Apprentissage automatique

Formulation probabiliste - rappel de la théorie des probabilités

Théorie des Probabilités

Sujets: variable aléatoire

- La théorie des probabilités est l'outil idéal pour formaliser nos hypothèses et incertitudes par rapport à nos données

- On va traiter nos données comme des **variables aléatoires**
  - la valeur d'une variable aléatoire est incertaine (avant de l'observer)
  - la loi de probabilité de la variable aléatoire caractérise notre incertitude par rapport à sa valeur

Théorie des Probabilités

Sujets: variable aléatoire discrète, probabilité jointe

- Soit X et Y des variables aléatoires **discrètes**
  - X peut prendre comme valeurs $x_1, \ldots, x_M$
  - Y peut prendre comme valeurs $y_1, \ldots, y_M$

- La **probabilité jointe** qu'on observe $X{=}x_i$ et $Y{=}y_j$ est notée

$$p(X = x_i, Y = y_j)$$

Théorie des Probabilités

Sujets: probabilité marginale

- Une **probabilité marginale** est lorsqu'on ne s'intéresse pas à toutes les variables aléatoire qu'on a défini
  - exemple : la probabilité marginale d'observer $X = x_i$

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Théorie des Probabilités

Sujets: probabilité conditionnelle

- Une **probabilité conditionnelle** est lorsqu'on s'intéresse la valeur d'une variable aléatoire «étant donnée» une valeur assignée à d'autres variables
  - exemple : la probabilité que $Y = y_j$ si on suppose que $X = x_i$

$$p(Y = y_j | X = x_i) = \frac{p(Y = y_j, X = x_i)}{p(X = x_i)}$$

  - utile si on veut raisonner par rapport à Y, après avoir observé que $X = x_i$

Théorie des Probabilités

Sujets: règle du produit

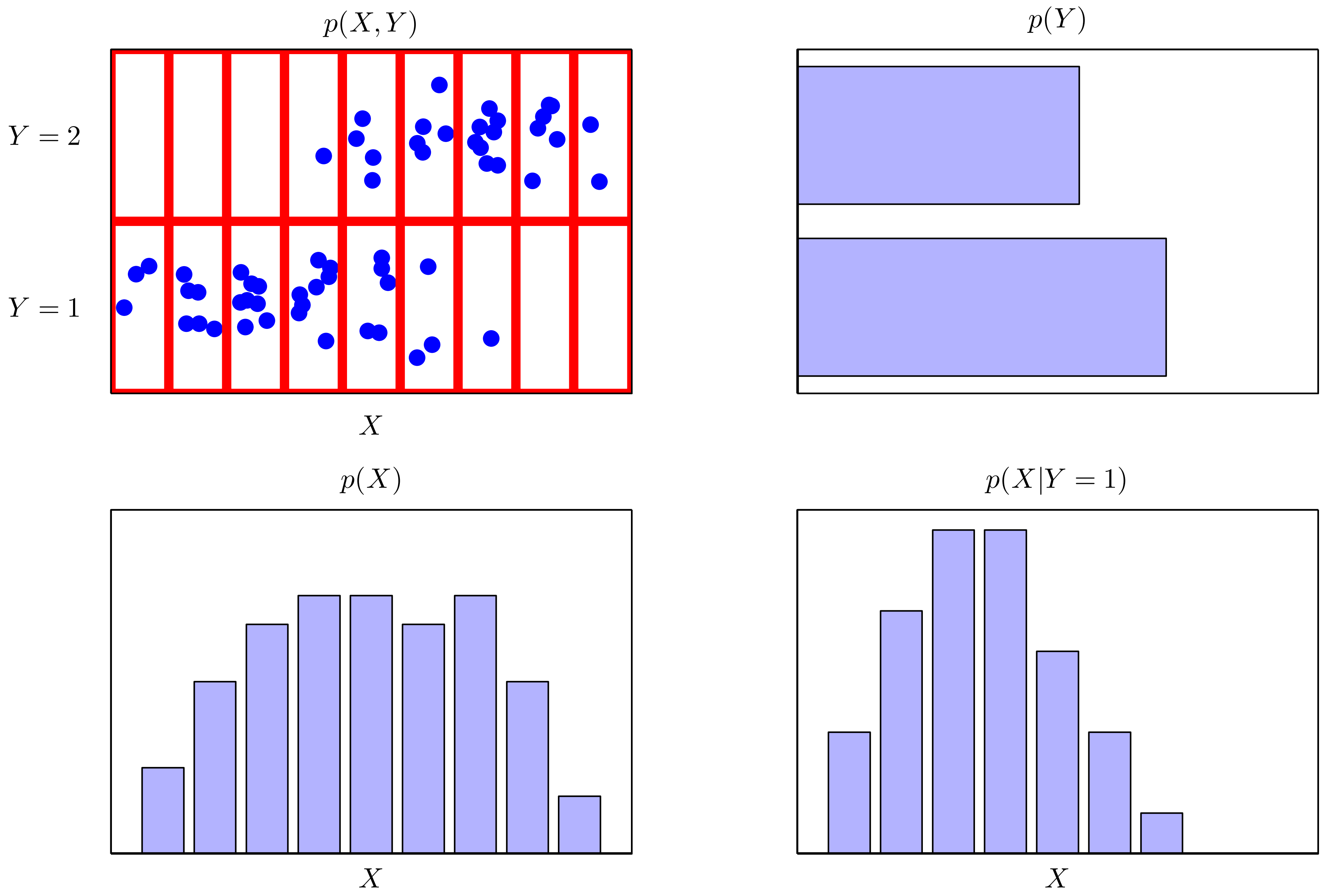
- Une probabilité jointe peut toujours être décomposée dans le produit d'une probabilité conditionnelle et marginale

$$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i) p(X = x_i)$$

- En mots :
  - la probabilité d'observer $X{=}x_i$ et $Y{=}y_j$, c'est la probabilité d'observer $X{=}x_i$ multipliée par la probabilité d'observer $Y{=}y_i$ étant donné que $X{=}x_i$

Théorie des Probabilités

Sujets: probabilités jointes, marginales et conditionnelles



Théorie des probabilités

Sujets: règle de Bayes, loi a priori

x_i et y_j ont disparu, seulement pour simplifier la notation

- La **règle de Bayes** permet d'inverser l'ordre de la conditionnelle

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \text{, où } p(X) = \sum_Y p(X|Y)p(Y)$$

  - $p(Y)$ est appelée loi de probabilité a priori (*prior*)
  - $p(Y|X)$ est appelée loi de probabilité a posteriori (*posterior*)

Théorie des Probabilités

Sujets: indépendance

- Deux variables aléatoires X et Y sont indépendantes si
  - $p(X,Y) = p(X)\,p(Y)$ ou
  - $p(Y \mid X) = p(Y)$ ou
  - $p(X \mid Y) = p(X)$
- En mots : observer la valeur d'une variable ne nous apprend rien sur la valeur de l'autre

Hugo Larochelle

Apprentissage automatique

Formulation probabiliste - variable aléatoire continue

Théorie des Probabilités

Sujets: variable aléatoire continue, fonction de densité

- Soit X une **variable aléatoire continue**
  - X peut prendre un nombre infini de valeurs possibles (e.g. $\mathbb{R}$)
  - X est associée à une **fonction de densité** de probabilité $p(x)$
  - la probabilité que X appartienne à un intervalle (a,b) est

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

Théorie des Probabilités

Sujets: variable aléatoire continue, fonction de densité

- Soit X une **variable aléatoire continue**
  - la fonction de densité doit satisfaire

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\, \mathrm{d}x &= 1 \end{aligned}$$

  - à noter que, contrairement aux probabilités d'une variable discrète, la fonction de densité peut être > 1.
  - peut être vu comme la probabilité que X appartienne à un intervalle infinitésimalement petit centrée en x

Théorie des Probabilités

Sujets: fonction de probabilité cumulative

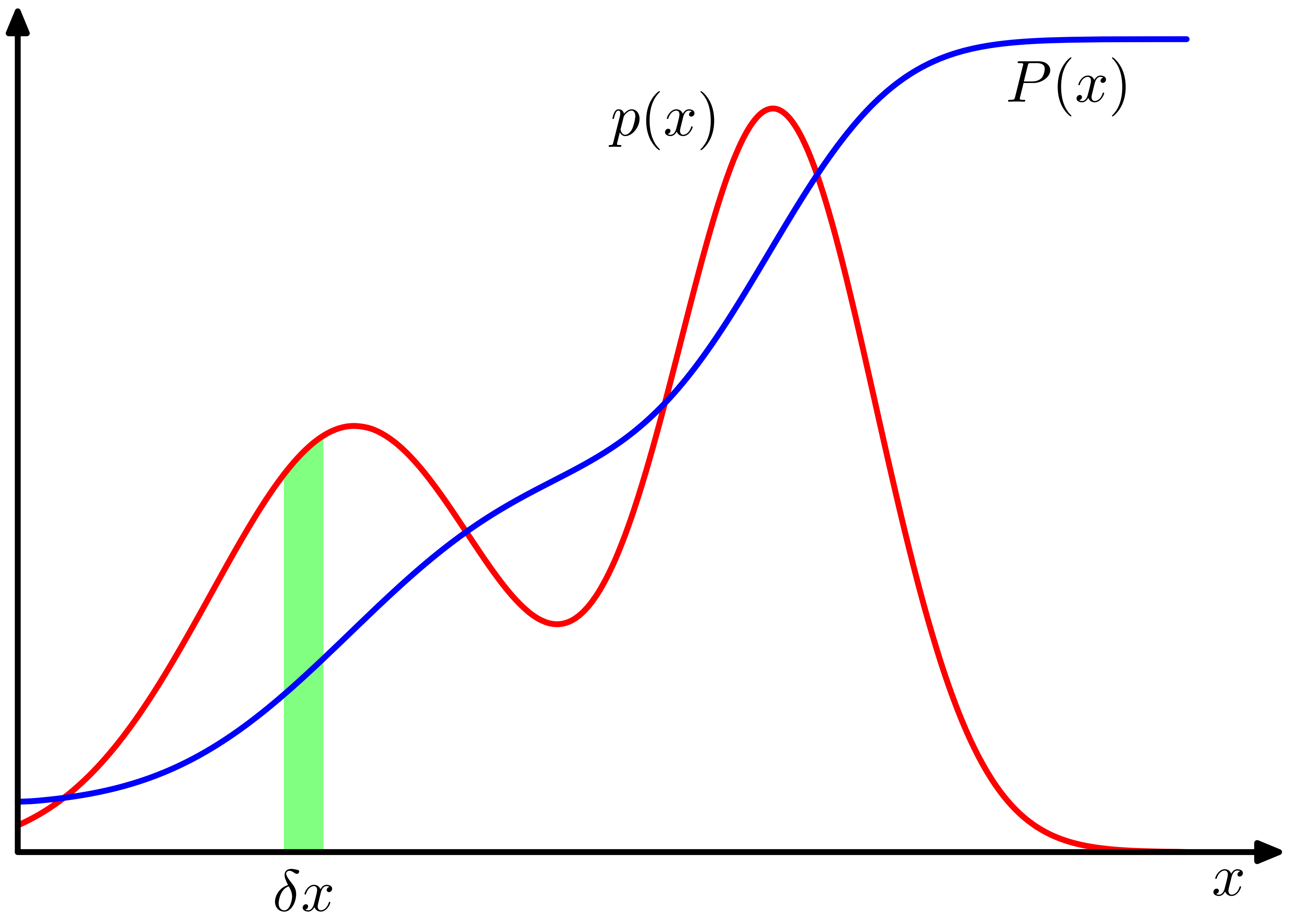
- Soit X une **variable aléatoire continue**
  - la **fonction de répartition** $P(z)$ (*cumulative distribution function*) donne la probabilité que X appartienne à l'intervalle $(-\infty, z)$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

  - les mêmes règles des probabilités marginales et conditionnelles s'appliquent à la fonction de densité
    - les sommes sont remplacées par des intégrales

Théorie des probabilités

Sujets: variable aléatoire continue



Théorie des Probabilités

Sujets: fonction de densité jointe

- Soit X et Y deux **variables aléatoires continues**
  - elles sont associées à une **fonction de densité jointe** $p(x,y)$ telle que :

$$p(x \in (a_x, b_x), y \in (a_y, b_y)) = \int_{a_x}^{b_x} \int_{a_y}^{b_y} p(x, y) \mathrm{d}y \mathrm{d}x$$

Hugo Larochelle

Théorie des Probabilités

Sujets: fonction de densité marginale et conditionnelle

- Soit X et Y deux **variables aléatoires continues**
  - la **fonction de densité marginale** s'obtient en intégrant l'autre variable :

$$p(x) = \int p(x, y) \mathrm{d}y$$

  - la **fonction de densité conditionnelle** s'obtient en divisant par la marginale :

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

Apprentissage automatique

Formulation probabiliste - espérance, variance et covariance

Théorie des Probabilités

Sujets: espérance

- L'**espérance** d'une fonction f d'une variable X est

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \text{(cas discret)}$$

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x \qquad \text{(cas continu)}$$

  - donne une «idée générale» de la valeur de $f(x)$

THÉORIE DES PROBABILITÉS

Sujets: variance

- La **variance** d'une fonction f d'une variable X est

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)] \right)^2 \right]$$

  - mesure à quelle point les valeurs de $f(x)$ varient autour de l'espérance

Théorie des Probabilités

Sujets: propriétés de l'espérance et la variance

- L'espérance d'une transformation linéaire satisfait

$$\mathbb{E}\left[ax+by\right] = a\mathbb{E}\left[x\right] + b\mathbb{E}\left[y\right]$$

a et b sont des constantes

- La variance d'une transformation linéaire satisfait

$$\mathrm{var}\left[ax+by\right] = a^2\mathrm{var}\left[x\right] + b^2\mathrm{var}\left[y\right]$$

seulement si X et Y sont indépendantes

Théorie des Probabilités

Sujets: espérance et variance conditionnelle

- L'espérance et la variance se généralise au cas conditionnel :

$$\mathbb{E}[f(x)|y] = \int f(x)p(x|y)\mathrm{d}x \quad \text{(cas continu)}$$

$$\mathrm{var}[f(x)|y] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)|y])^2 \,|y\right]$$

Sujets: covariance

- La **covariance** entre deux variables X et Y est

$$\begin{aligned} \mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy]-\mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

  - mesure à quel point on peut prédire X à partir de Y (linéairement), et vice-versa
  - si X et Y sont indépendantes, alors la covariance est 0
  - l'inverse n'est pas nécessairement vrai

Théorie des probabilités

Sujets: variable aléatoires multidimensionnelles

- Une variable aléatoire peut être un vecteur
  - la loi de probabilité du vecteur discret est une probabilité jointe

$$p(\mathbf{X} = \mathbf{x}) = p(X_1 = x_1, \ldots, X_D = x_D)$$

  - la fonction de densité d'un vecteur continu intègre à 1

$$\int p(\mathbf{x})\ \mathrm{d}\mathbf{x} = \int_{x_1} \ldots \int_{x_D} p(\mathbf{x})\ \mathrm{d}x_D \ldots \mathrm{d}x_1 = 1$$

où $p(\mathbf{x}) \geq 0$

Théorie des Probabilités

Sujets: variable aléatoires multidimensionnelles

- Une variable aléatoire peut être une vecteur
  - l'espérance d'un vecteur et le vecteur des espérances

$$\mathbb{E}[\mathbf{x}] = (\mathbb{E}[x_1], \ldots, \mathbb{E}[x_D])^{\mathrm{T}}$$

  - la covariance entre deux vecteurs est la matrice des covariances

$$\begin{aligned}\operatorname{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}].\end{aligned}$$

  - on note $\operatorname{cov}[\mathbf{x}] \equiv \operatorname{cov}[\mathbf{x}, \mathbf{x}]$

Théorie des Probabilités

Sujets: variable aléatoires multidimensionnelles

- L'espérance d'une transformation linéaire satisfait

$$\mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mathbb{E}[\mathbf{x}] + \mathbf{b}$$

- La covariance d'une transformation linéaire satisfait

$$\mathrm{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{C}\mathbf{y} + \mathbf{d}] = \mathbf{A}\mathrm{cov}[\mathbf{x}, \mathbf{y}]\mathbf{C}^{\mathrm{T}}$$

$$\mathrm{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mathrm{cov}[\mathbf{x}]\mathbf{A}^{\mathrm{T}}$$

Apprentissage automatique

Formulation probabiliste - loi gaussienne / normale

Loi de probabilité gaussienne

Sujets: loi gaussienne (loi normale)

- La **loi gaussienne** (aussi appelée loi normale) est une loi simple et pratique pour exprimer notre incertitude sur une quantité continue
  - assigne la densité de probabilité la plus élevée à une valeur moyenne μ
  - notre incertitude est exprimée par la variance σ^2 (ou l'écart-type σ)
  - exemple : «la réclamation des clients prend une valeur autour de μ \$, mais varie selon un écart-type de σ \$»

Loi de Probabilité Gaussienne

Sujets: loi gaussienne (loi normale)

- Une variable aléatoire suivant une loi gaussienne a la fonction de densité suivante :

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

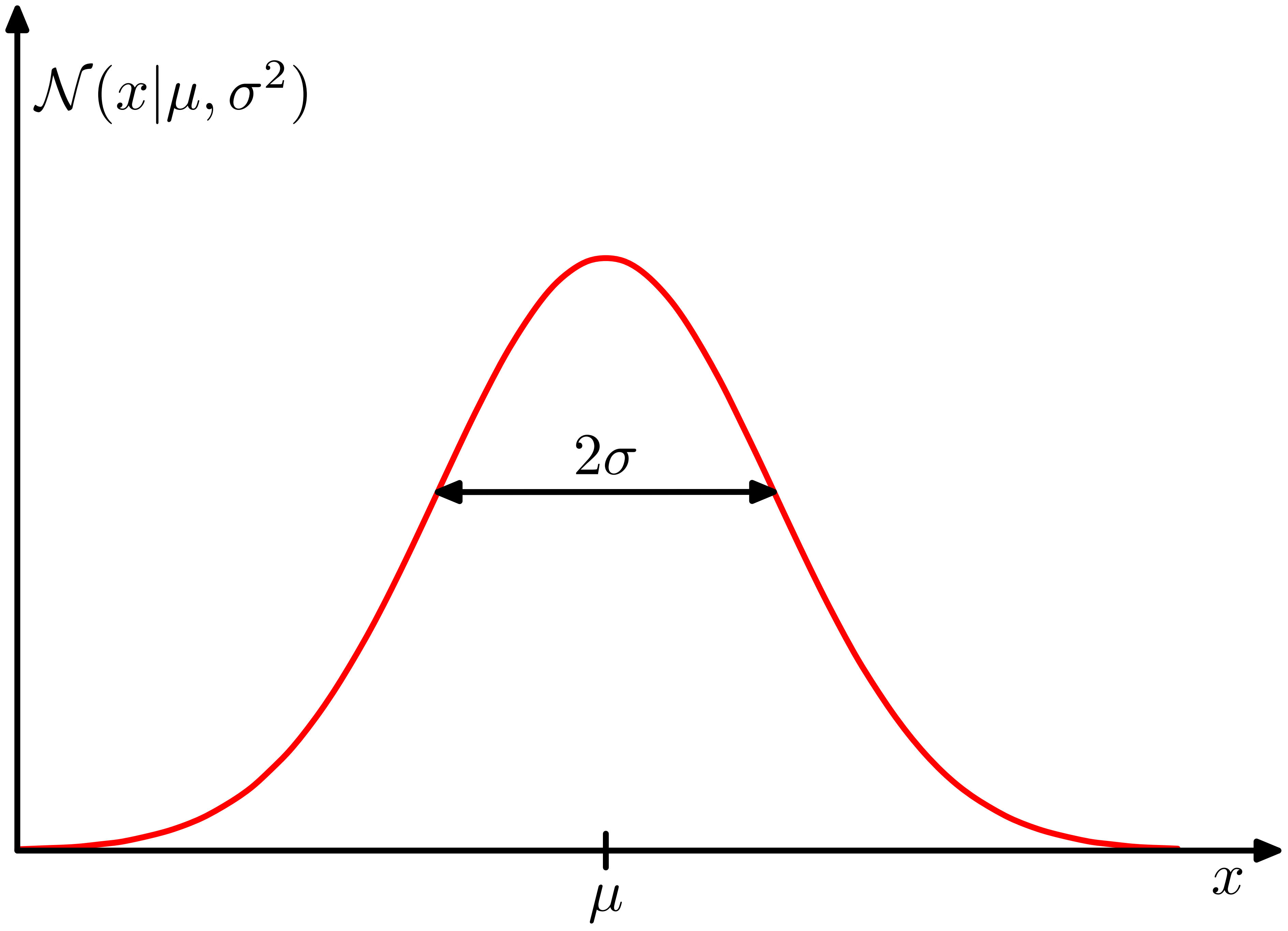
  - paramétrée par sa moyenne μ et sa variance σ^2

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

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)(x-\mu)^2 \,\mathrm{d}x = \sigma^2$$

LOI DE PROBABILITÉ GAUSSIENNE

Sujets: loi gaussienne (loi normale)



Loi de Probabilité Gaussienne

Sujets: loi gaussienne multidimensionnelle

- La version multidimensionnelle a une forme similaire

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\}$$

  - paramétrée par sa moyenne $\boldsymbol{\mu}$ et sa matrice de covariance $\boldsymbol{\Sigma}$

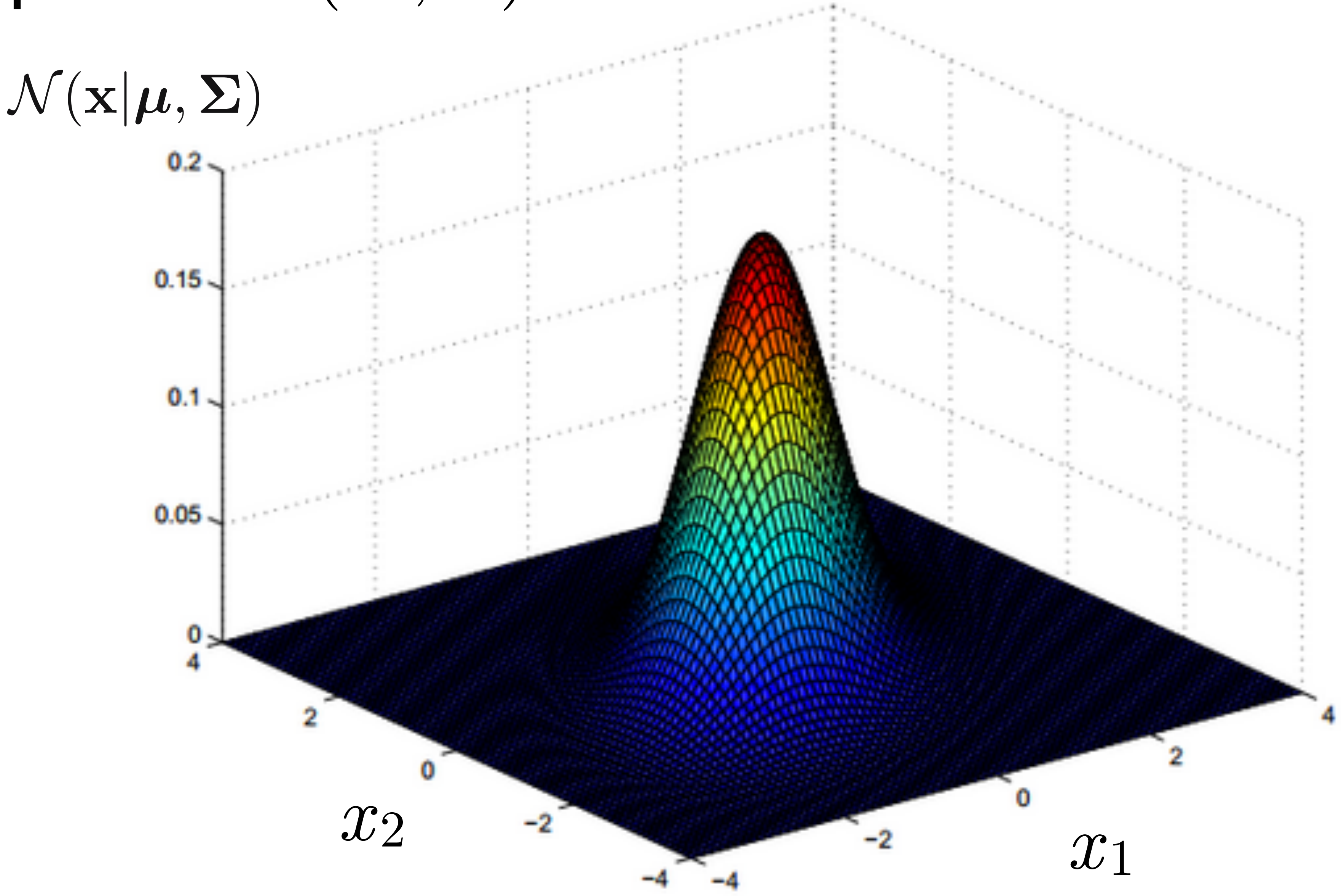
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \qquad\qquad \operatorname{cov}[\mathbf{x}] = \boldsymbol{\Sigma}$$

  - $\boldsymbol{\Sigma}$ permet de représenter des dépendances entre les élément de $\mathbf{x}$

LOI DE PROBABILITÉ GAUSSIENNE

Sujets: loi gaussienne multidimensionnelle

- Exemple : $\mathbf{x} = (x_1, x_2)$



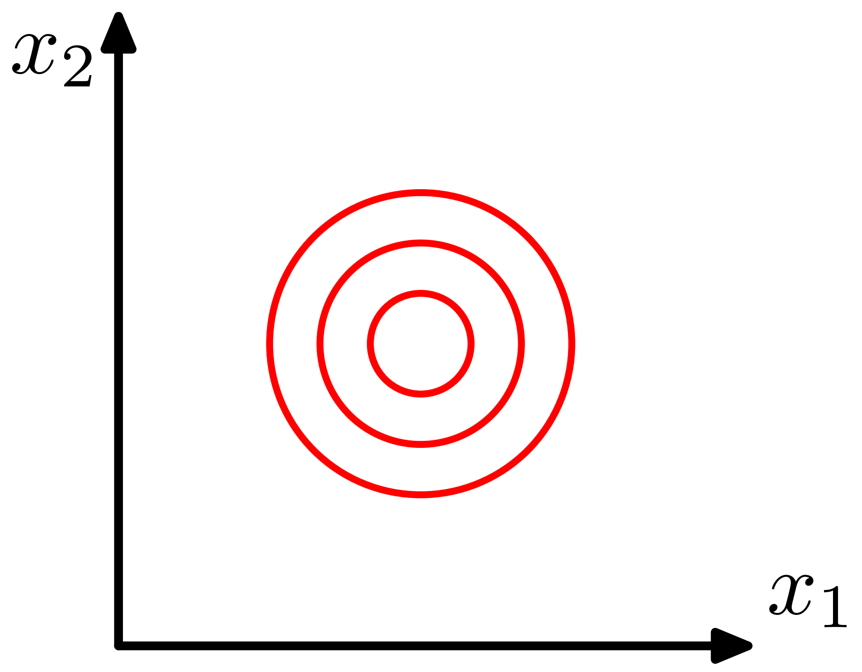
Loi de probabilité gaussienne

Sujets: loi gaussienne multidimensionnelle

- Exemple : $\mathbf{x} = (x_1, x_2)$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Contours de $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$



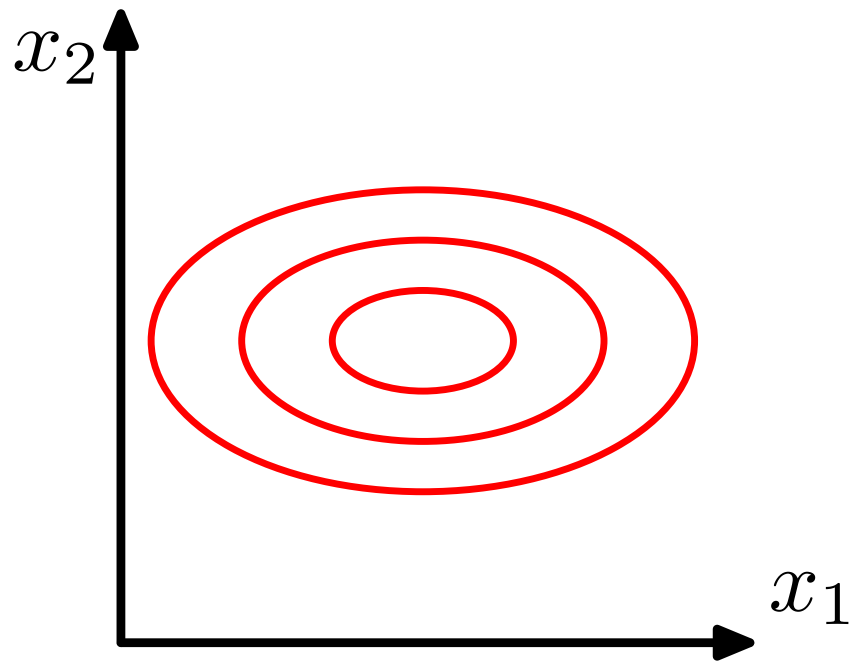
Loi de Probabilité Gaussienne

Sujets: loi gaussienne multidimensionnelle

- Exemple : $\mathbf{x} = (x_1, x_2)$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Contours de $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$



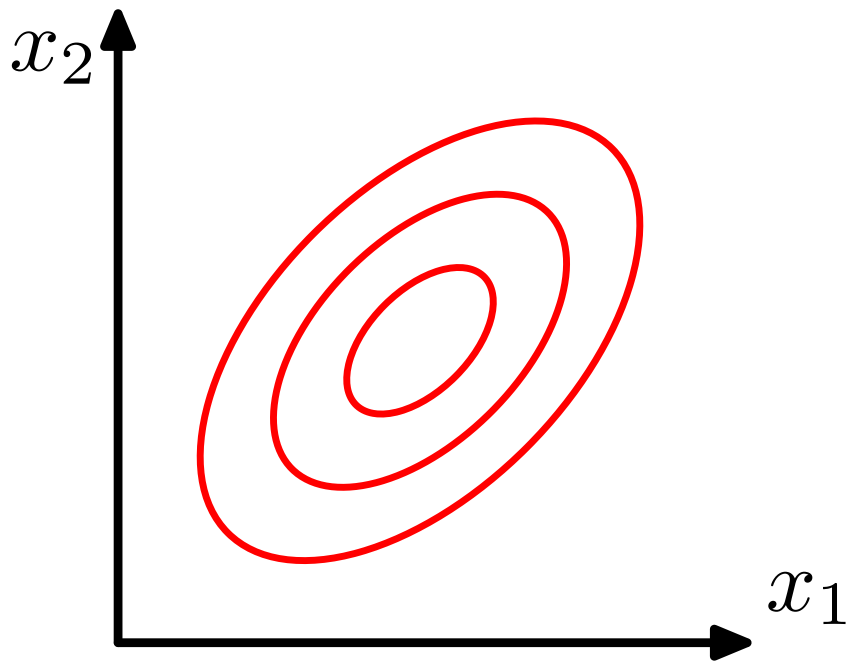
LOI DE PROBABILITÉ GAUSSIENNE

Sujets: loi gaussienne multidimensionnelle

- Exemple : $\mathbf{x} = (x_1, x_2)$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}$$

Contours de $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$



LOI DE PROBABILITÉ GAUSSIENNE

Sujets: combinaison de variables gaussiennes

- Une combinaison linéaire de variables aléatoires gaussiennes est également gaussienne

- Exemple
  - soit x une variable gaussienne de moyenne μ_1 et variance ${\sigma_1}^2$
  - soit y une variable gaussienne de moyenne μ_2 et variance ${\sigma_2}^2$
  - alors $ax + by$ suit une loi gaussienne de moyenne $a\mu_1 + b\mu_2$ et variance $a^2{\sigma_1}^2 + b^2{\sigma_2}^2$ (x et y sont indépendantes)

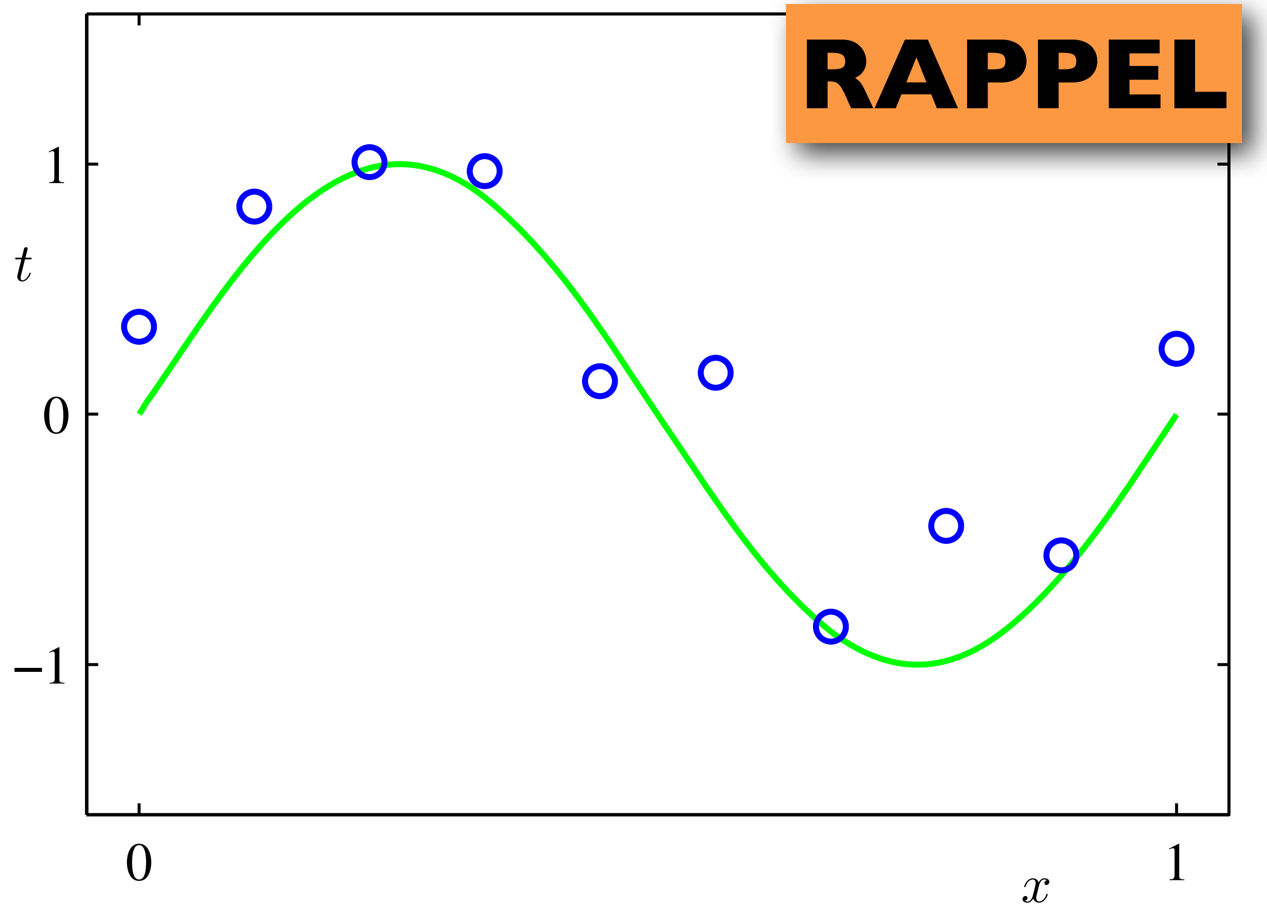
Apprentissage automatique

Formulation probabiliste - régression polynomiale revisitée

EXEMPLE: RÉGRESSION

Sujets: formulation probabiliste de la régression

- Retournons à notre exemple de régression
  - entrée : scalaire x
  - cible : scalaire t
- Données d'entraînement $\mathcal{D}$ contiennent :
  - $\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$
  - $\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$
- Objectif :
  - faire une prédiction $\hat{t}$ pour une nouvelle entrée $\hat{x}$

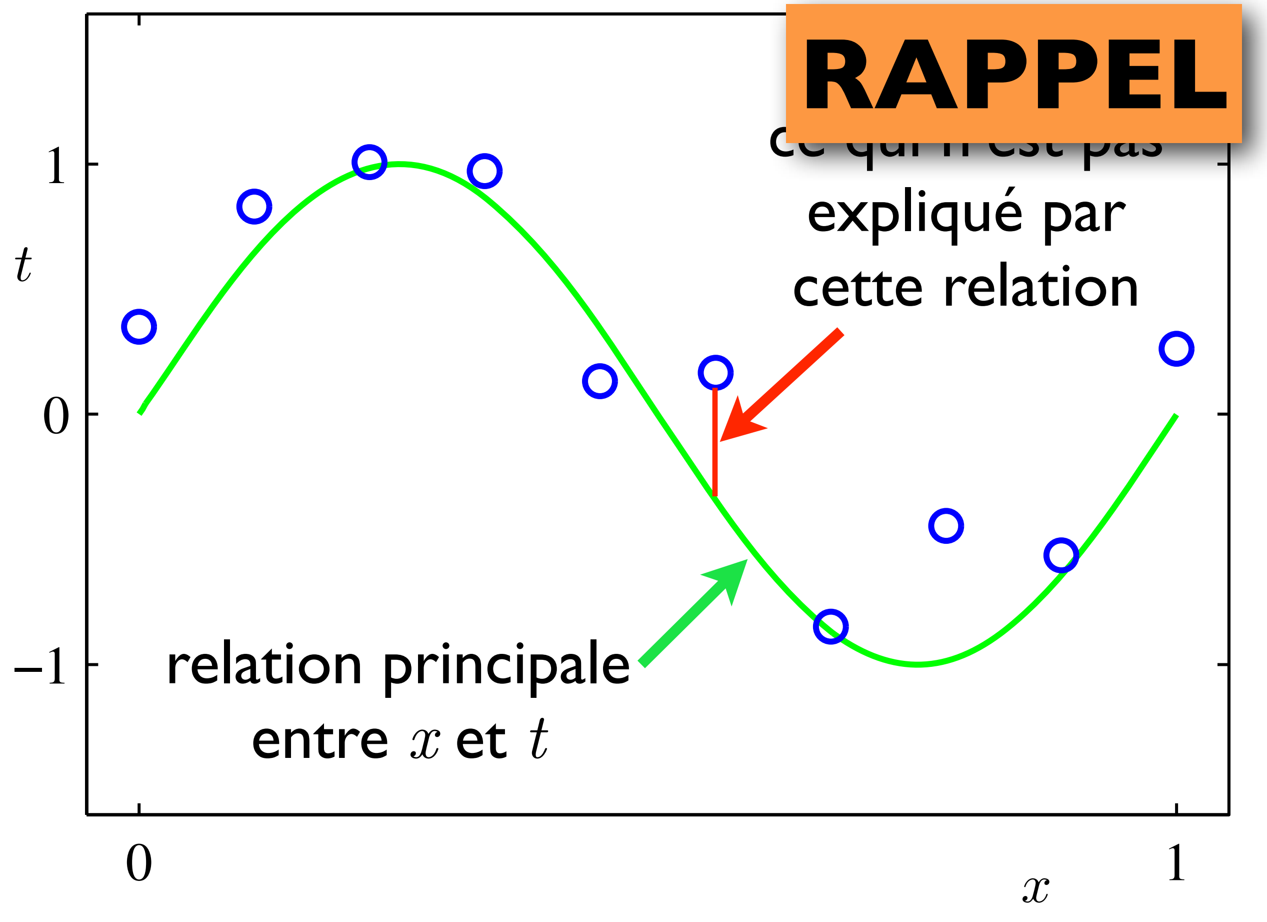


EXEMPLE: RÉGRESSION

Sujets: formulation probabiliste de la régression

- On va supposer qu'une bonne prédiction aurait une forme polynomiale

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$
$$= \sum_{j=0}^{M} w_j x^j$$



- $y(x, \mathbf{w})$ est notre **modèle**
  - représente nos hypothèses sur le problème à résoudre
  - a normalement des paramètres, qu'on doit trouver ($\mathbf{w}$ ici)

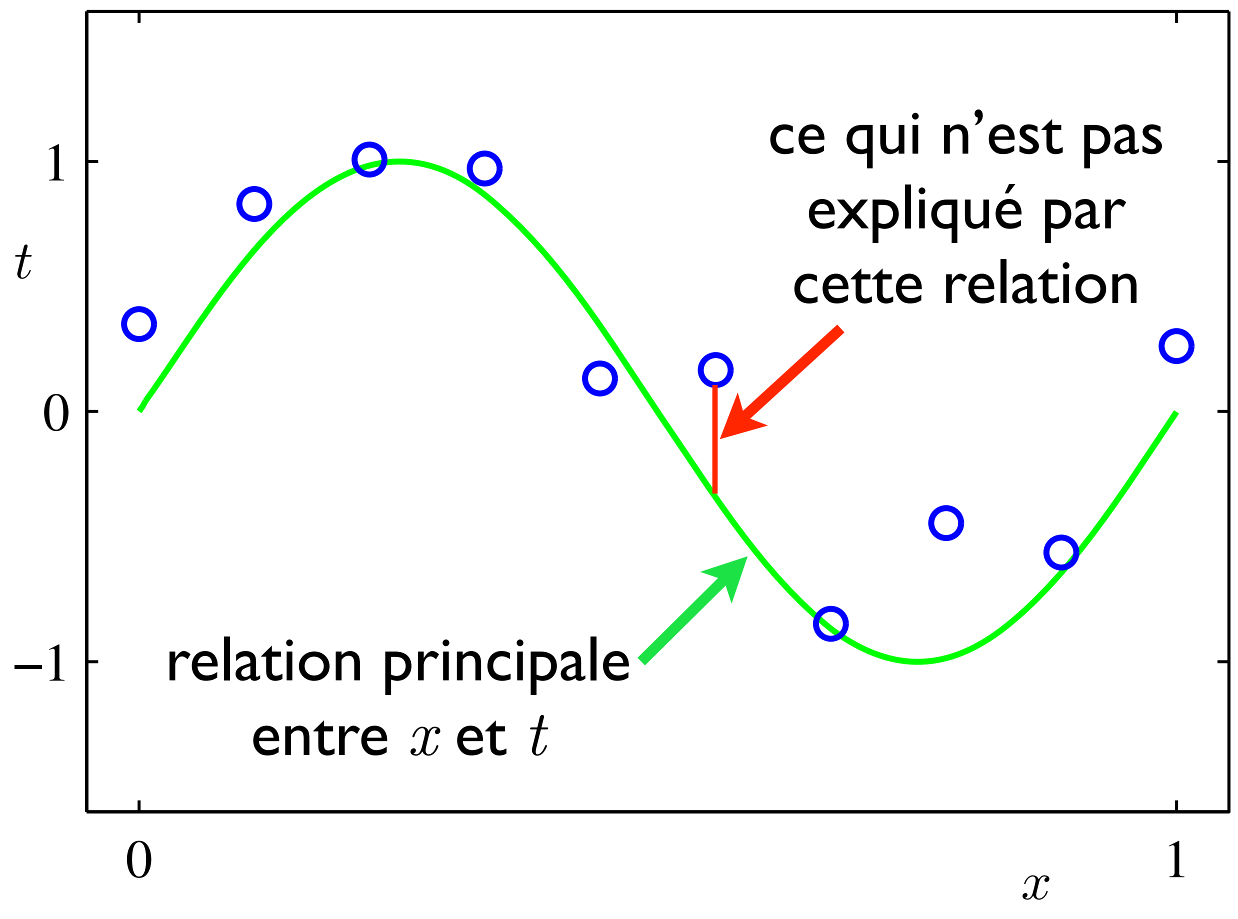
EXEMPLE: RÉGRESSION

Sujets: loi gaussienne conditionnelle

- On va formuler ce qui n'est pas expliqué par de façon probabiliste

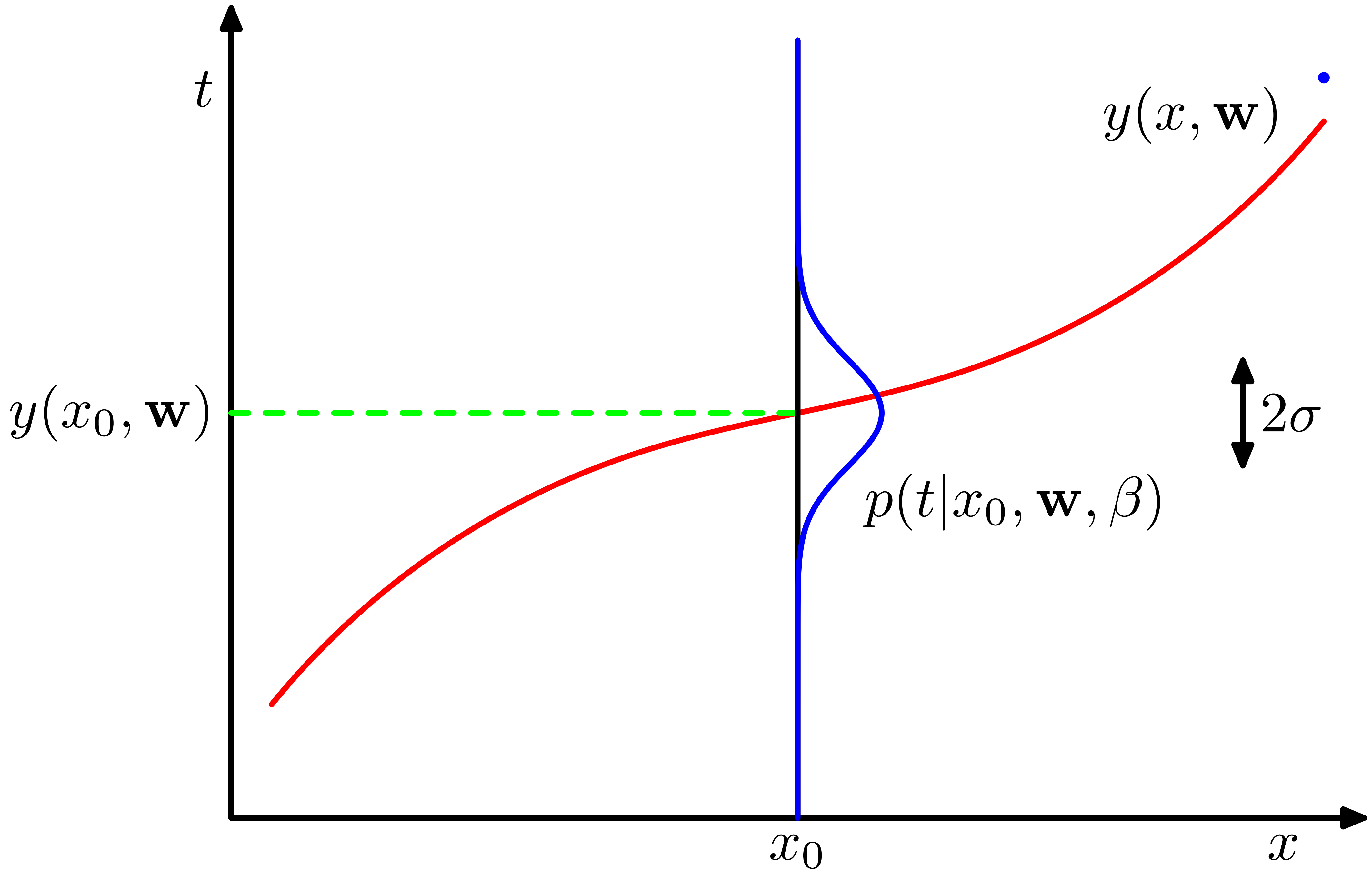
$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$

  - t a été générée selon une loi gaussienne de moyenne $y(x,\mathbf{w})$ et variance $\beta^{-1} = \sigma^2$
  - c'est une **loi gaussienne conditionnelle**



EXEMPLE: RÉGRESSION

Sujets: formulation probabiliste de la régression



Exemple: Régression

Sujets: hypothèse i.i.d.

- On va supposer que chaque cible a été générée indépendamment

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- Hypothèse de variables **indépendantes et identiquement distribuées (i.i.d.)**

Exemple: Régression

Sujets: maximum de vraisemblance

- Un bon modèle serait un modèle (w) qui associe la plus haute (log-)probabilité possible à nos données
  - on appelle ça la solution **maximum de vraisemblance**

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- Maximiser cette expression est équivalent à minimiser

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Apprentissage automatique

Formulation probabiliste - maximum a posteriori

Exemple: Régression

Sujets: maximum de vraisemblance

- Un bon modèle serait un modèle (w) qui associe la plus haute (log-)probabilité possible à nos données
  - on appelle ça la solution **maximum de vraisemblance**

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- Maximiser cette expression est équivalent à minimiser

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

EXEMPLE: RÉGRESSION

Sujets: loi a priori et loi a posteriori

- Et si on a une petite idée a priori de la solution $\mathbf{w}$
  - exemple : $\mathbf{w}$ n'est pas très loin de 0
- On peut formuler également de façon probabiliste
  - exemple : $\mathbf{w}$ suit une loi gaussienne centrée à 0

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

à quel point $\mathbf{w}$ s'éloigne de 0

$$= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

Exemple: Régression

Sujets: loi a priori et loi a posteriori

- $p(\mathbf{w}|\alpha)$ exprime notre croyance a priori sur la valeur de $\mathbf{w}$
  - c'est une **loi a priori** (*prior*)
- Lorsqu'on observe des données, on peut mettre à jour notre croyance

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

  - c'est la **loi a posteriori** (*posterior*)

Exemple: Régression

Sujets: maximum a posteriori

- On pourrait choisir le modèle $\mathbf{w}$ qui est le plus (log-)probable selon nos croyances a posteriori $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$
  - on appelle ça la solution **maximum a posteriori**

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

- Équivalent à la perte régularisée si $\lambda = \alpha/\beta$

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Apprentissage automatique

Formulation probabiliste - théorie de l'information

Théorie de l'Information

Sujets: théorie de l'information

- Les probabilités sont également utiles pour quantifier l'information présente dans des données
  - exemple : quel est le nombre minimum de bits nécessaire pour encoder un message ?
- Cette question est intimement liée à la probabilité d'observer ce message
  - plus le message est «surprenant» (improbable), plus on aura besoin de bits

THÉORIE DE L'INFORMATION

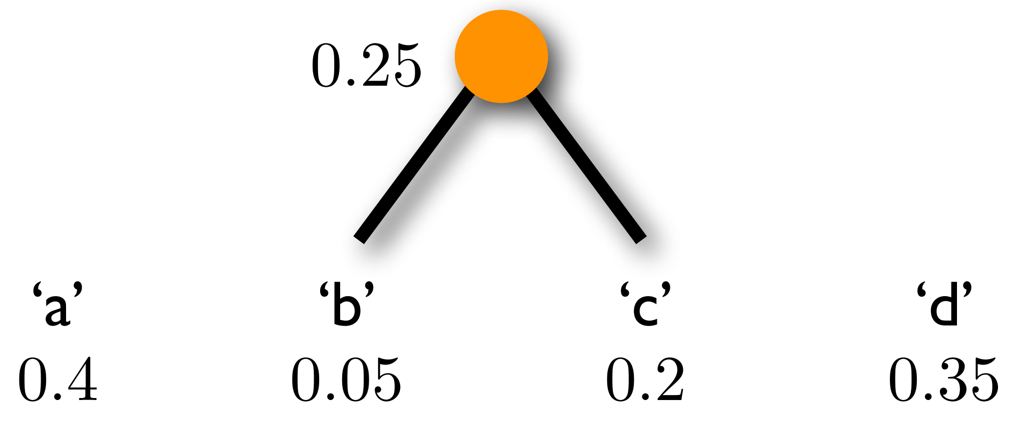
Sujets: codage de Huffman

- Codage de Huffman :
  - façon optimale d'encoder des symboles indépendants de façon binaire
  - plus un symbole est «fréquent» (probable), plus son code sera court

'a'	'b'	'c'	'd'
0.4	0.05	0.2	0.35

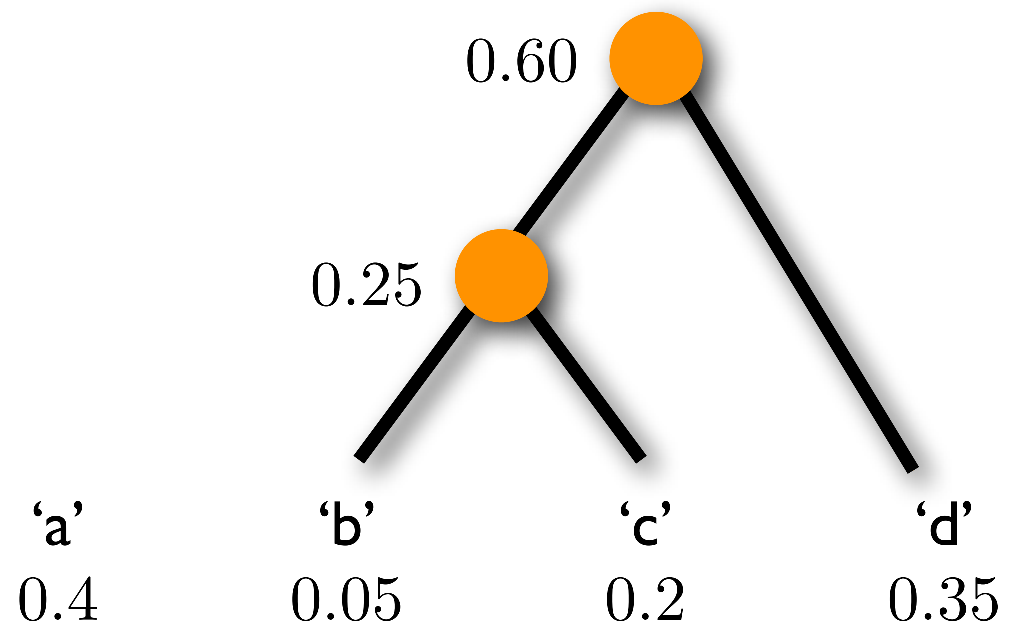
Sujets: codage de Huffman

- Codage de Huffman :
  - façon optimale d'encoder des symboles indépendants de façon binaire
  - plus un symbole est «fréquent» (probable), plus son code sera court



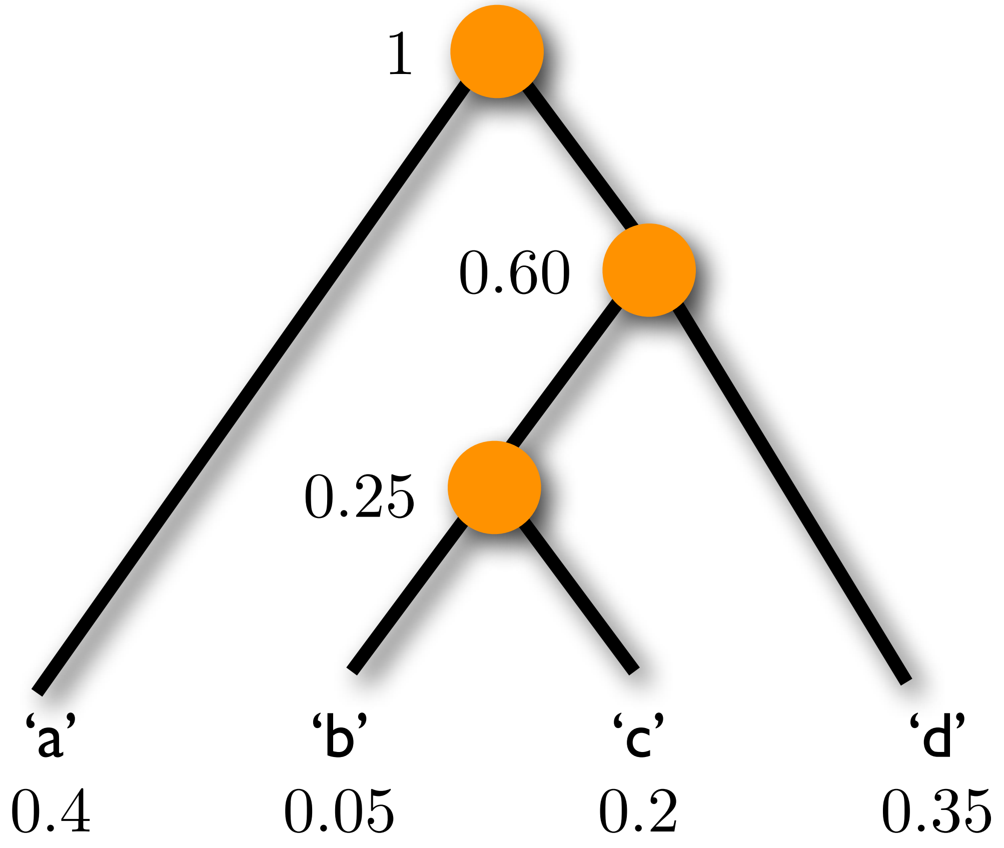
Sujets: codage de Huffman

- Codage de Huffman :
  - façon optimale d'encoder des symboles indépendants de façon binaire
  - plus un symbole est «fréquent» (probable), plus son code sera court



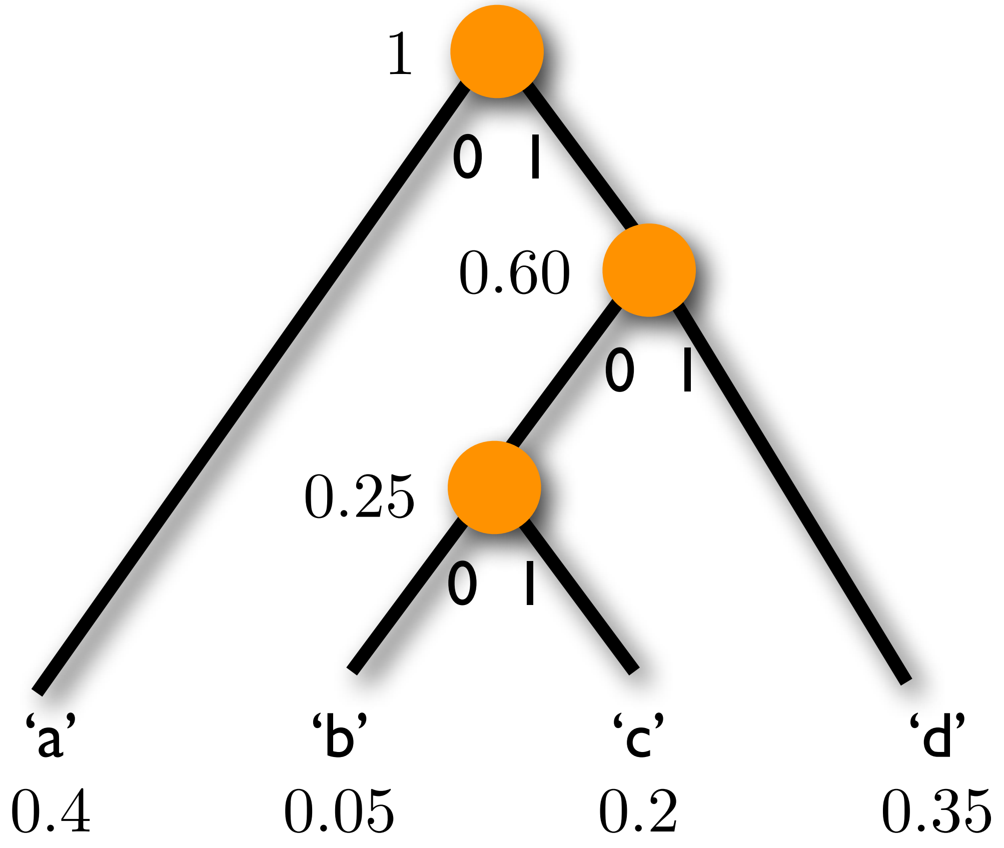
Sujets: codage de Huffman

- Codage de Huffman :
  - façon optimale d'encoder des symboles indépendants de façon binaire
  - plus un symbole est «fréquent» (probable), plus son code sera court



Sujets: codage de Huffman

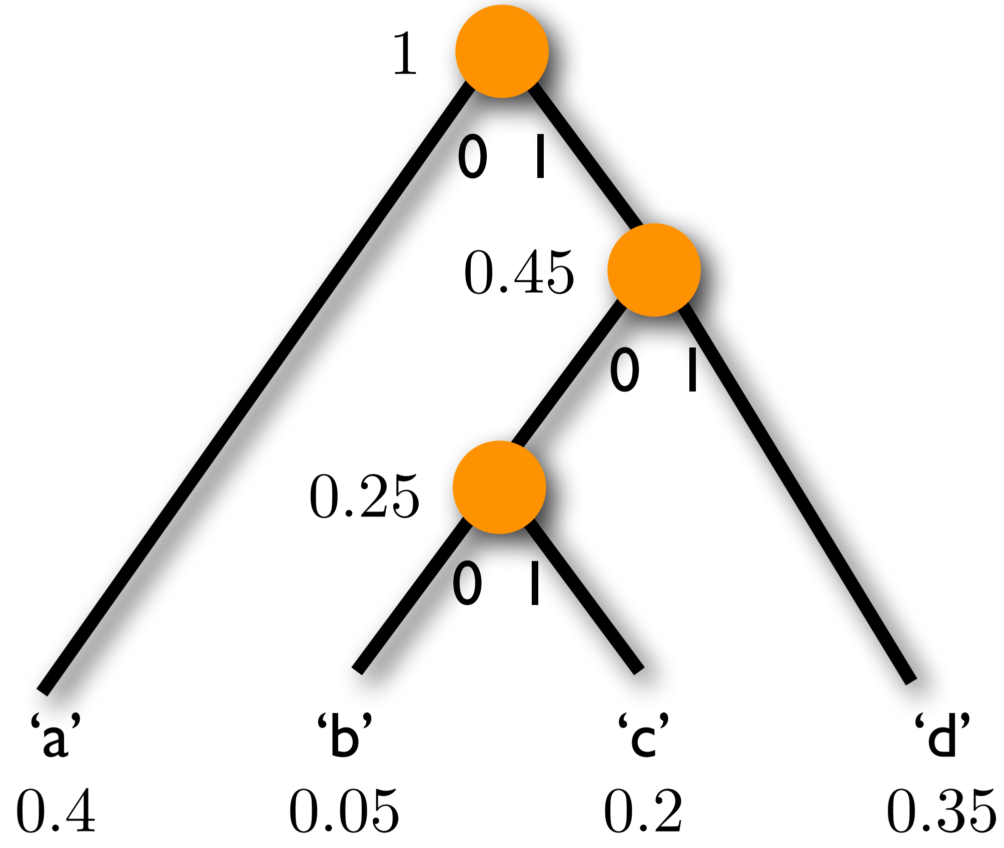
- Codage de Huffman :
  - façon optimale d'encoder des symboles indépendants de façon binaire
  - plus un symbole est «fréquent» (probable), plus son code sera court



Symbole	Code
'a'	0
'b'	100
'c'	101
'd'	11

THÉORIE DE L'INFORMATION

Sujets: entropie, information



- Soit $p(x)$ la probabilité d'observer le symbole x
  - la taille moyenne du code d'un symbole est

    $$0.4 \times 1 + 0.05 \times 3 + 0.2 \times 3 + 0.35 \times 2 = 1.85 \text{ (bits)}$$

- **Entropie** :

  $$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x) \approx 1.7393$$

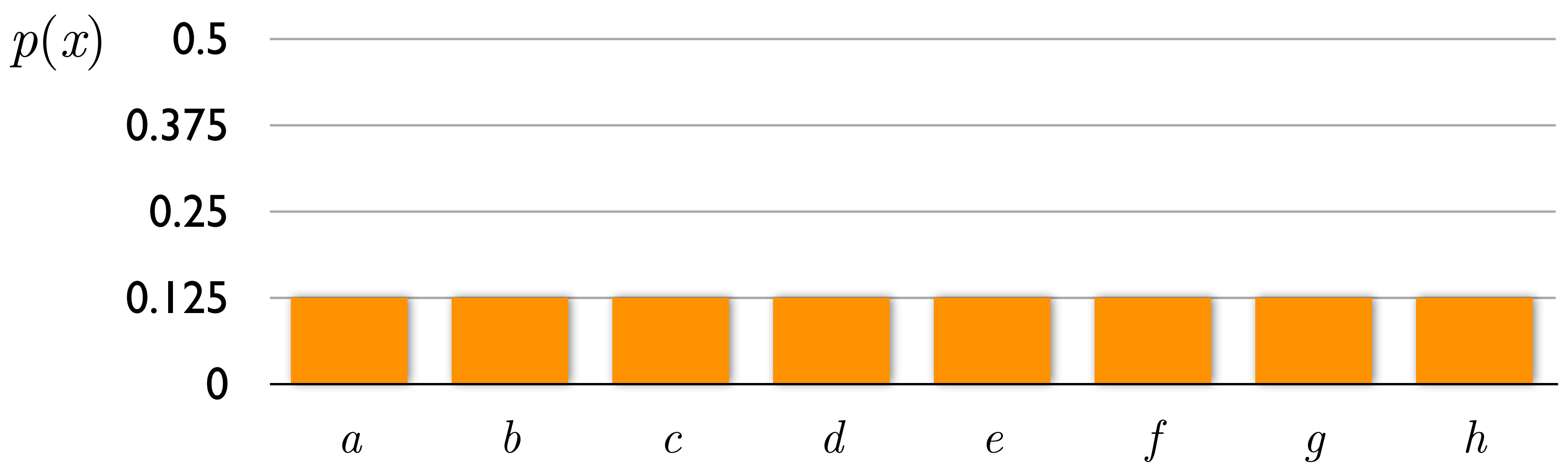
  - Claude Shannon a démontré qu'il est impossible de compresser l'information dans un plus petit code moyen
  - $-\log_2 p(x)$ est l'information contenue par x

HUGO LAROCHELLE

THÉORIE DE L'INFORMATION

Sujets: entropie

- L'entropie donne une façon standard de quantifier l'information moyenne contenue par une observation x
  - plus $p(x)$ est proche d'une loi uniforme, plus l'entropie est élevée
  - exemple : $x \in \{a,b,c,d,e,f,g,h\}$

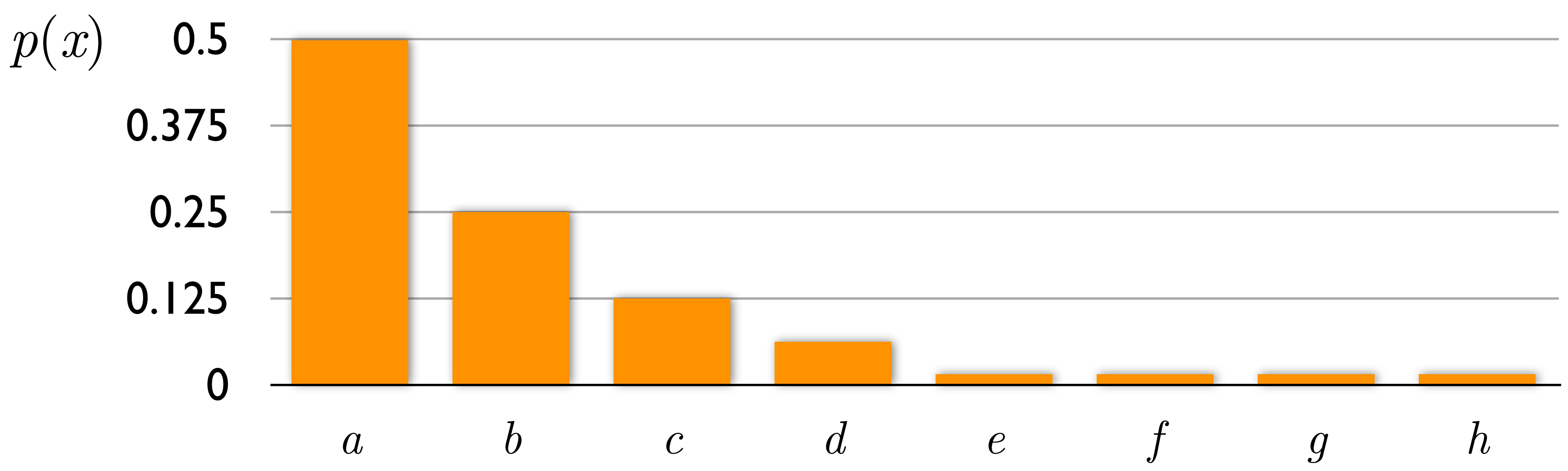


$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits}$$

THÉORIE DE L'INFORMATION

Sujets: entropie

- L'entropie donne une façon standard de quantifier l'information moyenne contenue par une observation x
  - plus $p(x)$ est proche d'une loi uniforme, plus l'entropie est élevée
  - exemple : $x \in \{a,b,c,d,e,f,g,h\}$



$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2 \text{ bits}$$

THÉORIE DE L'INFORMATION

Sujets: entropie

- L'entropie est une fonction d'une loi de probabilité
  - elle reflète l'incertitude représentée par la loi
  - si $p(x) = 1$ pour une seule valeur de x, l'entropie est 0

- On peut généraliser l'entropie à une loi jointe sur plusieurs variables

$$\mathrm{H}[x, y] = -\sum_x \sum_y p(x, y) \log_2 p(x, y)$$

Théorie de l'Information

Sujets: entropie conditionnelle

- L'entropie conditionnelle quantifie l'information **additionnelle** qu'apporte une nouvelle observation y

$$\mathrm{H}[y|x] = -\sum_{x}\sum_{y} p(x,y)\log_2 p(y|x)$$

- On peut démontrer que

$$\mathrm{H}[x,y] = \mathrm{H}[y|x] + \mathrm{H}[x]$$

Théorie de l'Information

Sujets: entropie différentielle

- On peut généraliser l'entropie au cas continu :

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

  - l'**entropie différentielle** peut être négative
  - lorsqu'on utilise le logarithme naturel, on parle de «nats» à la place de bits
  - l'interprétation comme mesure de l'incertitude associée à une loi de probabilité demeure pertinente

Sujets: entropie différentielle

- Plus la fonction de densité est «piquée», plus l'entropie sera basse
  - si $x \in [a,b]$, la loi uniforme a l'entropie maximale
  - si $x \in \mathbb{R}$ avec moyenne μ et variance σ^2, la loi gaussienne a l'entropie maximale

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\}$$

Apprentissage automatique

Formulation probabiliste - divergence de Kullback-Leibler

THÉORIE DE L'INFORMATION

Sujets: divergence de Kullback-Leibler ou entropie relative

- Si on ne connait pas $p(x)$, on va vouloir l'estimer

- Si $q(x)$ est notre estimation, on définit la **divergence de Kullback-Leibler** (K-L) comme suit :

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

  - dans le cas discret (avec des sommes), correspond au nombre de bits additionnels par rapport à ce qui serait optimal

Théorie de l'Information

Sujets: information mutuelle

- Utilisée comme «distance» entre deux lois
  - est toujours positive
  - n'est pas symétrique (contrairement à une vraie distance)
- Peut mesurer à quel point deux variables sont dépendantes

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] & \equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x}) p(\mathbf{y})) \\ & = -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned}$$

  - on appelle cette mesure l'**information mutuelle**