Modèles graphiques	**Automne 2007**
Cours 10 — 19 décembre	
Enseignant: Francis Bach	*Scribe: Sylvain Raybaud*

10.1 Sélection de Modèle

Dans les précédents cours on a vu comment estimer les paramètres d'un modèle. Mais il est une problématique qui n'a pas été abordée du tout : c'est le choix du graphe lui même. Il n'existe pas de solution universelle à ce problème.

10.1.1 Maximisation de la vraisemblance

Supposons que l'on veut choisir le meilleur modèle parmi k, $M_1(G_1, \Theta_1), .., M_k(G_k, \Theta_k)$. Une approche intuitive consiste à choisir le modèle pour lequel la vraisemblance des données d'apprentissage est la plus élevée ; soient $x_1, .., x_N$ les données (iid) :

$$k \leftarrow \arg\max_j \{max_\theta\{ln(\prod_{n=1}^{N} p(x_n|M_j, \theta))\}\}$$

Mais cela ne fonctionne pas en pratique, car c'est toujours le modèle "le plus flexible" (i.e. celui ayant le plus de paramètres - donc en général le graphe complet) qui est sélectionné.

10.1.2 Maximisation de la vraisemblance pénalisée par le nombre de paramètres

Pour palier à cet inconvénient, on peut pénaliser les modèles ayant beaucoup de paramètres ; soit $Q_.$ le nombre de paramètres du modèle :

$$k \leftarrow \arg\max_j \{max_\theta\{ln(\prod_{n=1}^{N} p(x_n|M_j, \theta)) - Q_j\}\}$$

On distingue les variantes suivantes :

Akaike Information Criterion (AIC)

$$k \leftarrow \arg\max_j \{max_\theta\{ln(\prod_{n=1}^{N} p(x_n|M_j, \theta)) - \frac{1}{2}Q_j\}\}$$

Bayesian Information Criterion (BIC)

$$k \leftarrow \arg\max_j \{max_\theta\{ln(\prod_{n=1}^{N} p(x_n|M_j,\theta)) - \frac{1}{2}ln(Q_j)\}\}$$

10.1.3 Validation croisée

L'approche la plus couramment utilisée est la *validation croisée* : les données disponibles sont divisées en un ensemble d'apprentissage $\mathcal{E}_{train}$ et un ensemble de test $\mathcal{E}_{test}$ (par exemple, respectivement 90% et 10% des données disponibles) ; on entraine chaque modèle sur l'ensemble d'apprentissage et on l'évalue sur l'ensemble de test. On choisit ensuite le modèle donnant la meilleure vraisemblance aux données de test.

$$\begin{aligned} \forall j \in \{1,..,k\} \quad \theta_j &\leftarrow \arg\max_\theta l(\mathcal{E}_{train}|\theta, M_j) \\ j^{opt} &\leftarrow \arg\max_j l(\mathcal{E}_{test}|\theta_j, M_j) \end{aligned}$$

Ceci a l'avantage de favoriser les modèles ayant de bonnes performances en généralisation.

Une variante est la *validation croisée à K passes* ("k-fold cross-validation") : on divise les données disponibles en K ensembles $\mathcal{E}^1,..,\mathcal{E}^K$ et on définit :

$$\forall j \in \{1,..,K\} \ . \ \begin{cases} \mathcal{E}^j_{train} = \cup_{i \neq j} \mathcal{E}^i \\ \mathcal{E}^j_{test} = \mathcal{E}^j \end{cases}$$

On définit ensuite :

$$\begin{aligned} \forall i \in \{1,..,k\}.\forall j \in \{1,..,K\}.\ \theta^j_i &\leftarrow \arg\max_\theta l(\mathcal{E}^j_{train}|\theta, M_i) \\ \forall i \in \{1,..,k\}.\forall j \in \{1,..,K\}.\ \alpha^j_i &\leftarrow l(\mathcal{E}^j_{test}|\theta^j_i, M_i) \end{aligned}$$

On choisit ensuite le modèle qui donne la meilleure vraisemblance en moyenne :

$$i^{opt} \leftarrow \arg\max_i \sum_{j=1}^{K} \alpha^j_i$$

10.1.4 Sélection dans le cadre Bayésien

Pour estimer les paramètres θ, on définissait la probabilité a posteriori

$$p(\theta|x)\alpha p(x|\theta,x)p(\theta)$$

On va également considérer le modèle M comme une variable aléatoire :

$$p(M|x) = \int_\theta p(M,\theta|x)d\theta = \int_\theta \frac{p(x|M,\theta)p(M,\theta)}{p(x)}d\theta$$

On choisit ensuite le modèle ayant la plus grande probabilité.