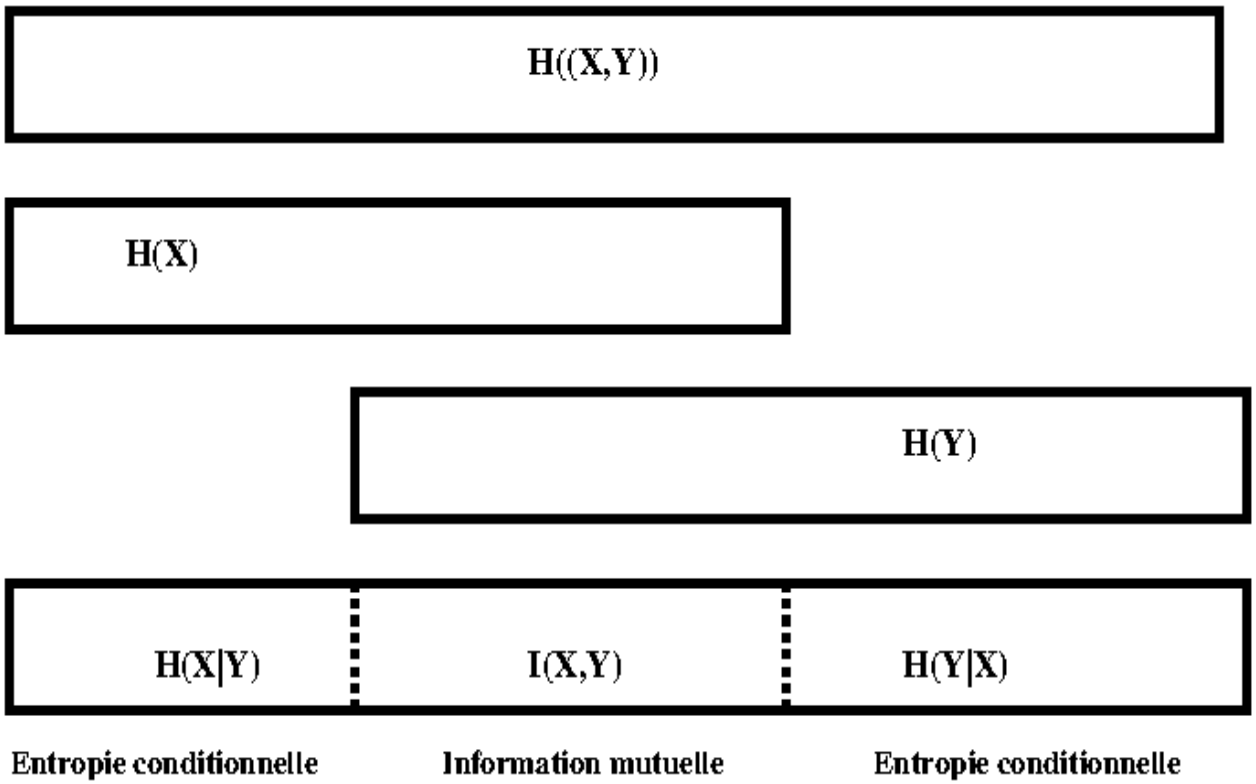
L'entropie différentielle d'une variable aléatoire X à densité $p(x)$ est définie par :

$$H(X) = -\int p(x) \ln(p(x)) dx = -\mathbb{E}_X[\ln(p)]$$

Peut être négative. Quelques exemples :

- loi uniforme $\mathcal{U}(a, b)$: $H(X) = \ln(b - a)$
- loi normale $\mathcal{N}(\mu, \sigma^2)$: $H(X) = \ln[\sigma\sqrt{2\pi e}]$
- loi normale multivariée dans $\mathbb{R}^n$: $\mathcal{N}(\mu, \Sigma)$: $H(X) = (1/2) \ln[(2\pi e)^n |\Sigma|]$

- Loi jointe (X, Y)
- Visualisation des relations entre les entropies des différentes lois
  - jointe $p(x, y)$, marginales $p(x), p(y)$, conditionnelles $p(y|x), p(x|y)$)



- Entropie conditionnelle $H(Y|X)$: réduction en moyenne d'entropie de Y sachant X

$$H(Y|X) = \mathbb{E}_X H(Y|X = x) \Rightarrow H(Y|X) = H((X, Y)) - H(X)$$

- Information mutuelle $I(X, Y)$: une mesure de leur dépendance probabiliste. Sur une des 2 variables, elle quantifie la réduction d'entropie apportée par la connaisssance de l'autre variable

$$I(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) \Rightarrow I(X, Y) = H(X) + H(Y) - H((X, Y))$$

- Divergence de Kullback-Leibler entre 2 lois de probabilité P, Q : une mesure de leur dissimilarité.

$$D_{KL}(P||Q) = \int_x p(x) \ln \frac{p(x)}{q(x)} dx \geq 0, \quad \texttt{non symétrique}, \quad = 0 \ \texttt{ssi} \ P = Q$$

- Entropie croisée entre 2 lois de probabilité P, Q :

$$H(P, Q) = -\int_x p(x) \ln q(x) dx = -\mathbb{E}_{p(x)} \ln q(x)$$

$$\underbrace{H(P, Q)}_{\texttt{entropie croisée}} = \underbrace{D_{KL}(P||Q)}_{\texttt{divergence KL}} - H(P)$$

- Divergence de Jensen-Shanon entre 2 lois de probabilité P, Q : une mesure de leur dissimilarité (utilisée pour l'apprentissage des modèles génératifs, les GANs).

$$D_{JS}(P||Q) = \frac{1}{2} D_{KL}(P||\frac{P+Q}{2}) + \frac{1}{2} D_{KL}(Q||\frac{P+Q}{2}) \in [0, \ln 2] \ \texttt{symétrique}$$

Apprentissage supervisé par minimisation de l'entropie croisée

Soit un modèle $p(y|x,w)$ inféré par apprentissage supervisé à partir d'un n-échantillon $D_n = \{(x^{(k)}, y^{(k)})_{[1:n]}\}$. L'estimation des paramètres w par maximum de vraisemblance est la valeur qui minimise la *loss function* :

$$\ell(w|D_n) \quad = \quad -\sum_{k=1}^{n} \ln p(y^{(k)}|x^{(k)}, w)$$

On peut interpréter la *loss function* par la moyenne empirique de la fonction $\ln p(y|x,w)$ sur les données tirées uniformément du n-échantillon D_n (au facteur $1/n$ près). Notons :

- $p_{\texttt{données}(D_n)}$ la distribution statistique des données (x,y)
- $p_{\texttt{modèle}(w)}$ celle inférée par le modèle $p(y|x,w)$

$$\ell(w|D_n) = \underbrace{-\mathbb{E}_{(x,y)\sim p_{\texttt{données}(D_n)}} \ln p_{\texttt{modèle}(w)}(x,y)}_{\texttt{Entropie croisée : } p_{\texttt{données}(D_n)} \times p_{\texttt{modèle}(w)}(x,y)}$$

Equivalence entre la moins log vraisemblance et l'entropie croisée entre la distribution des données et celle inférée par le modèle aussi bien en classification supervisée qu'en régression.

$$\ell(w|D_n) = H(p_{\texttt{données}(D_n)}, p_{\texttt{modèle}(w)})$$

De la relation entre l'entropie croisée et la divergence de Kullback-Liebler, on a :

$$\begin{aligned} \ell(w|D_n) \quad &= \quad D_{KL}(p_{\texttt{données}(D_n)} || p_{\texttt{modèle}(w)}) - H(p_{\texttt{données}(D_n)}) \\ \Rightarrow \arg\min_w \ell(w|D_n) \quad &= \quad \arg\min_w D_{KL}(p_{\texttt{données}(D_n)} || p_{\texttt{modèle}(w)}) \end{aligned}$$

Les estimateurs par maximum de vraisemblance, par minimisation de l'entropie croisée ou de la divergence de Kullback-Leibler sont donc équivalents.

Malédiction de la grande dimension

Complexité de la recherche d'un modèle de $y = f(x)$ à partir d'une base de données liée à la dimension des entrées

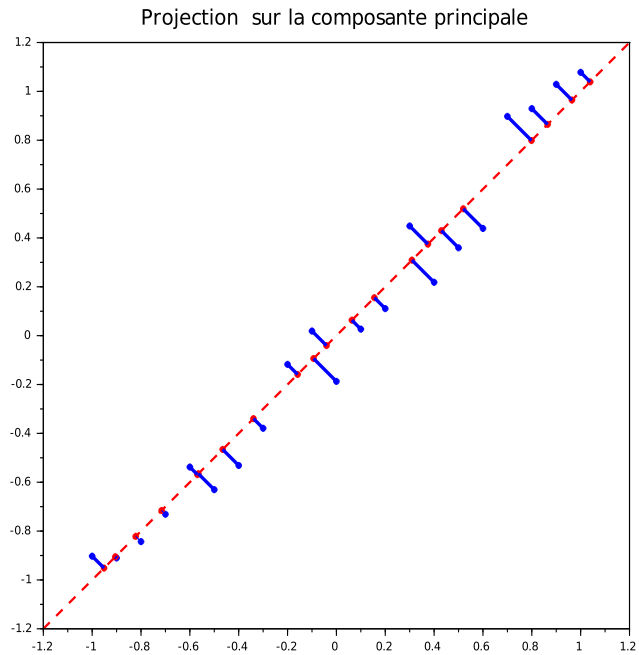
- exemple $x \in [0,1]^d$, n points par dimension $\rightarrow n^d$ au total
- dans la base MNIST (images de chiffres codés sur 784 pixels par $[0:255] \rightarrow x \in 256^{784}$! !

Réduction de dimension

En grande dimension des liaisons entre composantes de $x \rightarrow$ analyse et l'apprentissage simplifiés à condition de *trouver* le/un bon sous espace des x, par exemple par une transformation linéaire

Une solution : réduire la perte de l'approximation de l'inertie (mécanique), de la variance (statistique) du nuage de points par :

- ACP : Analyse en Composantes Principales,
- SVD : Décomposition en Valeurs Singulières, équivalente à l'ACP sur des données centrées



Singular Value Decomposition

Généralisation de la diagonalisation aux matrices rectangles. Toute matrice $X_{n\times p}$ vérifie $(n \geq p)$:

$$X = USV^t$$

- V matrice $(p \times p)$ des vecteurs propres de X^tX, orthogonale $V^tV = VV^t = I_{p\times p}$
- U matrice $(n \times p)$des p premiers vecteurs propres de XX^t, $U^tU = I_{p\times p}$
- S matrice diagonale $(p \times p)$ des p valeurs singulières $\sigma_i \geq 0$
- valeurs singulières, racines carrés des valeurs propres de X^tX, ordonnées $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$
- rang de X (nbr de colonnes lin. ind.) est p si $\sigma_p > 0$ sinon est $r-1$ si $\sigma_r = 0$

Décomposition de la variance/inertie sur les vecteurs propres

$$X^tX = VSU^tUSV^t = VS^2V^t \Rightarrow \texttt{Trace}(X^tX) = \texttt{Trace}(S^2) = \sum_{i=1}^{p} \sigma_i^2$$

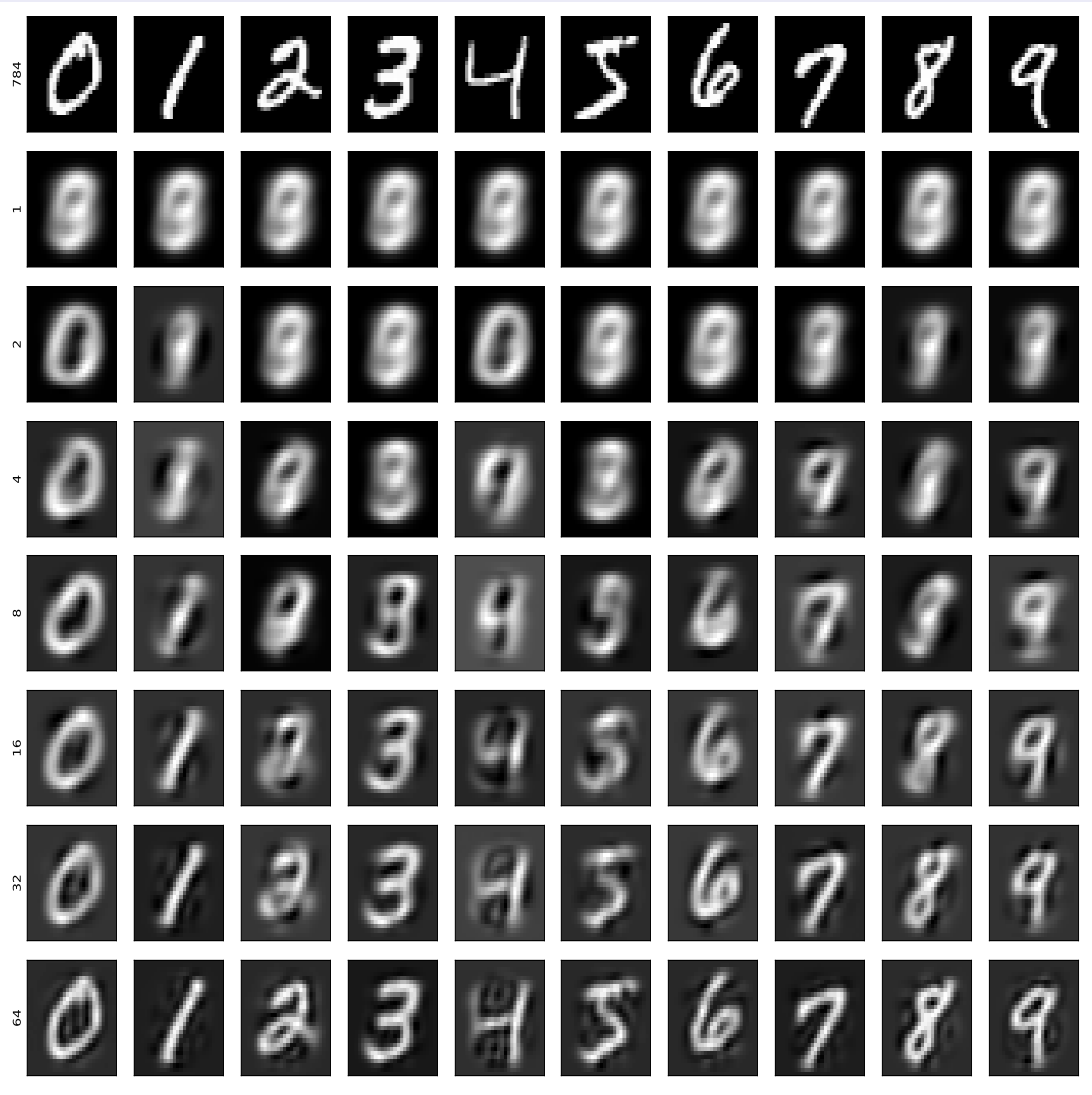
Données centrées $\Rightarrow$ Inertie $= n\mathrm{Var}(X) = \sum_{i,j} X_{i,j}^2 = \texttt{Trace}(X^tX) = \sum_{i=1}^{p} \sigma_i^2$

Meilleure réduction de dimension (critère de la variance) de $p \to p'$ (avec $p' < p$) est la projection sur les p' premiers vecteurs propres $\to$ axes principaux

Nouvelles coordonnées $\hat{X} = US = XV (u = V^tx)$ ou $\hat{X} = U = XVS^{-1} (u = S^{-1}V^tx)$ (données réduites)

Base MNIST : images de chiffres = matrices 28×28 pixels $\in \{0, 1, \ldots, 255\}$

Réduction de dimension par SVD : $(x^{(1)} x^{(2)} \ldots x^{(n)})^t = X_{n \times 784} = USV^t$
Images $x \in \mathbb{R}^{784}$ projetées dans $\mathbb{R}^p$, $p = 1, 2, 4, 8, 16, 32, 64$: $\hat{x} = (V_{784 \times p})^t x$



Entropie conditionnelle

Soit la loi jointe (X, Y). L'entropie conditionnelle $H(Y|X)$ est l'espérance de l'entropie de la loi conditionnelle $Y|X = x$.

$$H(Y|X) = \mathbb{E}_X H(Y|X = x)$$

Théorème

$$H(Y|X) = H((X, Y)) - H(X)$$

Démonstration

Par définition l'entropie conditionnelle $H(Y|X = x)$ est l'entropie de la loi $Y|X = x$. Soient $p(x, y), p(x), p(y|x)$ les densités respectives de la loi jointe, marginale et conditionnelle

$$\begin{aligned}
\mathbb{E}_X H(Y|X = x) &:= -\mathbb{E}_X \int_y p(y|x) \ln(p(y|x)) dy = -\int_{x,y} p(y|x) \ln(p(y|x)) p(x) dxdy \\
&= -\int_{x,y} p(x, y) \ln(p(y|x)) dxdy = -\int_{x,y} p(x, y) \ln \frac{p(x, y)}{p(x)} dxdy \\
&= -\int_{x,y} p(x, y) \ln(p(x, y)) dxdy + \int_{x,y} p(x, y) \ln(p(x)) dxdy \\
&= \underbrace{-\int_{x,y} p(x, y) \ln(p(x, y)) dxdy}_{H((X,Y))} + \underbrace{\int_x p(x) \ln(p(x)) dx}_{-H(X)}
\end{aligned}$$

Information mutuelle

L'information mutuelle $I(X,Y)$ de 2 variables aléatoires est une mesure de leur dépendance probabiliste. Sur une des 2 variables, elle quantifie la réduction d'entropie apportée par la connaisssance de l'autre variable

$$I(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

On en déduit facilement la seconde formulation $I(X,Y) = H(X) + H(Y) - H((X,Y))$

Théorème

$$I(X,Y) = \int_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} dxdy$$

Propriété : $I(X,Y) \geq 0$ et $I(X,Y) = 0 \iff X \perp\!\!\!\perp Y$ (indépendance)

Démonstration

$$\begin{aligned}
I(X,Y) &= \int_{x,y} p(x,y) \ln p(x,y) dxdy - \int_{x,y} p(x,y)(\ln p(x) + \ln p(y)) dxdy \\
&= \int_{x,y} p(x,y) \ln p(x,y) dxdy - \int_{x,y} p(y|x)p(x) \ln p(x) dxdy - \int_{x,y} p(x|y)p(y) \ln p(y) dxdy \\
&= \underbrace{\int_{x,y} p(x,y) \ln p(x,y) dxdy}_{-H((X,Y))} \underbrace{- \int_x p(x) \ln p(x) dx}_{H(X)} \underbrace{- \int_y p(y) \ln p(y) dy}_{H(Y)}
\end{aligned}$$

Information mutuelle entre lois gaussiennes

- Soient 2 vecteurs gaussiens (X, Y) de loi :

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix}\right) = \mathcal{N}(\mu, \Sigma)$$

- L'information mutuelle de X et de Y se déduit à partir de la formule de l'entropie d'une loi normale multivariée $H(X) = (1/2)\ln[(2\pi e)^n|\Sigma|]$

$$I(X, Y) = \frac{1}{2} \ln \frac{|\Sigma_X||\Sigma_Y|}{|\Sigma|}$$

- On vérifie que si les vecteurs X et Y ne sont pas corrélés ($\Sigma_{XY} = 0 \rightarrow |\Sigma| = |\Sigma_X||\Sigma_Y|$) l'information mutuelle est nulle, les variables sont donc indépendantes. En effet dans le cas gaussien il y a équivalence entre indépendance et non corrélation
- Dans le cas à 2 dimensions, l'information mutuelle s'exprime en fonction du coefficient de corrélation

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$$

$$I(X, Y) = \frac{1}{2} \ln \frac{1}{1 - \rho^2}$$

- Dans ce cas, l'information mutuelle est nulle lorsque le coefficient de corrélationn est nul

Divergence de Kullback-Leibler

La divergence de Kullback-Leibler entre 2 lois de probabilité P et Q est une mesure de leur dissimilarité.

$$D_{KL}(P||Q) = \int_x p(x) \ln \frac{p(x)}{q(x)} dx$$

- Attention ce n'est pas une distance car non symétrique
- $D_{KL}(P||Q) \geq 0$ avec égalité si et seulement si $P = Q$ (preque partout)

Théorème : relation entre divergence de Kullback-Liebler et information mutuelle

Soient (X, Y) deux variables alétoires de loi jointe $p(x, y)$ et de lois marginales $p(x), p(y)$:

$$I(X, Y) = D_{KL}(p(x, y)||p(x)p(y))$$

L'information mutuelle de 2 variables X et Y est donc égale à la divergence de Kullback-Leibler entre la loi jointe (X, Y) et la loi produit des 2 lois marginales X et Y.

Démonstration

Evident par définition de l'information mutuelle :

$$I(X, Y) = \underbrace{\int_{x,y} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dxdy}_{D_{KL}(p(x,y)||p(x)p(y))}$$

Entropie croisée

L'entropie croisée entre 2 lois de probabilité P et Q est définie par :

$$H(P,Q) = -\int_x p(x) \ln q(x) dx = -\mathbb{E}_{p(x)} \ln q(x)$$

- Attention à la non symétrie et à sa notation pour la distinguer de la loi jointe notée $H((P,Q))$.

Théorème : relation entre divergence de Kullback-Liebler et entropie croisée

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

Démonstration

$$D_{KL}(P||Q) = \int p(x) \ln \frac{p(x)}{q(x)} dx = \underbrace{\int p(x) \ln p(x) dx}_{-H(P)} \underbrace{- \int p(x) \ln q(x) dx}_{H(P,Q)}$$

- Divergence Kullback-Leibler comme critère en apprentissage supervisé
- Pour les modèles génératifs comme les GANs (Generative Adversial Networks), le critère à minimiser par le Générateur est la divergence de Jensen-Shannon (hyp. Discriminateur optimal)

Divergence de Jensen-Shannon

La divergence Jensen-Shannon est une version symétrisée et lissée de la divergence de Kullback-Leibler

$$D_{JS}(P\|Q) = \frac{1}{2} D_{KL}(P\|\frac{P+Q}{2}) + \frac{1}{2} D_{KL}(Q\|\frac{P+Q}{2})$$

Propriété

$D_{JS}(P\|Q) \in [0, \ln 2]$ et $D_{JS}(P\|Q) = H(\frac{P+Q}{2}) - \frac{H(P)+H(Q)}{2}$

Démonstration

$$\begin{aligned}
D_{JS}(P\|Q) &= \frac{1}{2}\int p(x)\ln(\frac{2p(x)}{p(x)+q(x)})dx + \frac{1}{2}\int q(x)\ln(\frac{2q(x)}{p(x)+q(x)})dx \\
&\leq \frac{1}{2}\int p(x)\ln(2)dx + \frac{1}{2}\int q(x)\ln(2)dx = \ln 2 \\
&= \underbrace{-\int \frac{p(x)+q(x)}{2}\ln(\frac{p(x)+q(x)}{2})dx}_{H(\frac{p+q}{2})} + \frac{1}{2}\underbrace{\int p(x)\ln(p(x))dx}_{-H(p)} + \underbrace{\frac{1}{2}\int q(x)\ln(q(x))dx}_{-H(q)}
\end{aligned}$$

Deux classes $\{\omega_1, \omega_2\}$ d'individus caractérisés par $x \in \mathbb{R}^d$. Loi conditionnelle $\mathbb{P}(x|\omega_i)$ modélise la variabilité des individus au sein de la classe ω_i. Cette loi est supposée normale mutivariée de dimension d :

$$\mathbb{P}(x|\omega_i) = \frac{1}{2\pi^{d/2}|\Sigma_i|^{1/2}} \exp -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)$$

La probabilité *a posteriori* $\mathbb{P}(\omega_1|x)$ (probabilité conditionelle de la classe ω_1 sachant x) est donnée par la formule de Bayes en fonction des probabilités *a priori* $\mathbb{P}(\omega_1), \mathbb{P}(\omega_2)$:

$$\begin{aligned}
\mathbb{P}(\omega_1|x) &= \frac{\mathbb{P}(x|\omega_1)\mathbb{P}(\omega_1)}{\mathbb{P}(x|\omega_1)\mathbb{P}(\omega_1) + \mathbb{P}(x|\omega_2)\mathbb{P}(\omega_2)} \\
&= \frac{1}{1 + \frac{\mathbb{P}(x|\omega_2)\mathbb{P}(\omega_2)}{\mathbb{P}(x|\omega_1)\mathbb{P}(\omega_1)}} \\
\frac{\mathbb{P}(x|\omega_2)\mathbb{P}(\omega_2)}{\mathbb{P}(x|\omega_1)\mathbb{P}(\omega_1)} &= \frac{\mathbb{P}(\omega_2)|\Sigma_1|^{1/2}}{\mathbb{P}(\omega_1)|\Sigma_2|^{1/2}} \exp[-\frac{1}{2}((x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1))] \\
&= \exp[-\frac{1}{2}((x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)) + \ln\frac{\mathbb{P}(\omega_2)|\Sigma_1|^{1/2}}{\mathbb{P}(\omega_1)|\Sigma_2|^{1/2}}] \\
&= \exp[-\frac{1}{2}(x^t(\Sigma_2^{-1} - \Sigma_1^{-1})x + \mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1 - 2(\mu_2^t\Sigma_2^{-1} - \mu_1^t\Sigma_1^{-1})x) + \ln\frac{\mathbb{P}(\omega_2)|\Sigma_1|^{1/2}}{\mathbb{P}(\omega_1)|\Sigma_2|^{1/2}}]
\end{aligned}$$

Les termes quadratiques s'annulent pour $\Sigma_1 = \Sigma_2$:

$$\begin{aligned}
&= \exp[-\frac{1}{2}(\mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1 - 2(\mu_2^t\Sigma_2^{-1} - \mu_1^t\Sigma_1^{-1})x) + \ln\frac{\mathbb{P}(\omega_2)|\Sigma_1|^{1/2}}{\mathbb{P}(\omega_1)|\Sigma_2|^{1/2}}] \\
\frac{\mathbb{P}(x|\omega_2)\mathbb{P}(\omega_2)}{\mathbb{P}(x|\omega_1)\mathbb{P}(\omega_1)} &= \exp -[\underbrace{(\mu_1^t\Sigma_1^{-1} - \mu_2^t\Sigma_2^{-1})}_{w}x + \underbrace{\frac{\mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1}{2} + \ln\frac{\mathbb{P}(\omega_1)|\Sigma_2|^{1/2}}{\mathbb{P}(\omega_2)|\Sigma_1|^{1/2}}}_{w_0}] \\
\Rightarrow \mathbb{P}(\omega_1|x) &= \frac{1}{1 + e^{-(w^tx + w0)}}
\end{aligned}$$