Modèles graphiques probabilistes (et pas que)

Jill-Jênn Vie

18 janvier 2022

Inférence bayésienne (Bayes†, 1763) (Laplace, 1774)

Définitions

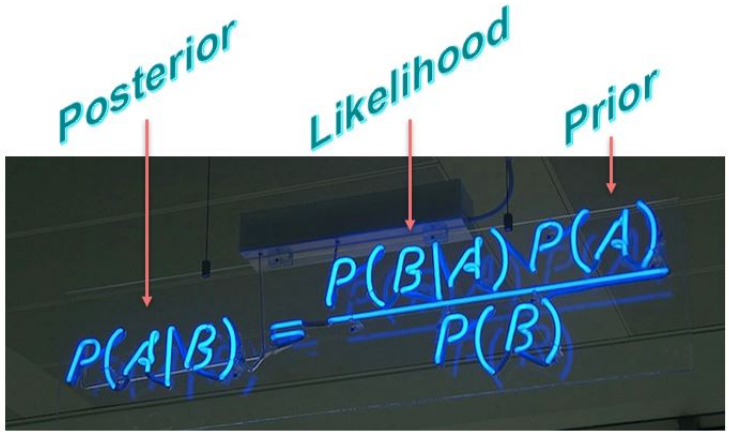
Jointe $p(z, x)$

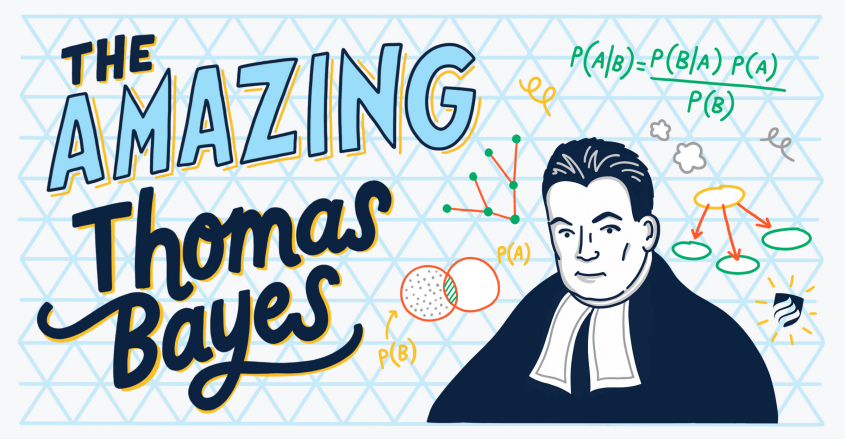
Marginale $p(x)$

Vraisemblance $p(x|z)$

A priori $p(z)$

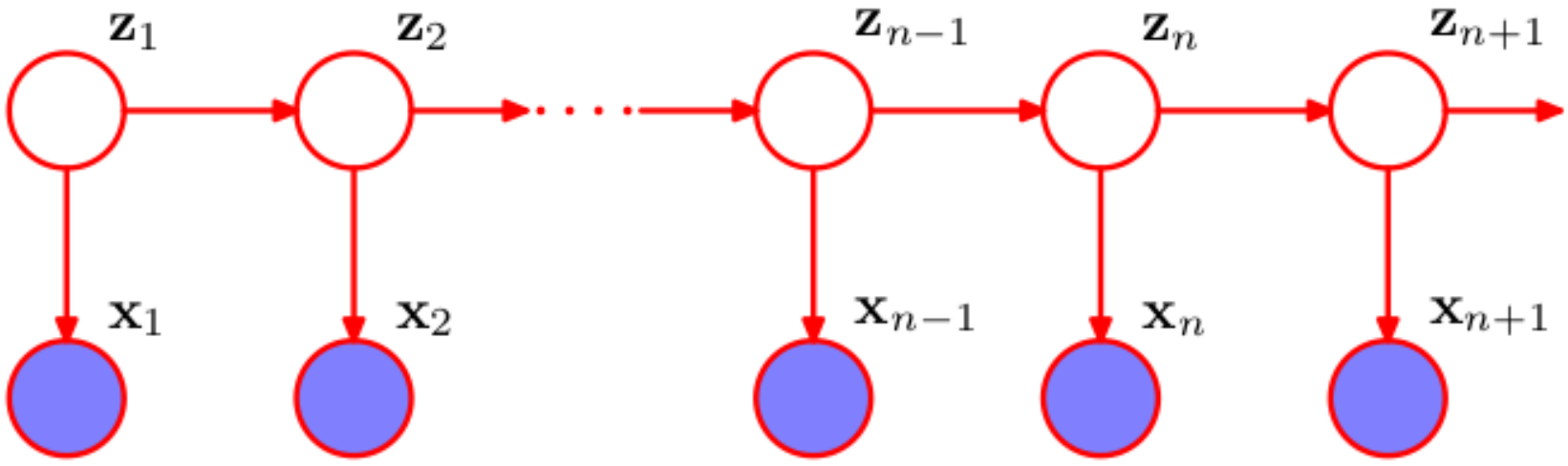
A posteriori $p(z|x)$





†Bayes est décédé en 1761, Price a amélioré ses notes et les a publiées

Pourquoi les modèles graphiques probabilistes ?



$$p(x_{1:N}, z_{1:N}) = p(z_1) \prod_{n=2}^{N} p(z_n|z_{n-1}) \prod_{n=1}^{N} p(x_n|z_n)$$

$$p(V) = \prod_{v \in V} p(v|\text{antécédents}(v))$$

Algorithme de Viterbi

But (en français)

Décoder la séquence de variables latentes la plus probable

But (en maths)

$$\underset{z_{1:N}}{\operatorname{argmax}}\, p(z_{1:N}|x_{1:N})$$

Applications

Traitement du langage (reconnaissance de parole, synthèse vocale), décodage de codes convolutionnels (satellite, Wi-Fi, radio), bio-informatique

Références

(Russell et Norvig, 2020) mais c'est mieux fait dans (Murphy, 2012)

Mais alors ??

Multiplication de matrice	$\sum_k a_{ik} b_{kj}$	$(+, \times)$	(Binet, 1812)
Plus court chemin	$\min_k d_{ik} + d_{kj}$	$(\min, +)$	(Bellman, 1958)
Maximum a posteriori	$\max_k p_{ik} p_{kj}$	$(\max, \times)$	(Viterbi, 1967)

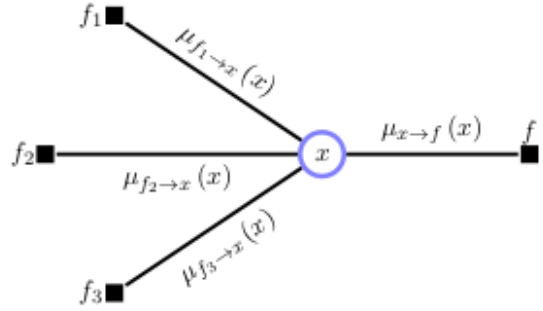
Grâce à (Bellman et) la distributivité sur les semi-anneaux

Algorithme de propagation des convictions (somme-produit)

Par Pearl (1982). Bien fait (Barber, 2012), détaillé (Bishop, 2006).

Variable to Factor message

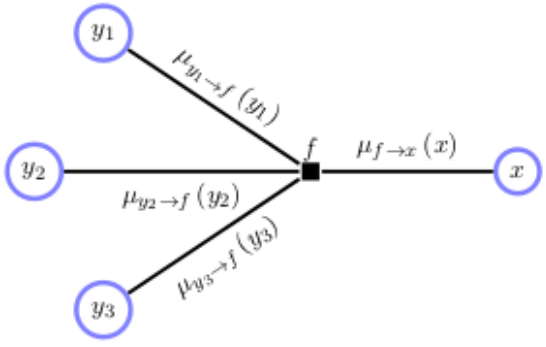
$$\mu_{x\to f}(x) = \prod_{g\in\{\mathrm{ne}(x)\setminus f\}} \mu_{g\to x}(x)$$



Factor to Variable message

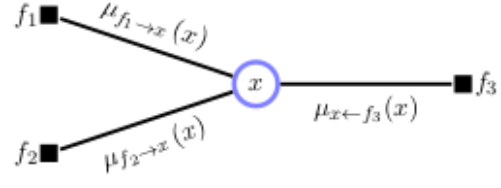
$$\mu_{f\to x}(x) = \sum_{\mathcal{X}_f\setminus x} \phi_f(\mathcal{X}_f) \prod_{y\in\{\mathrm{ne}(f)\setminus x\}} \mu_{y\to f}(y)$$

We write $\sum_{\mathcal{X}_f\setminus x}$ to denote summation over all states in the set of variables $\mathcal{X}_f\setminus x$.



Marginal

$$p(x) \propto \prod_{f\in\mathrm{ne}(x)} \mu_{f\to x}(x)$$



Optimisation

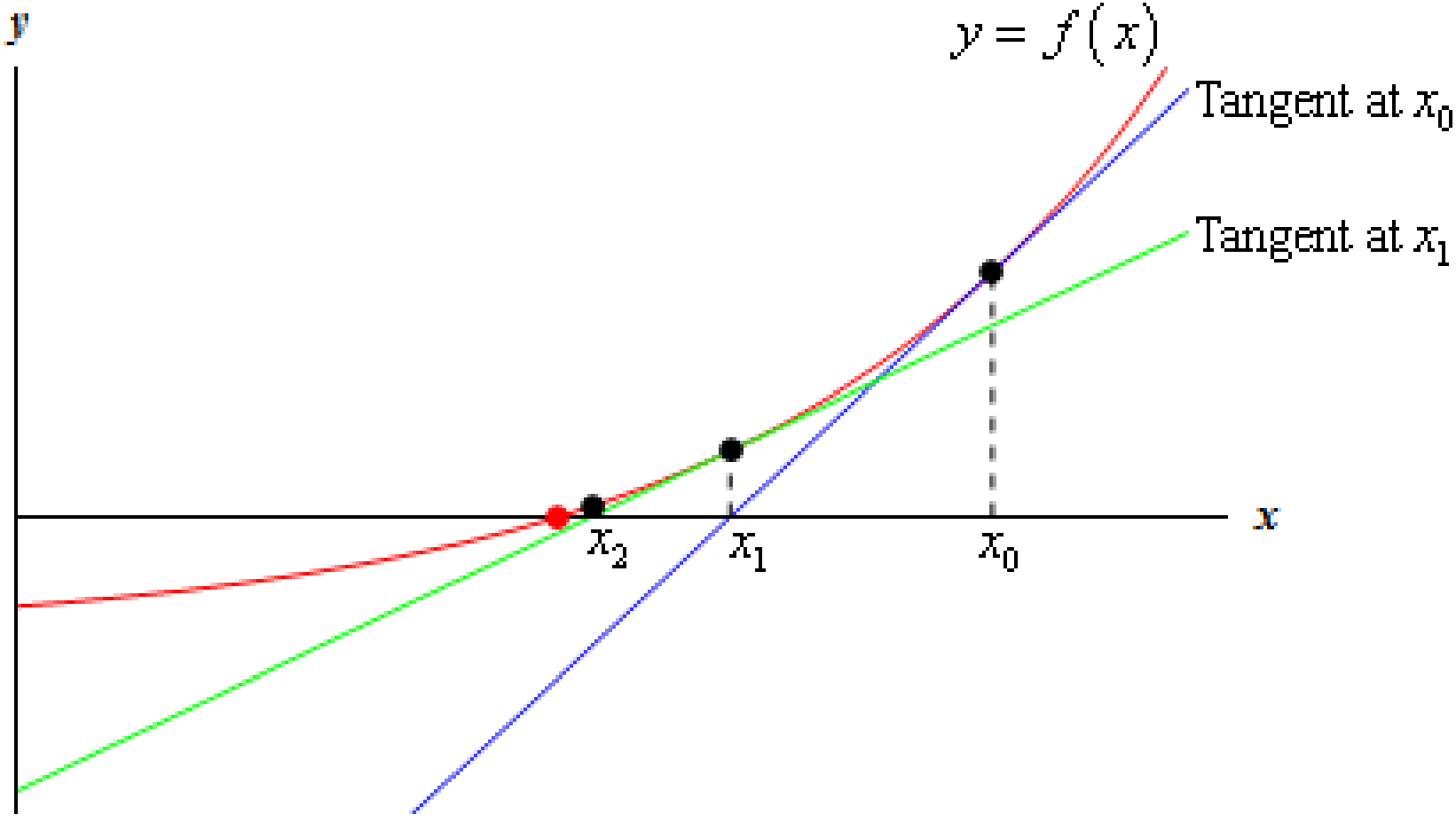
Trouver “les meilleurs” paramètres pour atteindre un but

En général, minimiser une erreur

$\Leftrightarrow$ Trouver les zéros d'une fonction (la dérivée)

Comment trouver les zéros d'une fonction ?

Méthode de Newton : trouver x tel que $f(x) = 0$



$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

Convergence quadratique

$$\exists C > 0, |x_{t+1} - \ell| \leq C|x_t - \ell|^2$$

En plus grande dimension

Et si $g : \mathbf{R}^d \to \mathbf{R}$ avec $d >> 1$?

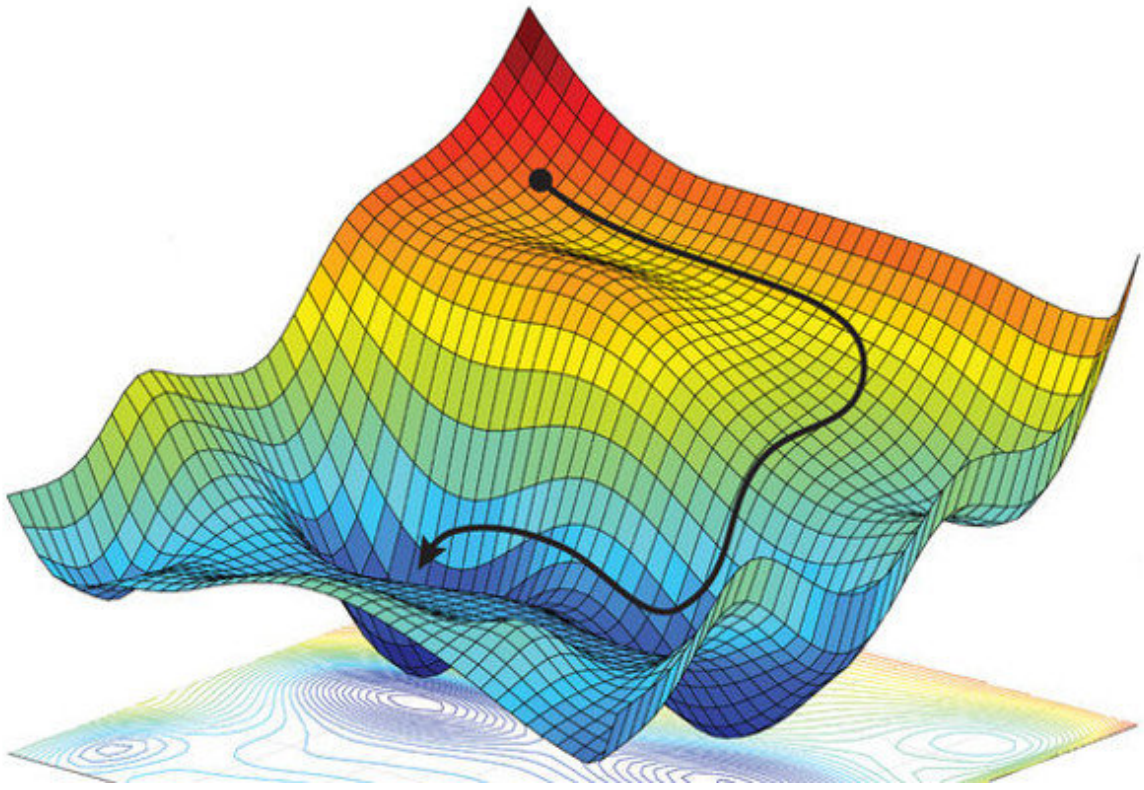
Que signifie $g'(\boldsymbol{x})$? Quelle est sa taille ? On le note $\frac{\partial g}{\partial \boldsymbol{x}}$ ou $\nabla_{\boldsymbol{x}} g$.

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \underbrace{g''(\boldsymbol{x}_t)^{-1}}_{\in \mathbf{R}^{n\times n}, O(n^3)} \underbrace{g'(\boldsymbol{x}_t)}_{\in \mathbf{R}^n}$$

Application

Régression logistique : $\mathcal{L} = \sum_i y_i \log y_i + (1 - y_i) \log(1 - y_i)$

Descente de gradient



$$x_{t+1} = x_t - \gamma \nabla_x(x_t)$$

```
for each epoch:
    for x, y in dataset:
        compute gradients ∂ℒ/∂θ (y, f(x))   # also noted ∇_θ ℒ
        θ ← θ − γ∇_θ ℒ   # γ is the learning rate
```

- Barber, David (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
- Bayes, Thomas (1763). "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S". In : *Philosophical transactions of the Royal Society of London* 53, p. 370-418. URL : https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053.
- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*.
- Murphy, Kevin P (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Russell, Stuart et Peter Norvig (2020). *Artificial Intelligence: A Modern Approach*.