Révision des probabilités

Roland Memisevic

Probabilités en IA

- Les probabilités nous permettent de quantifier des incertitudes :
- Au lieu d'une *valeur*, on utilise une distribution sur plusieurs valeurs alternatives.
- Exemple : À la place de '$x = 4$', on peut définir toutes les valeurs

$$p(x = 1), p(x = 2), p(x = 3), p(x = 4), p(x = 5)$$

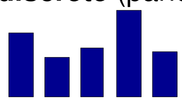
- Avantages :
  1. Robustesse (un modèle peut communiquer toutes ses connaissances)
  2. Mesure de l'incertitude (on a des "error bars")
- On peut toujours représenter '$x = 4$' comme un cas particulier (comment le feriez-vous ?)

Variables aléatoires

- L'attribut le plus important d'une variable aléatoire est sa *distribution*.
- $p(x)$ est une distribution si $p(x) \geq 0 \text{ and } \sum_x p(x) = 1$
- Bizarreries de notation :
  - Le symbole p est très "surchargé" : Par exemple, dans l'expression "$p(x, y) = p(x)p(y)$", chaque p a une signification différente !
  - Parfois on écrit X pour signifier la variable aléatoire et x pour les valeurs qu'elle prend : $p(X = x)$.
  - $\sum_x \ldots$ signifie la somme sur toutes les valeurs que x peut avoir.
- Pour des variables continues, on remplace $\sum$ par $\int$ (il y a d'autres techniques différentes (dues à la théorie de la mesure) qu'on peut généralement ignorer.
- Parfois, l'expression "fonction de densité de probabilité" est utilisée pour se référer à une variable aléatoire continue et "distribution" pour se référer à une variable discrète.

Certaines distributions courantes (1d)

Discrète

- **Bernoulli :** $p^x(1-p)^{1-x}$ où x est 0 ou 1
- **Distribution discrète** (parfois également appelé "multinoulli") : 
- Distribution **binomiale**, **multinomiale** : Somme sur Bernoulli/Discrète. (Parfois "multinomiale" est utilisé pour se référer à une distribution discrète...)
- **Poisson** : $p(k) = \frac{\lambda^k \exp(-\lambda)}{k!}$

Continue

- **Densité uniforme** : 
- **Densité gaussienne (1d)** :
  $p(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

Représenter des valeurs discrètes

- Un moyen très utile pour représenter une variable qui prend une valeur k parmi K valeurs possibles :
- Vecteur de dimension K, contenant $(K-1)$ 0, et un 1

$$\mathbf{x} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

- Codage **1-de-K**, **codage orthogonal**, **one-hot encoding**
- Notez que nous pouvons interpréter $\mathbf{x}$ comme une distribution !

Multinoulli utilisant le codage orthogonal

- L'encodage orthogonal nous permet d'écrire la distribution discrète :

$$p(\mathbf{x}) = \prod_k \mu_k^{x_k}$$

où μ_k signifie la probabilité d'état k.
- Cela peut considérablement simplifier les calculs.

Propriétés

- Les propriétés d'une variable aléatoire sont des propriétés de sa distribution.
- La moyenne :

Propriétés

- Les propriétés d'une variable aléatoire sont des propriétés de sa distribution.
- La moyenne :

$$\mu = \sum_x p(x)x = E[x]$$

- La variance :

Propriétés

- Les propriétés d'une variable aléatoire sont des propriétés de sa distribution.
- La moyenne :

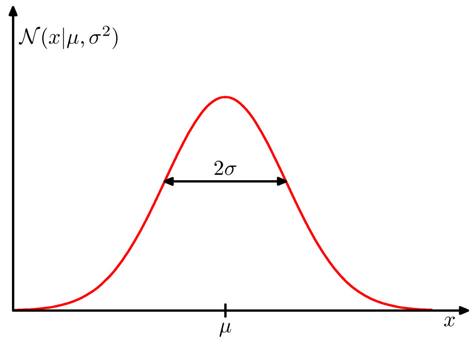
$$\mu = \sum_x p(x)x = E[x]$$

- La variance :

$$\sigma^2 = \sum_x p(x)(x-\mu)^2 = E\left[(x-\mu)^2\right]$$

- (L'écart-type : $\sigma = \sqrt{\sigma^2}$)

La Gaussienne (1d)



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

Plusieurs variables

- La **distribution jointe** $p(x,y)$ de deux variables x, y satisfait aussi

$$p(x,y) > 0 \;\text{ and }\; \sum_{x,y} p(x,y) = 1,$$

- Ou également

$$p(\boldsymbol{x}) > 0 \;\text{ and }\; \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1,$$

  pour vecteur $\boldsymbol{x}$
- Pour des variables discrètes, imaginez un *tableau*.

Probabilité conditionnelle et marginale

- Tout ce qu'on pourrait désirer savoir à propos d'un vecteur aléatoire peut être dérivé de la distribution jointe.
- **distribution marginale** :

$$p(x) = \sum_y p(x,y) \quad \text{and} \quad p(y) = \sum_x p(x,y)$$

- Imaginez l'effondrement de toutes les valeurs dans une dimension du tableau.
- **distribution conditionelle** :

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

- Imaginez une famille de distributions, indexée par la variable conditionnante. (On peut également écrire $p(y|x)$ comme $p_x(y)$).

Propriétés

- La moyenne :

$$\boldsymbol{\mu} = \sum_{\boldsymbol{x}} p(\boldsymbol{x})\boldsymbol{x} = E[\boldsymbol{x}]$$

- La covariance :

$$\mathrm{cov}(x_i, x_j) = E\left[(x_i - \mu_i)(x_j - \mu_j)\right]$$

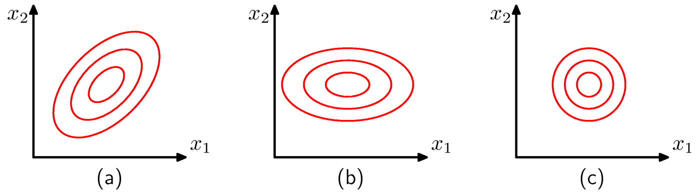
- Matrice de covariance :

$$\Sigma_{ij} = \mathrm{cov}(x_i, x_j) \quad \left(\Sigma = \sum_{\boldsymbol{x}} p(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\right)$$

- Le coefficient de corrélation :

$$\mathrm{corr}(x_i, x_j) = \frac{\mathrm{cov}(x_i, x_j)}{\sqrt{\sigma_i^2 \sigma_j^2}}$$

La Gaussienne multivariée



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

Une formule fondamentale

$$p(x|y)p(y) = p(x, y) = p(y|x)p(x)$$

- Peut être généralisée à plus de deux variables ("règle de la chaîne des probabilités").
- Un cas particulier est **la règle de Bayes :**

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

Indépendance

- Deux variables aléatoires sont **indépendantes** si

$$p(x, y) = p(x)p(y)$$

- Notez que cette définition implique

$$p(y|x) = p(y)$$

- (Être indépendant implique d'être décorrélé, mais pas l'inverse)
- Deux variables aléatoires sont **conditionnellement indépendantes**, sachant une troisième z, si

$$p(x, y|z) = p(x|z)p(y|z)$$

- (Notez que tous ces concepts sont toujours des propriétés de la distribution jointe.)

L'utilité de l'indépendance

- Imaginez un ensemble de variables, $x_1, x_2, \ldots, x_K$
- Simplement définir leur densité jointe (sans parler de faire des calculs avec elle) est tout à fait impossible pour une grande valeur de K !
- Mais si tous les x_i sont indépendants ?
- Dans ce cas, il faut simplement définir K probabilités, car *la densité jointe en est le produit.*
- On peut étendre cette idée (très loin) en utilisant l'indépendance conditionnelle ("modèles graphiques").

Maximum de vraisemblance

- Une autre propriété utile de l'indépendence :
- Imaginez que nous avons un ensemble de données

$$(x_1, \ldots, x_N)$$

  et que nous voulons construire un modèle du processus qui a généré ces données.
- Approche : Ajuster un modèle probabiliste $p(x; \boldsymbol{w})$, qui contient des paramètres $\boldsymbol{w}$.
- Comment ? Maximisez la probabilité de "voir" ces données selon ce modèle !

Maximum de vraisemblance

- Cela est facile si les exemples sont indépendants et distribués identiquement ("**iid**") :

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{w}) = \prod_i p(\boldsymbol{x}_i; \boldsymbol{w})$$

- Au lieu de maximiser la probabilité, nous pouvons maximiser la log-probabilité, car le log est monotone :

$$L(\boldsymbol{w}) := \log \prod_i p(\boldsymbol{x}_i; \boldsymbol{w}) = \sum_i \log p(\boldsymbol{x}_i; \boldsymbol{w})$$

- Chaque exemple $\boldsymbol{x}_i$ contribue un terme additif à l'objectif.

Maximum de vraisemblance pour la moyenne d'une Gaussienne

- Il faut maximiser

$$L(\mu) = \sum_i \log p(x_i; \mu) = \sum_i \left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) - \text{const.}$$

- La dérivée est :

$$\frac{\partial L(\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_i (x_i - \mu) = \frac{1}{\sigma^2} \Big(\sum_i x_i - N\mu \Big)$$

- Mettre à zero :

$$\mu = \frac{1}{N} \sum_i x_i$$