Hiver 2019

Méthodes d'apprentissage

IFT603

Formulation probabiliste

Par
Pierre-Marc Jodoin
et
Hugo Larochelle

1

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

2

Variable aléatoire

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

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

$$P(X = x_i, Y = y_i)$$

et se lit comme la « probabilité d'observer à **<u>la fois</u>** x_i **<u>et</u>** y_j .

- Note:

$$P(X = x_i, Y = y_i) = P(Y = y_i, X = x_i)$$

3

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 $X = x_i$

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

4

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

$$P(X = x_i \mid Y = y_j)$$

Se lit : la probabilité que X= x_j étant donné que $Y= y_i$

https://www.npr.org/2016/11/14/501737150/rural-voters-played-a-big-part-in-helping-trump-defeat-clinton

5

Probabilité conditionnelle

Exemple, élections américaines 2016

$$P(Voter\ républicain) = 46.1\%$$

VS

$$\begin{cases} P(Voter\ républicain \mid Zone\ urbaine) = 35\% \\ P(Voter\ républicain \mid Zone\ rurale) = 62\% \\ P(Voter\ républicain \mid Banlieu) = 50\% \end{cases}$$

https://www.npr.org/2016/11/14/501737150/rural-voters-played-a-big-part-in-helping-trump-defeat-clinton

6

Produit des probabilités

x_i et y_j ont disparu, seulement pour simplifier la notation

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

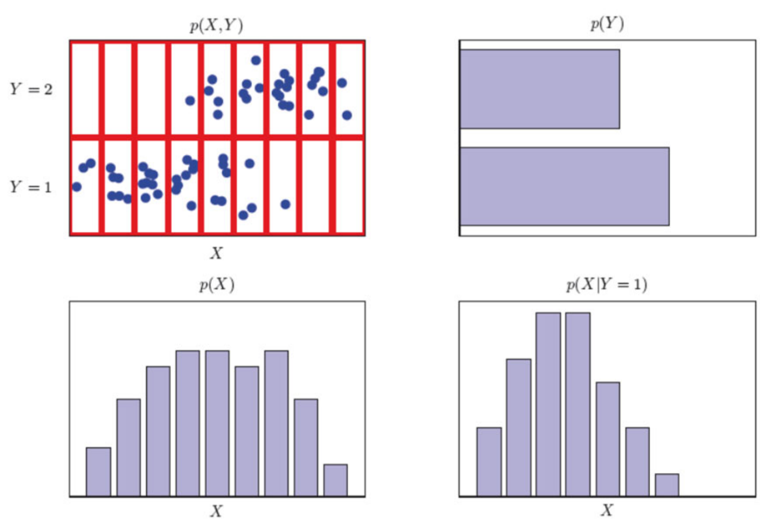
$$P(X,Y) = P(X \mid Y)P(Y)$$

En mots :

la probabilité d'observer $X = x_i$ **ET** $Y = y_j$, c'est la probabilité d'observer $Y = y_i$ multipliée par la probabilité d'observer $X = x_i$ **étant donné que** $Y = y_i$

7

Probabilités jointes, marginale et conditionnelles



Crédit : Bishop

8

Bayes

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

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

$p(Y)$ est appelée loi de probabilité ***a priori*** (*prior*)

$p(Y \mid X)$ est appelée loi de probabilité ***a posteriori*** (*posterior*)

9

Indépendance

Deux variables aléatoires X et Y sont indépendantes si

- $P(X, Y) = P(X)P(Y)$ ou
- $P(X \mid Y) = P(X)$ ou
- $P(Y \mid X) = P(Y)$

- Observer la valeur d'une variable ne nous apprend rien sur la valeur de l'autre

10

Variable aléatoire continue

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

$$p\big(x \in (a,b)\big) = \int_a^b p(x)dx$$

11

Variables aléatoires continues

Soit X une **variable aléatoire continue**

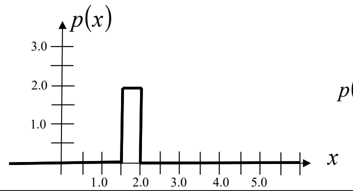
- la fonction de densité doit satisfaire

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

à noter que, contrairement aux probabilités d'une variable discrète, la fonction de densité peut être > 1.

Exemple



$$p(x) = \begin{cases} 2 & \text{is } x \in [1.5, 2.0] \\ 0 & \text{sinon} \end{cases}$$

12

Variables aléatoires continues

Soit X une **variable aléatoire continue**

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

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

13

Variables aléatoires continues

Soient X et Y deux **variables aléatoires continues**

- elles sont associées à une **fonction de densité jointe** $p(x,y)$ telle que :

$$p(x \in [a,b], y \in [c,d]) = \int_a^b \int_c^d p(x,y)dxdy$$

14

Variables aléatoires continues

Soient X et Y deux **variables aléatoires continues**

- La **function de densité marginale** s'obtient en intégrant l'autre variable

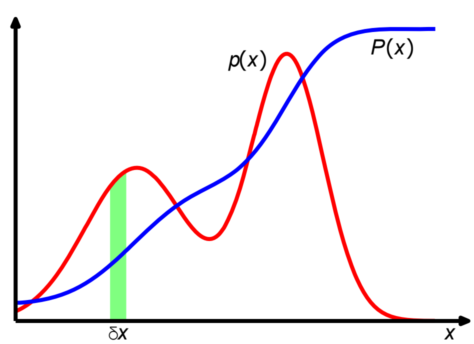
$$p(x) = \int p(x,y)dy$$

- La **function de densité conditionnelle** s'obtient comme auparavant

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

15

Variables aléatoires continues



16

Expérance mathématique

L'**espérance** d'une **variable *X*** est la moyenne qu'on obtient si on répète un grand nombre de fois une expérience

$$\mathrm{E}[X] = \sum_x x p(x) \quad \text{(cas discret)}$$

$$\mathrm{E}[X] = \int x p(x) dx \quad \text{(cas continu)}$$

17

Expérance mathématique

L'**espérance** d'une **fonction *f(x)*** est la moyenne qu'on obtient si on génère un grand nombre de valeurs pour cette fonction

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

$$\mathrm{E}[f] = \int f(x) p(x) dx \quad \text{(cas continu)}$$

18

Variance

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

$$\operatorname{var}[X] = \mathrm{E}\left[\left(X - \mathrm{E}[X]\right)^2\right]$$

- La **variance** d'une **fonction *f (x)*** est

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

La variance mesure à quel point les valeurs varient autour de l'espérance

19

Propriétés de l'espérance et de la variance

Transformation linéaire de l'espérance

$$\mathrm{E}[ax+by] = \sum_x \sum_y (ax+by)\,p(x,y) \qquad \text{a,b sont réels}$$

$$= a\mathrm{E}[x] + b\mathrm{E}[y] \qquad \text{Si x, y indépendants}$$

Transformation linéaire de la variance

$$\operatorname{var}[ax+by] = a^2 \operatorname{var}[x] + b^2 \operatorname{var}[y] \qquad \text{Si x, y indépendants}$$

20

Espérance et variance conditionnelles

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

$$\mathrm{E}[x \mid y] = \sum_x x p(x \mid y)$$

$$\mathrm{E}[f(x) \mid y] = \sum_x f(x) p(x \mid y)$$

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

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

21

Covariance

La covariance entre 2 variables aléatoires X et Y

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

mesure à quel point on peut prédire X à partir de Y (linéairement), et vice-versa

si X etY sont indépendantes, alors la covariance est 0

22

Variables aléatoires multidimensionnels

Une variable aléatoire peut être un vecteur

L'espérance d'un vecteur est le vecteur des espérances

$$\mathrm{E}[\vec{x}] = \left(\mathrm{E}[x_1], \ldots, \mathrm{E}[x_D]\right)^{\mathrm{T}}$$

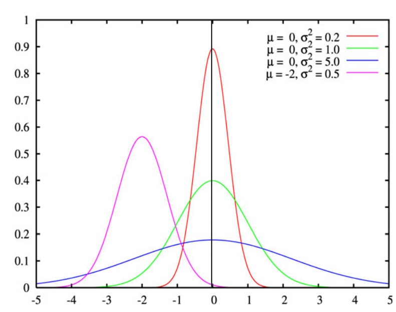
Et la covariance de deux vecteurs est

$$\mathrm{cov}[\vec{x}, \vec{y}] = \mathrm{E}_{\vec{x}\vec{y}}\left[\vec{x}\vec{y}^{\mathrm{T}}\right] - \mathrm{E}_{\vec{x}}[\vec{x}]\mathrm{E}_{\vec{y}}[\vec{y}]$$

23

Loi de probabilité gaussienne

$$N(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Moyenne : $\mathrm{E}[x] = \int_{-\infty}^{\infty} \mathrm{N}(x; \mu, \sigma) x dx = \mu$

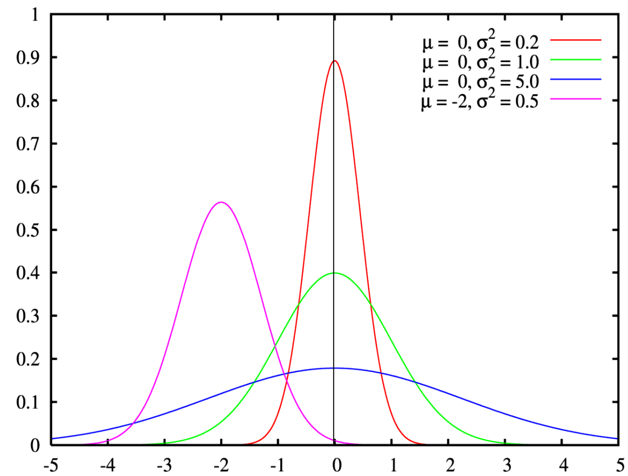
Variance : $\mathrm{var}[x] = \int_{-\infty}^{\infty} \mathrm{N}(x; \mu, \sigma)(x - \mu)^2 dx = \sigma^2$

Écart type : $\sqrt{\mathrm{var}[x]} = \sigma$

24

Loi de probabilité gaussienne

σ	68.26894921371%
2σ	95.44997361036%
3σ	99.73002039367%



25

Gaussienne **multivariée**

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

Moyenne : $\mathrm{E}[\vec{x}] = \vec{\mu}$

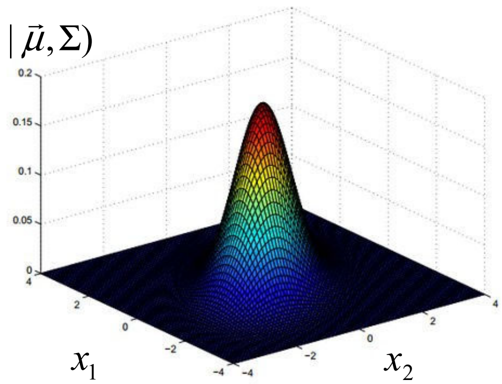
Variance : $\mathrm{cov}[\vec{x}] = \Sigma$

26

Gaussienne **multivariée**

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

$N(\vec{x} \mid \vec{\mu}, \Sigma)$



x_1 x_2

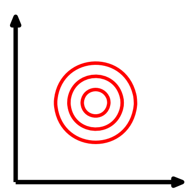
27

Gaussienne **multivariée**

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

Courbes de niveaux de $N(\vec{x} \mid \vec{\mu}, \Sigma)$

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



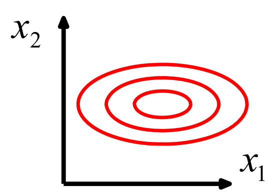
28

Gaussienne **multivariée**

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

Courbes de niveaux de $N(\vec{x} \mid \vec{\mu}, \Sigma)$

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



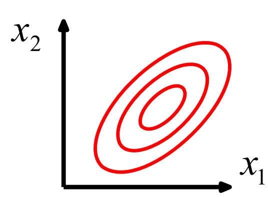
29

Gaussienne **multivariée**

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

Courbes de niveaux de $N(\vec{x} \mid \vec{\mu}, \Sigma)$

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



30

Gaussienne **multivariée**

Une **combinaison linéaire** de variables aléatoires gaussiennes est également gaussienne

- Exemple
  - soit x une variable gaussienne de moyenne μ_1 et variance σ_1^2
  - soit y une variable gaussienne de moyenne μ_2 et variance σ_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)

31

Régression:

Maximum de vraisemblance

vs

Maximum a posteriori

32

32

Introduction au tableau

Maximum de vraisemblance

vs

Maximum a posteriori

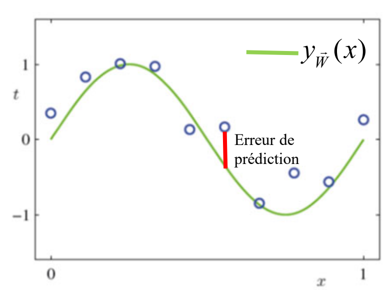
33

33

Régression 1D

Retournons à notre exemple de régression

- **Données**
  - entrée : scalaire x
  - cible : scalaire t
- **Ensemble d'entraînement** D contient:
  - $X = (x_1, \ldots, x_N)^T$
  - $T = (t_1, \ldots, t_N)^T$



- **Objectif :**
  - Faire une prédiction $\hat{t}$ pour chaque nouvelle entrée $\hat{x}$

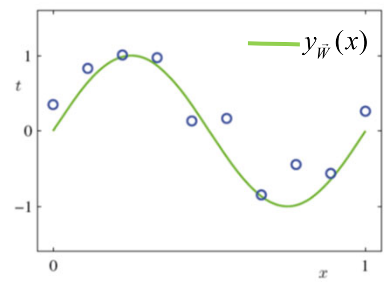
34

34

Régression polynomiale

➢ On va supposer que nos données suivent une **forme polynomiale**

$$y_{\vec{W}}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$
$$= \sum_{i=0}^{M} w_i x^i$$



➢ $y_{\vec{W}}(x)$ est notre **modèle**

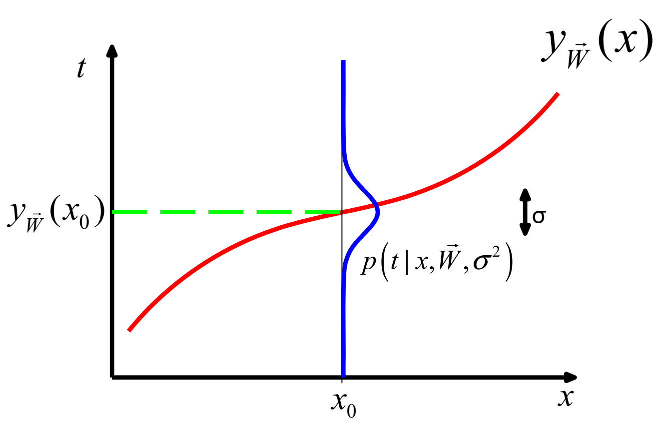
- ✓ Représente nos hypothèses sur le problème à résoudre
- ✓ Un modèle a toujours des paramètres qu'on doit trouver (ici $\vec{W}$)

35

35

Régression polynomiale

Loi gaussienne conditionnelle

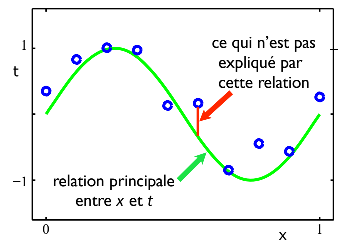


36

Régression polynomiale

Loi gaussienne conditionnelle

On va formuler l'hypothèse que chaque donnée t a été générée selon une gaussienne de moyenne $y_{\vec{W}}(x)$ et de variance σ^2



$$p\left(T \mid X, \vec{W}, \sigma^2\right) = \prod_{n=1}^{N} p\left(t_n \mid x_n, \vec{W}, \sigma^2\right)$$

$$= \prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), \sigma^2\right)$$

Variables indépendantes et identiquement distribuées (i.i.d)

37

Régression polynomiale

Maximum de vraisemblance
(cf. notes au tableau)

Cherche les meilleurs paramètres $\vec{W}$ en **maximisant la vraisemblance**

Connue **Inconnue**

$$W = \arg\max_W p\left(T \mid X, \vec{W}, \sigma^2\right)$$

$$= \arg\max_W \prod_{n=1}^{N} N\left(t_n \mid x_n, \vec{W}, \sigma^2\right) \quad \text{=> Données gaussienns et (i,i.d)}$$

$$= \arg\max_W \ln\left[\prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), \sigma^2\right)\right]$$

$$= \arg\max_W \sum_{n=1}^{N} \ln N\left(t_n \mid y_W(x_n), \sigma^2\right)$$

38

Régression polynomiale

Maximum de vraisemblance
(cf. notes au tableau)

Cherche les meilleurs paramètres W en **maximisant la vraisemblance**

$$W = \arg\max_W \sum_{n=1}^{N} \ln N\left(t_n \mid y_W(x_n), \sigma^2\right)$$

$$= \arg\max_W \sum_{n=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(t_n - y_W(x_n))^2}{2\sigma^2}}\right]$$

Indépendant de W

$$= \arg\max_W \sum_{n=1}^{N} -\frac{(t_n - y_W(x_n))^2}{2\cancel{\sigma^2}} + \cancel{\ln\left[\frac{1}{\sqrt{2\pi\sigma}}\right]}$$

$$= \arg\min_W \frac{1}{2}\sum_{n=1}^{N} (t_n - y_W(x_n))^2$$

39

Régression polynomiale

Maximum de vraisemblance

Cherche les meilleurs paramètres W en **maximisant la vraisemblance**

$$\boxed{W = \arg\min_W \frac{1}{2}\sum_{n=1}^{N} (t_n - y_W(x_n))^2}$$

40

Régression polynomiale

Maximum *a posteriori* (*MAP*)
(cf. notes au tableau)

Cherche les meilleurs paramètres W en **maximisant la probabilité *a posteriori***

Inconnue

Connues

$$W = \arg\max_W p\left(W \mid X, T, \sigma^2\right)$$

$$= \arg\max_W \frac{p\left(T \mid X, W, \sigma^2\right) p(W)}{P\left(X, T, \sigma^2\right)}$$ => Par le théorème de Bayes

Constante par rapport à W

$$= \arg\max_W p\left(T \mid X, W, \sigma^2\right) p(W)$$

$$= \arg\max_W \prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), W, \sigma^2\right) p(W)$$ => Données gaussiennes et (i,i.d)

41

Régression polynomiale

Maximum *a posteriori* (*MAP*)
(cf. notes au tableau)

Cherche les meilleurs paramètres W en **maximisant la probabilité *a posteriori***

$$W = \arg\max_W \prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), \sigma^2\right) p(W)$$

Si on présuppose que les paramètres W suivent une distribution gaussienne centrée à zéro avec une matrice de variance Σ

$$W = \arg\max_W \prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), \sigma^2\right) N(W \mid 0, \Sigma)$$

$$W = \arg\max_W \ln\left[\prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), \sigma^2\right) N(W \mid 0, \Sigma)\right]$$

42

Régression polynomiale

Maximum *a posteriori* (*MAP*)
(cf. notes au tableau)

Cherche les meilleurs paramètres W en **maximisant la probabilité *a posteriori***

$$W = \arg\max_W \ln\left[\prod_{n=1}^{N} N\left(t_n \mid y_W(x_n), \sigma^2\right) N(W \mid 0, \Sigma)\right]$$

$$= \arg\max_W \sum_{n=1}^{N} \ln\left[N\left(t_n \mid y_W(x_n), \sigma^2\right) N(W \mid 0, \Sigma)\right]$$

$$= \arg\max_W \sum_{n=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(t_n - y_W(x_n))^2}{2\sigma^2}}\right] + \ln\left[\frac{1}{(2\pi)^{1/M}|\Sigma|} e^{-\frac{W^{\mathrm{T}}\Sigma^{-1}W}{2}}\right]$$

$$= \arg\max_W \sum_{n=1}^{N} -\frac{(t_n - y_W(x_n))^2}{2\sigma^2} - \frac{W^{\mathrm{T}}\Sigma^{-1}W}{2} + \ln\left[\frac{1}{\sqrt{2\pi\sigma}}\right] + \ln\left[\frac{1}{(2\pi)^{1/M}|\Sigma|}\right]$$

43

Régression polynomiale

Maximum *a posteriori* (*MAP*)
(cf. notes au tableau)

Cherche les meilleurs paramètres W en **maximisant la probabilité *a posteriori***

$$W = \arg\max_W \sum_{n=1}^{N} -\frac{(t_n - y_W(x_n))^2}{2\sigma^2} - \frac{W^{\mathrm{T}}\Sigma^{-1}W}{2} + \cancel{\ln\left[\frac{1}{\sqrt{2\pi\sigma}}\right]} + \cancel{\ln\left[\frac{1}{(2\pi)^{1/M}|\Sigma|}\right]}$$

Constante par rapport à W

De plus, comme on ne connaît généralement pas Σ, on suppose qu'elle est isotropique

$$\Sigma = \begin{pmatrix} \alpha^2 & 0 & \cdots & 0 \\ 0 & \alpha^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \alpha^2 \end{pmatrix} = \alpha^2 I$$

44

Régression polynomiale

Maximum *a posteriori* (*MAP*)
(cf. notes au tableau)

Cherche les meilleurs paramètres W en **maximisant la probabilité *a posteriori***

$$W = \arg\max_W \sum_{n=1}^{N} -\frac{(t_n - y_W(x_n))^2}{2\sigma^2} - \frac{W^T \Sigma^{-1} W}{2}$$

$$= \arg\max_W \sum_{n=1}^{N} -\frac{(t_n - y_W(x_n))^2}{2\sigma^2} - \frac{W^T W}{2\alpha^2}$$

$$= \arg\min_W \sum_{n=1}^{N} \frac{(t_n - y_W(x_n))^2}{2} + \lambda \frac{W^T W}{2} \qquad \text{où } \lambda = \frac{\sigma^2}{\alpha^2}$$

45

Régression polynomiale

Maximum *a posteriori* (*MAP*)

Cherche les meilleurs paramètres W en **maximisant la probabilité *a posteriori***

$$\boxed{W = \arg\min_W \sum_{n=1}^{N} \frac{(t_n - y_W(x_n))^2}{2} + \lambda \frac{W^T W}{2}}$$

NOTE

Formule également connue sous le nom de
« **régression de *Ridge*** »

Voir sklearn pour une implémentation simple
scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

46

Théorie de l'information

47

47

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

48

48

Théorie de l'information

Codage de Huffman :

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

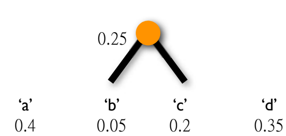
'a'	'b'	'c'	'd'
0.4	0.05	0.2	0.35

49

Théorie de l'information

Codage de Huffman :

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

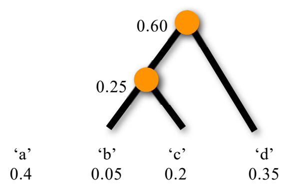


50

Théorie de l’information

Codage de Huffman :

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

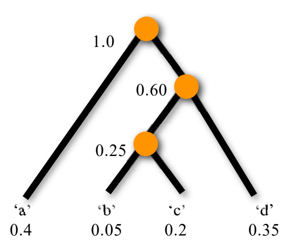


51

Théorie de l’information

Codage de Huffman :

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

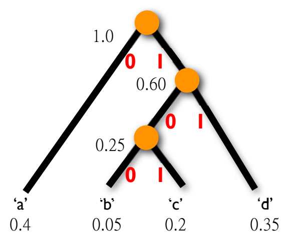


52

Théorie de l'information

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	Prob
'a'	0	40%
'b'	100	5%
'c'	101	20%
'd'	11	35%

53

Entropie

Symbole	Code	Prob
'a'	0	40%
'b'	100	5%
'c'	101	20%
'd'	11	35%

- Soit $p(x)$ la probabilité d'observer le symbole x

  la taille moyenne du code d'un symbole est

  $0.4 \times 1 + 0.05 \times 3 + 0.2 \times 3 + 0.35 \times 2 =$ **<u>1.85 (bits)</u>**

- **Entropie** :

$$H[x] = -\sum_x p(x) \log_2 p(x) \approx \underline{\textbf{1.739 (bits)}}$$

Claude Shannon a démontré qu'il est impossible de compresser l'information dans un plus petit code moyen **sans perte d'information**

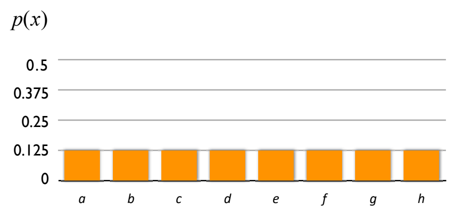
$-\log_2 p(x)$ est l'information contenue par x

54

Entropie

Plus $p(x)$ est proche d'une **loi uniforme**, plus **l'entropie est élevée**

exemple : $x \in \{a, b, c, d, e, f, g, h\}$



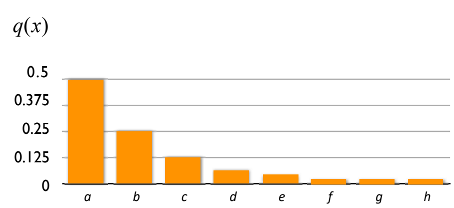
$$H[x] = -\sum_x p(x) \log_2 p(x)$$
$$= -\left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right)$$
$$= -8\left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right)$$
$$= 3 \text{ bits}$$

55

Entropie

Plus $q(x)$ s'éloigne d'une **loi uniforme**, plus **l'entropie est faible**

exemple : $x \in \{a, b, c, d, e, f, g, h\}$



$$H[x] = -\sum_x q(x) \log_2 q(x)$$
$$= -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) - \left(\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) - \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right) - \left(\frac{1}{16}\log_2\left(\frac{1}{16}\right)\right) - \left(\frac{1}{32}\log_2\left(\frac{1}{32}\right)\right) - 3\left(\frac{1}{64}\log_2\left(\frac{1}{64}\right)\right)$$
$$= 2.06 \text{ bits}$$

56

Entropie

L'entropie se généralise aux variables continues

$$H[x] = -\int_{-\infty}^{\infty} p(x)\log_2 p(x)dx$$

57

Entropie relative et divergence de Kullback-Leibler

- Si on ne connaît 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 :

$$KL(p(x) \| q(x)) = -\sum_x p(x)\log_2 q(x) - \left(-\sum_x p(x)\log_2 p(x) \right)$$
$$= \sum_x p(x)\log_2 \frac{p(x)}{q(x)}$$

➢ correspond au **nombre de bits additionnels** par rapport à ce qui serait optimal

58

Entropie jointe

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 **plusieurs variables**

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

59

Entropie conditionnelle

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

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

On peut démontrer que

$$H[x, y] = H[y \mid x] + H[x]$$

60

Information mutuelle

- Mesure à quel point deux variables sont indépendantes

$$I(x,y) = KL\left(p(x,y) \,\|\, p(x)p(y)\right)$$
$$= \sum_{x,y} p\left(x,y\right) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

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

61