

Introduction aux modèles probabilistes



Guillaume Obozinski

INRIA - Ecole Normale Supérieure - Paris



E.S. Apprentissage artificiel
Ecole des Mines, 29 avril 2011

Apprentissage automatique

Objet

- Extraire des “relations statistiques” entre
 - un grand nombre de variables d'entrées / prédicteurs / descripteurs
 - une ou plusieurs variables de sorties / décision(s)
- Construire une **connaissance empirique**:
Transformation d'information empirique en connaissance statistique

Spécificités par rapport aux autres approches en IA

- 1 Connaissance construite essentiellement à partir des données
- 2 Capacité de **généralisation**

Spécificité par rapport aux statistiques classiques

But

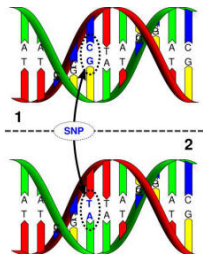
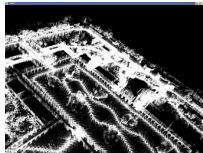
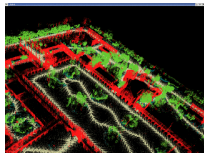
Modèle prédictif/d'action vs modèle explicatif de la réalité.

Difficulté

Nécessité d'intégrer un **un très grand nombre de variables**

- Vision artificielle: 10^7 dimensions par image
- Imagerie Cérébrale: 10^5 dimensions par volume
- Traitement automatique des langues: $10^4 - 10^{15}$ paramètres
- Génétique: 10^4 gènes, 10^5 SNPs/ microsatellites, 10^9 bases d'ADN

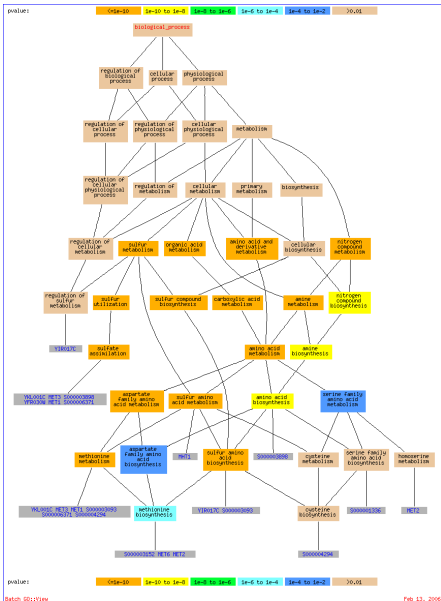
Des problèmes structurés de grande dimension



SNPs or SNPs =

sites of variation in the genome
(spelling mistakes)

Karen	AGCTTGAC	TCCA	TGATGATT
Debo	AGCTTGAC	GCCAT	TGATGATT
Jose	AGCTTGAC	TCCC	TGATGATT
Thomas	AGCTTGAC	GCCC	TGATGATT
Anupriya	AGCTTGAC	TCCA	TGATGATT
Robert	AGCTTGAC	GCCA	TGATGATT
Michelle	AGCTTGAC	TCCC	TGATGATT
Zhijun	AGCTTGAC	GCCC	TGATGATT



Fléau de la dimension (Curse of dimensionality)

Croissance exponentielle du "volume" avec la dimension

⇒ le nombre de paramètres des modèles croît exponentiellement.

Histogrammes

Construire l'histogramme de $X \in [0, 1]$ avec 10 intervalles

→ possible avec 100 observations

Construire l'histogramme de $X \in [0, 1]^{10}$

→ taille et nombre d'intervalles ?

→ a priori impossible avec 100 ou même 10^6 observations !

Modèle de SNPs

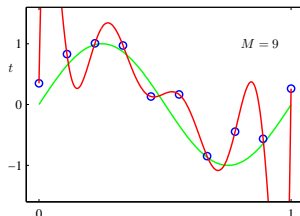
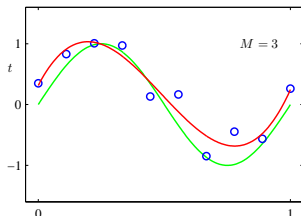
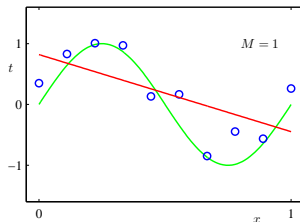
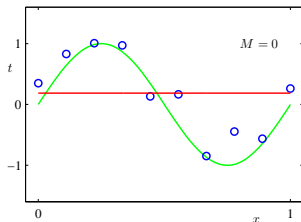
SNP: Single-Nucleotide Polymorphism

- Correspondent à 90% des variations génétiques humaines
- Nombre de loci $k > 10^5$
- Nombre de configurations $> 2^{10^5} \dots$

Sur-apprentissage (Overfitting)

Régression Linéaire $Y = w_0 + w_1X + w_2X^2 + \dots + w_pX^p + \varepsilon$

$$\min_{\mathbf{w}} \frac{1}{2n} \sum_{i=1}^n (y_i - (w_0 + w_1x_i + w_2x_i^2 + \dots + w_px_i^p))^2$$



Notations, formules, définitions

- Loi jointe de X_A et X_B : $p(x_A, x_B)$
- Loi marginale : $p(x_A) = \sum_{x_{A^c}} p(x_A, x_{A^c})$
- Loi conditionnelle : $p(x_A|x_B) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p(x_B) \neq 0$

Formule de Bayes

$$p(x_A|x_B) = \frac{p(x_B|x_A) p(x_A)}{p(x_B)}$$

→ La formule de Bayes **n'est pas** "bayésienne".

Espérances et Variances

- Espérance de X : $\mathbb{E}[X] = \sum_x x \cdot p(x)$
- Espérance de $f(X)$, pour f mesurable :

$$\mathbb{E}[f(X)] = \sum_x f(x) \cdot p(x)$$

- Variance :

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

- Espérance conditionnelle de X sachant Y :

$$\mathbb{E}[X|Y] = \sum_x x \cdot p(x|y)$$

- Variance conditionnelle :

$$\text{Var}(X | Y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2 | Y] = \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2$$

Notions d'indépendance

Indépendance: $X \perp\!\!\!\perp Y$

On dit que X et Y sont indépendantes et on note $X \perp\!\!\!\perp Y$ ssi:

$$\forall x, y, \quad P(X = x, Y = y) = P(X = x) P(Y = y)$$

Indépendance conditionnelle: $X \perp\!\!\!\perp Y \mid Z$

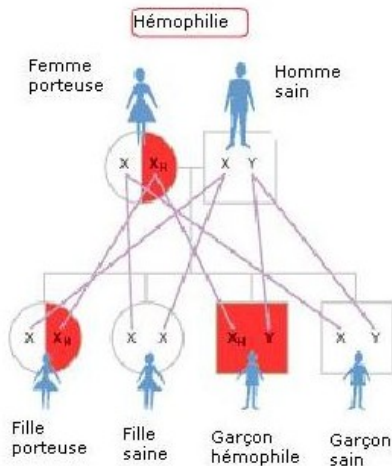
- On dit que X et Y sont indépendantes conditionnellement à Z et
- on note $X \perp\!\!\!\perp Y \mid Z$ ssi:

$\forall x, y, z,$

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) P(Y = y \mid Z = z)$$

Indépendance Conditionnelle: exemple

Exemple d'une "maladie récessive liée à l'X":
Transmission du gène de l'hémophilie



Risques de maladie pour des fils d'un père sain:

- dépendant pour deux frères.
- conditionnellement indépendant sachant si la mère est porteuse ou non.

Modèle statistique

Modèle paramétrique – Définition:

Ensemble de distributions paramétré par un vecteur $\theta \in \Theta \subset \mathbb{R}^p$

$$\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$$

Modèle de Bernoulli: $X \sim \text{Ber}(\theta)$ $\Theta = [0, 1]$

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

Modèle Binomial : $X \sim \text{Bin}(n, \theta)$ $\Theta = [0, 1]$

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

Modèle Multinomial: $X \sim \mathcal{M}(n, \pi_1, \pi_2, \dots, \pi_K)$ $\Theta = [0, 1]^K$

$$p(x|\theta) = \binom{n}{x_1, \dots, x_k} \pi_1^{x_1} \dots \pi_k^{x_k}$$

Modèle gaussien

Loi gaussienne réelle : $X \sim \mathcal{N}(\mu, \sigma^2)$

X une v.a. réelle, et $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$.

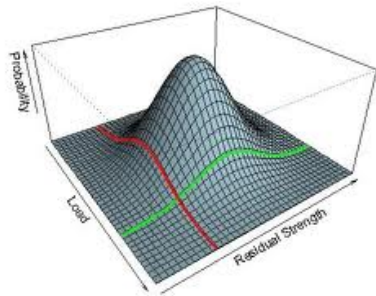
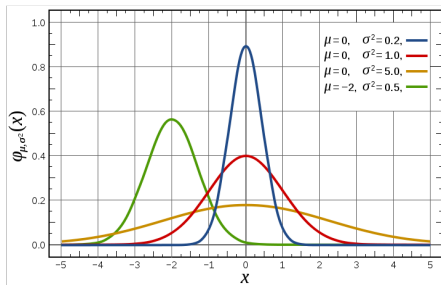
$$p_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right)$$

Loi gaussienne multidimensionnelle: $X \sim \mathcal{N}(\mu, \Sigma)$

X v.a. à valeur dans \mathbb{R}^d . Si \mathcal{K}_n est ensemble des matrices $n \times n$ définies positives, et $\theta = (\mu, \Sigma) \in \Theta = \mathbb{R}^d \times \mathcal{K}_n$.

$$p_{\mu, \Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

Densités gaussiennes



Principe du Maximum de vraisemblance

- Soit un modèle $\mathcal{P}_\Theta = \{p(x|\theta) \mid \theta \in \Theta\}$
- Soit une observation x

Vraisemblance:

$$\begin{aligned}\mathcal{L} : \Theta &\rightarrow \mathbb{R}_+ \\ \theta &\mapsto p(x|\theta)\end{aligned}$$

Estimateur du maximum de vraisemblance:

$$\hat{\theta}_{\text{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}} p(x|\theta)$$

Cas de données i.i.d

Pour $(x_i)_{1 \leq i \leq n}$ un *échantillon* de données i.i.d de taille n :

$$\hat{\theta}_{\text{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^n p(x_i|\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^n \log p(x_i|\theta)$$



Sir Ronald Fisher
(1890-1962)

Exemples de calculs de l'EMV

- Modèle de Bernoulli
- Modèle Multinomial
- Modèle Gaussien

Estimation bayésienne

On traite le paramètre θ comme une **variable aléatoire**.

A priori

On dispose d'un *a priori* $p(\theta)$ sur les paramètres du modèle.

A posteriori

Les observations contribuent via la vraisemblance: $p(x|\theta)$.

La probabilité *a posteriori* du modèle est alors

$$p(\theta|x) = \frac{p(x|\theta) p(\theta)}{p(x)} \propto p(x|\theta) p(\theta).$$

→ L'estimateur bayésien est donc une distribution de probabilité sur les paramètres.

On parle d'*inférence* bayésienne.

Entropie et Divergence de Kullback-Leibler

Entropie

$$H(p) = - \sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)]$$

→ Espérance de la moins log-vraisemblance

Divergence de Kullback-Leibler

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right]$$

→ Espérance du log-rapport de vraisemblance

→ Propriété: $KL(p\|q) \geq 0$

La régression linéaire

Le classifieur naïf de Bayes

Données

Etiquette de classe:

$$Z \in \{1, \dots, K\}$$

Descripteurs X_j , $j = 1, \dots, D$

Modèle

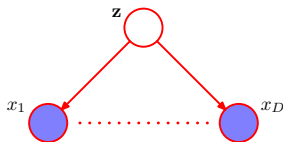
Modèle multinomial:

$p(z = k) = \pi_k$ Modèle pour

$p(x_1, \dots, x_D | z = k)$?

“Hypothèse naïve”

$$\begin{aligned} p(x_1, \dots, x_D | z = k) \\ = \prod_{j=1}^D p(x_j | z = k; \theta_{k,j}) \end{aligned}$$



Apprentissage (estimation)

$$\hat{\pi} = \operatorname{argmax}_{\pi: \pi^\top \mathbf{1} = 1} \prod_{k,j} \pi_k^{\delta(z^{(i)}, k)}$$

$$\hat{\theta}_{k,j} = \operatorname{argmax}_{\theta_{k,j}} \sum_{i=1}^n \log p(x_j^{(i)} | z^{(i)} = k; \theta_{k,j})$$

Le classifieur naïf de Bayes (suite)

Prédiction (décodage):

$$\hat{z} = \operatorname{argmax}_z \frac{\prod_{j=1}^D p(x_j|z)p(z)}{\sum_{z'} \prod_{j=1}^D p(x_j|z')p(z')}$$

Propriétés

- Ignore la corrélation entre descripteurs
- Prédiction requiert seulement la règle de Bayes
- Modèle peut-être appris massivement en parallèle
- Complexité en $\mathcal{O}(nD)$

Problème du clustering et K-means

Modèle de Mixture Gaussien

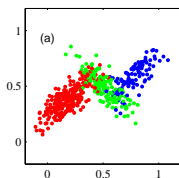
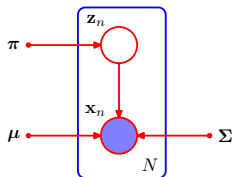
- K composantes
- \mathbf{z} indicateur de la composante
- $\mathbf{z} = (z_1, \dots, z_K)^\top \in \{0, 1\}^K$
- $\mathbf{z} \sim \mathcal{M}(1, (\pi_1, \dots, \pi_K))$

- $$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

- $$p(\mathbf{x}|\mathbf{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^K z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- $$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Estimation:
$$\operatorname{argmax}_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k} \log \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$$



Espérance-Maximisation

$$\begin{aligned}\log p(\mathbf{x}; \boldsymbol{\theta}) &= \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z})} \\ &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z})} =: F(q, \boldsymbol{\theta})\end{aligned}$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) - KL(q || p(\cdot | \mathbf{x}; \boldsymbol{\theta})) = \log \left[\sum_{\mathbf{z}} q(\mathbf{z}) p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \right] + H(q)$$

- Etape E $\operatorname{argmax}_q \mathcal{L}(q, \boldsymbol{\theta}) = p(\cdot | \mathbf{x}; \boldsymbol{\theta})$
- Etape M $\operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_q [\log p(\mathbf{Z}, \mathbf{x}; \boldsymbol{\theta})]$

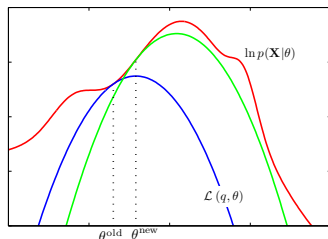
Algorithme EM

- Etape E $\operatorname{argmax}_q \mathcal{L}(q, \theta) = p(\cdot | \mathbf{x}; \theta)$
- Etape M $\operatorname{argmax}_\theta \mathcal{L}(q, \theta) = \operatorname{argmax}_\theta \mathbb{E}_q [\log p(\mathbf{Z}, \mathbf{x}; \theta)]$

Algorithme

Itérer jusqu'à convergence:

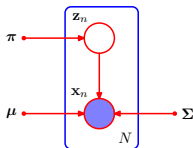
- 1 Etape E:
 - $q^{t+1} = p(\cdot, \mathbf{x}; \theta^t)$
 - $Q(\theta, \theta^t) = \mathbb{E} [\log p(\mathbf{Z}, \mathbf{x}; \theta) | \theta^t]$
- 2 Etape M:
 - $\theta^{t+1} = \operatorname{argmax}_\theta Q(\theta, \theta^t)$



Algorithme EM pour la Mixture Gaussienne

Soit $\theta^t = (\pi^t, (\mu_k^t, \Sigma_k^t)_k)$.

$$\prod_{i=1}^n p(\mathbf{z}^i, \mathbf{x}^i; \theta) = \prod_{i=1}^n \prod_{k=1}^K \pi_k^{z_k^i} \left(\mathcal{N}(\mathbf{x}^i; \mu_k, \Sigma_k) \right)^{z_k^i}$$



Etape E:

$$p(\mathbf{z}^1, \dots, \mathbf{z}^n | \mathbf{x}^1, \dots, \mathbf{x}^n; \theta^t) = \prod_{i=1}^n p(\mathbf{z}^i | \mathbf{x}^i; \theta^t)$$

$$q_k^i = P(z_k^i = 1 | \mathbf{x}^i; \theta^t) = \frac{p(\mathbf{x}^i | z_k^i = 1; \theta^t) P(z_k^i = 1; \theta^t)}{p(\mathbf{x}^i; \theta^t)} = \frac{\pi_k^t \mathcal{N}(\mathbf{x}^i; \mu_k^t, \Sigma_k^t)}{\sum_{\ell} \pi_{\ell}^t \mathcal{N}(\mathbf{x}^i; \mu_{\ell}^t, \Sigma_{\ell}^t)}$$

$$\mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x} | \theta)] = \mathbb{E}_q \left[\sum_{i,k} z_k^i (\log \pi_k + \log \mathcal{N}(\mathbf{x}^i; \mu_k, \Sigma_k)) \right]$$

$$= \sum_{i,k} q_k^i \log \pi_k - \frac{1}{2} q_k^i (\mathbf{x}_i - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}_i - \mu_k) - \frac{1}{2} q_k^i \log((2\pi)^d |\Sigma_k|)$$

Algorithme EM pour la Mixture Gaussienne II

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \sum_{i,k} q_k^i \log \pi_k - \frac{1}{2} q_k^i (x_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (x_i - \boldsymbol{\mu}_k) - \frac{1}{2} q_k^i \log((2\pi)^d |\boldsymbol{\Sigma}_k|)$$

Etape M:

$$\max_{\pi, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k} Q\left((\pi, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k), \boldsymbol{\theta}^t\right) \quad \text{s.t.} \quad \sum_k \pi_k = 1$$

Après calculs:

$$n_k^{t+1} = \sum_i q_k^i$$

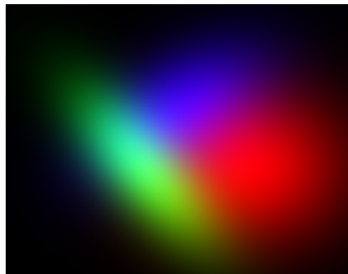
$$\pi_k^{t+1} = \frac{n_k^{t+1}}{n}$$

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i q_k^i x_i$$

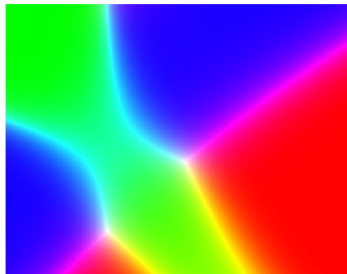
$$\boldsymbol{\Sigma}_k^{t+1} = \frac{1}{n_k^{t+1}} \sum_i q_k^i (x_i - \boldsymbol{\mu}_k^{t+1})(x_i - \boldsymbol{\mu}_k^{t+1})^\top$$

Algorithme EM pour la Mixture Gaussienne III

$$p(\mathbf{x}|\mathbf{z})$$

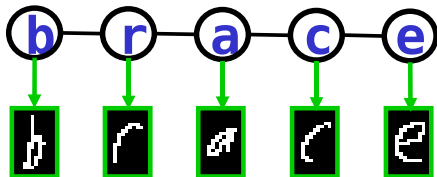


$$p(\mathbf{z}|\mathbf{x})$$

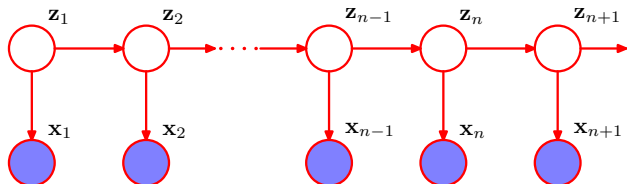


Chaîne de Markov Cachée (HMM)

- reconnaissance vocale
- langage naturel
- reconnaissance d'écriture manuscrite
- séquence biologiques (protéines, DNA)



Chaîne de Markov Cachée (HMM)



$$p(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{z}_1, \dots, \mathbf{z}_N) = p(\mathbf{z}_1) \prod_{n=2}^N p(\mathbf{z}_n | \mathbf{z}_{n-1}) \prod_{n=1}^N p(\mathbf{x}_n | \mathbf{z}_n)$$

Chaîne de Markov homogène

- $\mathbf{z}_n \in \{0, 1\}^K$ indicateur d'état ($1, \dots, K$)
- chaîne de Markov *homogène*: $\forall n, p(\mathbf{z}_n | \mathbf{z}_{n-1}) = p(\mathbf{z}_2 | \mathbf{z}_1)$
- \mathbf{x}_n symbole émis ($\{0, 1\}^K$) / observation (\mathbb{R}^d)

Chaîne de Markov Cachée (HMM)

Paramétrisation

distribution de l'état initial

$$p(\mathbf{z}_1; \pi) = \prod_{k=1}^K \pi_k^{z_{1k}}$$

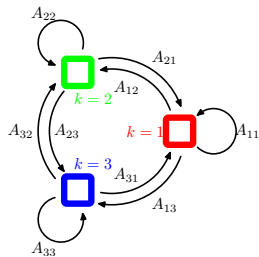
matrice de transition

$$p(\mathbf{z}_n | \mathbf{z}_{n-1}; A) = \prod_{j=1}^K \prod_{k=1}^K A_{jk}^{z_{n-1,j} z_{nk}}$$

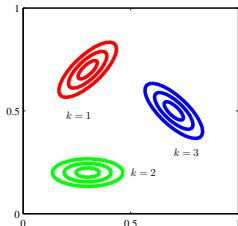
probabilités d'émission

$p(\mathbf{x}_n | \mathbf{z}_n; \phi)$ e.g. Gaussian Mixture

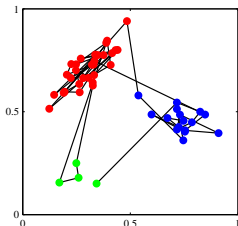
Interprétation



Transitions de \mathbf{z}_n



$p(\mathbf{x}_n | \mathbf{z}_n)$



Trajectoire de \mathbf{x}_n

Maximum de vraisemblance pour les HMM

Application de l'algorithme EM

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{X}, \theta^t) \quad \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}, \theta^t)$$

Espérance de la log-vraisemblance:

$$Q(\theta, \theta^t) = \sum_{k=1}^K \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^N \sum_{j=1}^K \sum_{k=1}^K \xi(z_{n-1,j}, z_{nk}) \log A_{jk} + \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \log p(x_n | \phi_k)$$

En maximisant par rapport aux paramètres $\{\pi, A\}$ on obtient

$$\pi_k^{t+1} = \frac{\gamma(z_{1k})}{\sum_{j=1}^K \gamma(z_{1j})}$$

$$A_{jk}^{t+1} = \frac{\sum_{n=2}^N \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^K \sum_{n=2}^N \xi(z_{n-1,j}, z_{nl})}$$

Si les émissions sont Gaussiennes on a aussi:

$$\mu_k^{t+1} = \frac{\sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n}{\sum_{n=1}^N \gamma(z_{nk})} \quad \Sigma_k^{t+1} = \frac{\sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k) (\mathbf{x}_n - \mu_k)^\top}{\sum_{n=1}^N \gamma(z_{nk})}$$

Maximum de vraisemblance pour les HMM

Application de l'algorithme Somme-Produit

Dans le cadre des HMM, l'algorithme est connu sous le nom *aller-retour* ou algorithme de *Baum-Welch*.

On propage les messages

- forward $\alpha(\mathbf{z}_n) = p(\mathbf{x}_n | \mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n | \mathbf{z}_{n-1})$
- backward $\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1} | \mathbf{z}_{n+1}) p(\mathbf{z}_{n+1} | \mathbf{z}_n)$

qui satisfont les propriétés:

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{z}_n) \quad \beta(\mathbf{z}_n) = p(\mathbf{x}_{n+1}, \dots, \mathbf{x}_N | \mathbf{z}_n)$$

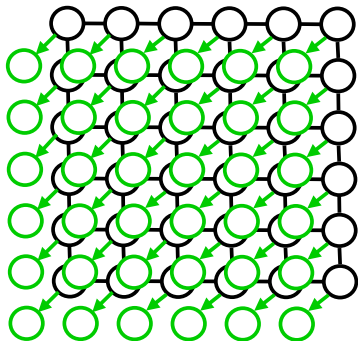
Finalement on obtient les marginales:

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{X}, \theta^t) = \frac{\alpha(\mathbf{z}_n) \beta(\mathbf{z}_n)}{p(\mathbf{X} | \theta^t)}$$

et

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \frac{\alpha(\mathbf{x}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1}) \beta(\mathbf{x}_n)}{p(\mathbf{X} | \theta^t)}$$

Champ de Markov Caché



Original image



Segmentation

Les **modèles graphiques**

- 1 permettent de modéliser des distributions sur un grand nombre de variables en utilisant la structure du problème.
- 2 permettent
 - l'**estimation** fréquentiste avec le maximum de vraisemblance
 - l'**estimation** (ou inférence) bayésienne

Au travers de deux exemples

- l'algorithme EM pour l'estimation dans les modèles de mixture
- l'estimation des Chaînes de Markov Cachées

nous avons vu la nécessité de faire des calculs de probabilités et d'espérance qui utilisent la structure du graphe

→ problème de l'**inférence probabiliste**.

- Une bonne partie des illustrations et des notations de cet exposé proviennent du très bon livre de Christopher Bishop:

Pattern Recognition and Machine Learning, 2006, Springer.

<http://research.microsoft.com/~cmbishop/PRML/>

- Pour aller plus loin:

Daphne Koller et Nir Friedman, Probabilistic Graphical Models - Principles and Techniques, 2009, MIT Press.