Introduction aux modèles graphiques **2010/2011**

Cours 2 — 6 octobre

Enseignant: Francis Bach *Scribe: Nicolas Cheifetz, Issam El Alaoui*

2.1 Maximum de vraisemblance pour une loi Gaussienne multivariée

Soit x, un vecteur de k observations i.i.d. Le vecteur x est la réalisation d'une variable aléatoire X qui suit une loi normale :

$$X \sim \mathcal{N}(\mu, \Sigma), \quad \text{où } \mu \in \mathbb{R}^k \text{ et } \Sigma \in \mathbb{R}^{k \times k}$$

Afin de calculer les paramètres de la loi, on cherche à maximiser la log-vraisemblance :

$$\begin{aligned} l(\theta) &= \log\ p(x_1, \ldots, x_n \mid \theta) \\ &= \sum_{i=1}^{n} \log\ p(x_i \mid \theta) \\ &= \sum_{i=1}^{n} -\frac{k}{2} \cdot \log(2\pi) - \frac{1}{2} \cdot \log \mid \Sigma \mid -\frac{1}{2} \cdot (x_i - \mu)^T \cdot \Sigma^{-1} \cdot (x_i - \mu) \end{aligned}$$

Il s'agit d'optimiser $l(\theta)$; il faut donc calculer successivement deux gradients :

$$\nabla_\mu l(\theta) = \sum_{i=1}^{n} \left(-\frac{1}{2}\right) \cdot 2 \cdot \Sigma^{-1} \cdot (\mu - x_i)$$

$$\nabla_\mu l(\theta) = 0 \quad \Leftrightarrow \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Notons à présent $\Lambda = \Sigma^{-1}$. Dans ce cas, le gradient par rapport Λ est :

$$\nabla_\Lambda l(\theta) = \Lambda^{-1} - \hat{\Sigma} \quad \text{, car } \nabla_\Lambda \log \mid \Lambda \mid = \Lambda^{-1}$$

$$\nabla_\Lambda l(\Lambda) = 0 \quad \Leftrightarrow \quad \hat{\Sigma} = \Lambda^{-1} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu) \cdot (x_i - \mu)^T$$

Remarque: On retrouve la moyenne et la matrice de covariance empirique, ce qui montre bien que le maximum de vraisemblance pour les modèles paramétriques classiques permet d'obtenir les estimateurs classiques.

2.1.1 Régression linéaire

Remarque: un noeud dans un modèle graphique représente une variable aléatoire.

Considérons deux noeuds X et Y dont on observe une répartition sur un axe (X, Y) : On a par exemple envie de dire que Y va dépendre linéairement de X. Il faut donc chercher la loi de $Y \mid X$ en maximisant la vraisemblance.

Supposons que :

$$Y \mid X \sim \mathcal{N}(\theta^T X + \theta_0, \sigma^2)$$

Une petite astuce nous permet de nous affranchir de θ_0.
En écrivant $\tilde{x} = \binom{x}{1}$, où x est en général de dimension k et $\tilde{\theta} = \binom{\theta}{\theta_0}$.
On est ramené à

$$Y \mid X \sim \mathcal{N}(\theta^T X, \sigma^2)$$

Les données du problème sont :

- x_i^j : i-ème observation de la j-ème variable ($x_i \in \mathbb{R}^k$) et $i \in \{1, \ldots, n\}$,$j \in \{1, \ldots, k\}$
- $X \in \mathbb{R}^{n \times k}$ est dite **matrice de design**

On supposera les données indépendantes identiquement distribuées (*i.i.d.*).

Le but est de maximiser la log vraisemblance :

$$\begin{aligned} l(\theta, \sigma^2) &= \sum_{i=1}^{n} \log\ \mathcal{N}(y_i \mid \theta^T \cdot x_i, \sigma^2) \\ &= \sum_{i=1}^{n} \frac{1}{2} \cdot \log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \cdot (\theta^T \cdot x_i - y_i)^2 \end{aligned}$$

On notera avec le symbole $\propto$ une égalité à une constante près.
On calcule la log vraisemblance de (θ, σ^2) :

$$l(\theta, \sigma^2) \propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \cdot \parallel y - X\theta \parallel^2$$

avec $\parallel y - X\theta \parallel^2 = (y - X\theta)^T \cdot (y - X\theta) = (y^T y + \theta^T X^T X\theta - 2y^T X\theta)$.

Tout d'abord, calculons le gradient de la log vraisemblance par rapport à θ :

$$\nabla_\theta \, l(\theta, \sigma^2) = -\frac{1}{2\sigma^2} \cdot 2X^T(X\theta - y)$$

$$ie \;\; \nabla_\theta \, l(\theta, \sigma^2) = 0 \;\Leftrightarrow\; \boxed{X^T X\theta = x^T y} \text{ “équation normale”}$$

Donc, $\theta = (X^T X)^{-1} \cdot X^T y$.

Calculons ensuite le gradient de la log vraisemblance par rapport à $\frac{1}{\sigma^2}$:

$$\nabla_{1/\sigma^2} \, l(\theta, \sigma^2) = \frac{n}{2} \cdot \frac{1}{1/\sigma^2} - \frac{1}{2} \cdot \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

$$\nabla_{1/\sigma^2} \, l(\theta, \sigma^2) = 0 \quad \Leftrightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{\theta}^T x_i)^2$$

Questions :

- $(X^T X)$ est-il inversible ? $\quad \rightarrow$ condition nécessaire : $n \geqslant k$ (si $n < k$, $rg(X^T X) \leqslant n < k$)

  En pratique :

  - on fait du QR
  - on remplace $(X^T X)$ par $(X^T X + \epsilon \cdot Id)$; autrement dit, on pénalise une norme sur un des paramètres pour assurer une convergence des résultats. (CF Cours Audibert *Apprentissage statistique*).
    Cecie est equivalent à d'ajouter $-\frac{\epsilon}{\sigma^2}\theta$ dans $\nabla_\theta l$. Cela revient à ajouter $-\frac{\epsilon}{\sigma^2}\theta||\theta||^2$ dans $l(\theta, \sigma^2)$
  - quand $n \gg k$, pas de problème

- Comment évaluer la qualité de l'estimateur du MV ?
  Soit $x \sim p_{\theta_0}(x)$. Dans le cas de l'estimateur de l'espérance par moyenne empirique, on rappelle que le théorème central limite permet de conclure de la convergence ainsi que de l'erreur commise. Ceci est plus général; en effet, sous certaines hypothèses, $\hat{\theta}_{MV} \xrightarrow[n\to\infty]{} \theta_0$. De plus, on a : $\| \hat{\theta}_{MV} - \theta_0 \| \propto \frac{1}{\sqrt{n}}$.

2.2 Régression logistique

Maintenant, on dispose de $X \in \mathbb{R}^k$ et de $Y \in \{1, ..., q\}$.
On cherche une séparation dans l'espace des paramètres.

Cas : q=2

Il faut construire un modèle prédictif du type de Y. Par exemple, $Y \in \{0, 1\}$.

Limitons-nous d'abord à $q = 2$ (pour $q > 2$, on parle de régression softmax). Un séparateur linéaire est par exemple $\theta^T x + \theta_0 = 0$; on veut en déterminer un modèle :

$$p(y = 1 \mid x) = \sigma(\theta^T x + \theta_0)$$

On choisit la fonction sigmoide : $\sigma(z) = \frac{1}{1+e^{-z}}$. On l'a choisi notamment car $\sigma' = \sigma(1-\sigma)$ et $\sigma(-z)+\sigma(z) = 1$. De plus, son logarithme est concave, ce qui permettra une maximisation plus stable.
Nous travaillons avec les hypothèses suivantes :

- (x_i, y_i) i.i.d. ,$\forall i \in \{1, \ldots, n\}$
- $x_i \in \mathbb{R}^k$ et $y_i \in \{0, 1\}$

$$\begin{aligned} l(\theta) &= \sum_{i/y_i=1} \log(\sigma(\theta^T x_i)) + \sum_{i/y_i=0} \log(1 - \sigma(\theta^T x_i)) \\ &= \sum_{i=1}^{n} y_i \cdot \log(\sigma(\theta^T x_i)) + (1 - y_i) \cdot \log(\sigma(-\theta^T x_i)) \end{aligned}$$

$l(\theta)$ est bien concave car $log(\sigma)$ est concave. Calculons le gradient, et on note : $\sigma(\theta^T x_i) = \eta_i$

$$\nabla_\theta \, l(\theta) = \sum_{i=1}^{n} y_i \cdot x_i \cdot (1 - \eta_i) - (1 - y_i) \cdot x_i \cdot \eta_i$$

$$\Leftrightarrow \quad \nabla_\theta \, l(\theta) = \sum_{i=1}^{n} x_i \cdot (y_i - \eta_i) = \sum_{i=1}^{n} x_i \cdot (y_i - \sigma(\theta^T x_i))$$

La présence de σ nous empêche d'inverser immédiatement. Il faut donc calculer la hessienne $\nabla^2_\theta \, l(\theta) \in R^{k \times k}$, matrice des dérivées secondes pour obtenir : (Cf. poly)

$$\nabla^2_\theta \, l(\theta) = -\sum_{i=1}^{n} x_i \cdot (x_i \cdot \eta_i \cdot (1 - \eta_i))^T = X^T \cdot Diag(\eta_i(1 - \eta_i)) \cdot X$$

2.2.1 Méthodes de descente de gradient

Lorsqu'il est question de minimiser $f : \mathbb{R}^k \to \mathbb{R}$ convexe, une condition nécessaire est d'avoir un gradient nul, mais cette condition n'est pas suffisante.
On peut donc être amené à utiliser des méthodes de descente de gradient.

Une descente produit une séquence $x^{(k)}$ telle que $x^{(k+1)} = x^{(k)} + \varepsilon^{(k)} d^{(k)}$ où $d^{(k)}$ est une direction de descente, $\varepsilon^{(k)} > 0$ est le pas, de sorte à avoir $f(x^{(k+1)}) < f(x^{(k)})$ (sauf pour le $x^{(k)}$ optimal).
Dans la cadre de la descente de gradient, on choisit $d^{(k)} = -\nabla_x f(x^{(k)})$.
Il existe plusieurs stratégies pour définir le pas $\varepsilon^{(k)} > 0$:

1. Le pas constant: $\varepsilon^{(k)} = \varepsilon$. Dans ce cas, l'algorithme n'est pas toujours convergent.
2. Un pas décroissant $\varepsilon^{(k)} \propto \frac{1}{k}$ (avec $\sum_k \varepsilon^{(k)} = \infty$ et $\sum_k {\varepsilon^{(k)}}^2 < \infty$). Toujours convergent mais parfois on peut dépasser le minimum (avec un pas de trop)
3. La "*Line Search*" qui cherche à trouver $\min_\varepsilon f(x^{(k)} + \varepsilon d^{(k)})$:
   - soit de manière exacte (en pratique, c'est une opération coûteuse et souvent inutile). Toujours convergent.
   - soit de manière approchée. Toujours convergent.

Remarque : La minimisation exacte de $\min_{\epsilon \in R} f(x_t - \epsilon \cdot \nabla f(x_t))$ est une perte de temps, sauf si il s'agit d'un problème de minimisation pour un polynôme (recherche des racines).

2.2.2 Méthodes du second ordre

Ici, f est convexe et surtout $f \in C^2$. On a $f : \mathbb{R}^k \to \mathbb{R}$:

$$f(x + \delta) = f(x) + \delta^T \cdot \nabla f(x) + o(||x||)$$

Le développement poussé à la hessienne donnerait : $\delta^T \cdot \nabla^2 f(x) \cdot \delta + o(||x||^2)$.
De plus, on a comme propriété que f est partout dérivable, et convexe si et seulement si $\nabla^2 f(x)$ est semi définie positive : $\forall z \in R^k,\ z^T \cdot \nabla^2 f(z) \cdot z \geq 0$.
En dérivant par rapport à δ :

$$\nabla f(x) + \nabla^2 f(x)\delta = 0$$

ie $\delta = -(\nabla^2 f(x))^{-1} \cdot \nabla f(x)$ est le point qui minimise l'approximation quadratique de $f(x)$.
On utilise la méthode de Newton pour trouver un x vérifiant cette équation pour δ tendant vers 0 :

Méthode du second ordre, méthode de Newton

L'idée sous-jacente de la méthode de Newton est de minimiser l'approximation quadratique de f en $x^{(k)}$, $x \mapsto \tilde{f}(x) = f(x^{(k)}) + \nabla_x f(x^{(k)})^T (x - x^{(k)}) + (x - x^{(k)})^T \nabla_x^2 f(x^{(k)})(x - x^{(k)})$.
On reprend le même schéma de descente décrit précédemment avec désormais, la direction de descente égale à $d^{(k)} = -(\nabla_x^2 f(x^{(k)}))^{-1} \nabla_x f(x^{(k)})$ (on suppose que la Hessienne est bien conditionnée en $x^{(k)}$).

Si la fonction n'est pas convexe, la méthode n'est pas globalement convergente, mais néanmoins localement convergente.

Même si la fonction est convexe, la méthode n'est pas globalement convergente. Pour rendre la méthode de Newton globalement convergente avec une fonction convexe, il est nécessaire d'avoir recours à une *Line Search*.

Cependant, pour la régression logistique, pour n assez grand (par rapport à p), la méthode de Newton est convergente. La complexité algorithmique de cette méthode est alors de l'ordre de $O(p^3 + p^2 n)$ (où p et n sont respectivement la dimension et le nombre de points), correspondant à la formation ainsi qu'à l'inversion de la Hessienne. Il existe de nombreuses extensions de méthodes Newtoniennes - telles que les méthodes dites Quasi-Newtoniennes - qui essaient de réduire la complexité de la procédure en approximant le calcul de la Hessienne.

Remarque: en pratique, lorsque l'on utilisera la méthode de Newton dans la cas de la régression logistique, on pourra constater une convergence en environ 20 itérations, sans line search. De plus, dans le cas de la régression logistique, si le nombre d'observations n est grand devant la dimension de la variable d'entrée p, alors la méthode de Newton est globalement convergente.

2.2.3 Algorithme IRLS pour la régression logistique

Reécrivons le gradient $\nabla_\theta\ l(\theta)$:

$$\nabla_\theta\ l(\theta) = \sum_i x_i \cdot (y_i - \eta_i) = -x^T \cdot (y - \eta)$$

$$\nabla^2_\theta\ l(\theta) = -x^T \cdot Diag(\eta_i(1 - \eta_i)) \cdot x$$

Lorsque cette dernière matrice est définie très négative, elle a une courbure très forte, ce qui impliquera que la méthode de Newton sera efficace. On parle alors de iRLS (iteratively Reweighted Least Squares).
On écrira alors en notant $W = Diag(\eta_i(1 - \eta_i))$

- $\theta_0 = 0$
- $\forall t, \quad \theta_{t+1} = \theta_t - (X^T W X)^{-1} \cdot (X^T(\eta - y))$

Que se passe-t-il lorsque on a plus que deux étiquettes ? Par exemple si $Y \in \{1, ..., q\}$:
On dispose alors d'un équivalent de σ (fonction sigmoide) dans ce cas :

$$p(Y = s \mid x) = \frac{e^{\theta_s^T \cdot x}}{\sum_{j=1}^q e^{\theta_j^T x}}$$

On utilisera alors typiquement une fonction softmax, ou bien une méthode 1 contre tous, ou 1 contre 1 (ce qui est assez naturel graphiquement - cf poly)

2.2.4 Liens avec la SVM

Pour la régression logistique, on prend l'ensemble $\{0, 1\}$, alors qu'en SVM, on choisira plutôt l'ensemble $\{-1, 1\}$, i.e., au lieu de prendre $y_i \in \{0, 1\}$, on considère $s_i = 2y_i - 1 \in \{-1, 1\}$.
Notons $f(u) = \log \frac{1}{\sigma(u)} = \log(1 + e^{-u})$
Alors, on a : $l(\theta) = \sum_{i=1}^{n} \log \left(\frac{1}{\sigma \cdot (s_i \theta^T x_i)} \right) = \sum_{i=1}^{n} f(s_i \theta^T x_i)$.

L'hyperplan séparateur est $\theta^T x = 0$, quelle est l'erreur comise ? On a une erreur de prédiction lorsque $s_i \theta^T x_i \leq 0$. Et le nombre d'erreurs est $\sum_{i=1}^{n} g_{0-1}(s_i \theta^T x_i)$, où $g_{0-1}(u) = 1$ if $u < 0$ and 0 sinon. Ceci est une fonction convexe. Plusieurs choix sont possibles pour la remplacer par une fonction convexe.

La SVM utilise $u \to max(0, 1-u)$ alors que la logistique utilise $f(u) = \log(1+e^{-u})$. Ces deux fonctions (et les prédictions associées) sont très proches.

Le modèle est dit bien "bien spécifié" lorsque $p(Y = 1 \mid x)$ est effectivement égale à $\sigma(\theta^T x)$. Notez que même si le modèle est mal spécifié, on peut toujours appliquer la régression logistique.

2.3 Classification générative vs. discriminative

La classification discriminative utilise un modèle $P(Y \mid X)$ sans chercher de modélisation de densité pour X. La classification générative va utiliser un modèle $P(Y)$ et un modèle $P(X \mid Y)$ pour modéliser la loi jointe $P(X, Y)$.

La règle de Bayes permet de relier ces deux types de classification :

$$P(Y = 1 \mid X) = \frac{P(X \mid Y = 1) \cdot P(Y = 1)}{P(X)}$$

$p(X \mid Y = 1)$ sera typiquement une gaussienne.

Pour une loi de Bernouilli π $(= p(y))$, on prendra :
$p(X \mid Y = 1) \propto \mathcal{N}(\mu_1, \Sigma_1)$
$p(X \mid Y = 0) \propto \mathcal{N}(\mu_0, \Sigma_0)$

Qu'obtient-on alors en terme de discrimination ?

$$p(Y = 1 \mid X) \propto \frac{\pi_1}{(2\pi)^{\frac{k}{2}} \cdot |\Sigma_1|^{\frac{1}{2}}} \exp(\frac{-1}{2}(X - \mu_1)^T \Sigma_1^{-1}(X - \mu_1))$$

$$p(Y = 0 \mid X) \propto \frac{\Pi_1}{(2\pi)^{\frac{k}{2}} \cdot |\Sigma_0|^{\frac{1}{2}}} \exp(\frac{-1}{2}(X - \mu_1)^T \Sigma_0^{-1}(X - \mu_1))$$

Sous l'hypothèse que $\Sigma_0 = \Sigma_1 = \Sigma$ (covariances identiques), on voit apparaître $p(Y = 1 \mid x) = \sigma(\theta^T x + \theta_c)$. On retombe sur la fonction logistique. Voir cours suivant pour plus de détails.