Modèles Graphiques

Barbara Gris & Nelle Varoquaux

October 26, 2011

Contents

1 Apprentissage dans les modèles discrets

Soit z et x deux variables aléatoires pouvant prendre respectivement N et K valeurs telles que: $p(z = m) = \pi_m$ et $p(x = k|z = m) = \theta_{m,k}$

On suppose que l'on observe n valeurs (x_i, z_i) iid. On note $N_{i,j}$ le cardinal de $\{k \in [\![1, n]\!] | (x_k, z_k) = (j, i)\}$ et N_i celui de $\{j \in [\![1, n]\!] | z_j = i\}$.

L'estimateur (π, θ) du maximum de vraisemblance est:

$(\pi, \theta) = \text{argmax}\{\prod_{k=1}^{n} p(x_k, z_k) | \sum_{i=1}^{M} \pi_i = 1; \forall m, \sum_{k=1}^{K} \theta_{m,k} = 1\}$

$(\pi, \theta) = \text{argmax}\{\prod_{i=1}^{M} \prod_{j=1}^{K} \pi_i^{N_{ij}} \theta_{i,j}^{N_{ij}} | \sum_{i=1}^{M} \pi_i = 1; \forall m, \sum_{k=1}^{K} \theta_{m,k} = 1\}$

$(\pi, \theta) = \text{argmax}\{\sum_{i=1}^{M} \sum_{j=1}^{K} N_{i,j} (\log \pi_i + \log(\theta_{i,j}) | \sum_{i=1}^{M} \pi_i = 1; \forall m, \sum_{k=1}^{K} \theta_{m,k} = 1\}$

$(\pi, \theta) = \text{argmax}\{\sum_{i=1}^{M} \sum_{j=1}^{K} N_{i,j} \log \pi_i + \sum_{i=1}^{M} \sum_{j=1}^{K} n_{i,j} \log(\theta_{i,j} | \sum_{i=1}^{M} \pi_i = 1; \forall m, \sum_{k=1}^{K} \theta_{m,k} = 1\}$

$(\pi, \theta) = \text{argmax}\{\sum_{i=1}^{M} \sum_{j=1}^{K} N_{i,j} \log \pi_i + \sum_{i=1}^{M} \sum_{j=1}^{K} N_{i,j} \log \theta_{i,j} | \sum_{i=1}^{M} \pi_i = 1; \forall m, \sum_{k=1}^{K} \theta_{m,k} = 1\}$

Comme la première double somme ne dépend que de π et la deuxième que de θ, on peut calculer les valeurs pour lesquelles elles sont maximales séparément, on a donc:

$$\pi = \text{argmax}\{\sum_{i=1}^{M} \sum_{j=1}^{K} N_{i,j} \log \pi_i | \sum_{i=1}^{M} \pi_i = 1\}$$

$$\pi = \text{argmax}\{\sum_{i=1}^{M} N_i \log \pi_i | \sum_{i=1}^{M} \pi_i = 1\}$$

et

$$\theta = (\theta_1, ...\theta_M)$$

avec $\theta_i = \text{argmax}\{\sum_{j=1}^{K} N_{i,j} \log \theta_{i,j} | \forall m, \sum_{k=1}^{K} \theta_{m,k} = 1\}$

1.1 Calcul du maximum de vraisemblance de π

On utilise la méthode du lagrangien.

On note donc $\mathcal{L}(\pi, \lambda) = \sum_{i=1}^{M} (N_i \log \pi_i) + \lambda((\sum_{i=1}^{M} \pi_i) - 1)$

$\mathcal{L}(\pi, \lambda) = \sum_{i=1}^{M} (N_i \log \pi_i + \lambda \pi_i + \frac{\lambda}{M})$.
Comme le i-ème terme de la somme ne dépend que de π_i, pour maximiser $\mathcal{L}(\pi, \lambda)$ par rapport à π_i il suffit de maximiser chaque i-ème terme de la somme par rapport à π_i.
Or $\frac{\partial}{\partial \pi_i}(N_i \log \pi_i + \lambda \pi_i + \frac{\lambda}{M}) = \frac{N_i}{\pi_i} + \lambda$ donc cette dérivée est nulle *ssi* $\pi_i = -\frac{N_i}{\lambda}$, négative pour $\pi_i > -\frac{N_i}{\lambda}$ et positive sinon.

Ainsi $\frac{\partial(N_i \log \pi_i + \lambda \pi_i + \frac{\lambda}{M})}{\partial \pi_i}$ est maximal en $\pi_i = -\frac{N_i}{\lambda}$. Donc à λ fixé, $\mathcal{L}(\pi, \lambda)$ est maximal pour $\hat{\pi}(\lambda) = (-\frac{N_1}{\lambda}, ..., -\frac{N_M}{\lambda})$.

De plus comme $\min\{\mathcal{L}(\hat{\pi}(\lambda), \lambda) | \lambda \in \Re\} = \max\{\sum_{i=1}^{M} \sum_{j=1}^{K} N_{i,j} \log \pi_i | \sum_{i=1}^{M} \pi_i = 1\}$ et que ce minimum est atteint pour λ tel que $\sum_{i=1}^{M} \hat{\pi}(\lambda)_i = 1$, il est atteint pour $\lambda = -n$ et donc finalement l'estimateur du maximum de vraisemblance de π est

$$\hat{\pi} = (\frac{N_1}{n}, ..., \frac{N_M}{n}).$$

1.2 Maximum de Vraisemblance de θ

De même que précédemment on peut calculer les valeurs de $\theta_i = (\theta_{i,1}, ..., \theta_{i,K})$ pour lesquelles $\sum_{j=1}^{K} N_{i,j} \log \theta_{i,j}$ est maximale avec $\forall m, \sum_{k=1}^{K} \theta_{m,k} = 1$ séparément. Un calcul analogue à celui de la partie précédente montre alors que $\hat{\theta}_i = (\frac{N_{i,1}}{n}, ..., \frac{N_{i,K}}{n})$

2 Classification linéaire

2.1 1.Modèle génératif (LDA)

On suppose que l'on observe N valeurs (x_i, y_i) iid. Calculons le maximum de vraisemblance de $\theta = (\pi, \Sigma, \mu_0, \mu_1)$.

La log-vraisemblance est par définition:

$l(\theta) = \log(\prod_{n=1}^{N} p(y_n|\pi)p(x_n|y_n, \theta))$

$l(\theta) = \sum_{n=1}^{N} \log p(y_n|\pi) + \sum_{n=1}^{N} \log p(x_n|y_n, \Sigma, \mu_0, \mu_1)$

On peut donc maximiser le premier terme indépendamment du deuxième terme.

La valeur de π pour laquelle le premier terme est maximal a déjà été calculé dans l'exercice 1 : l'estimateur du maximum de vraisemblance de π est

$$\hat{\pi} = \frac{\sum_{n=1}^{N} y_n}{N}.$$

Le deuxième terme s'écrit:

$l(\theta) = \sum_{n=1}^{N} \log(\frac{1}{2\pi|\Sigma|^{1/2}}(\exp(-\frac{1}{2}(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1)))^{y_n}(\exp(-\frac{1}{2}(x_n-\mu_0)^T\Sigma^{-1}(x_n-\mu_0)))^{1-y_n}$
$l(\theta) = \sum_{n=1}^{N}(\log(\frac{1}{2\pi|\Sigma|^{1/2}})-\frac{y_n}{2}(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1)-\frac{1-y_n}{2}(x_n-\mu_0)^T\Sigma^{-1}(x_n-\mu_0))$.

2.1.1 Estimation de μ_1 et μ_0

$l(\Sigma, \mu_0, \mu_1)$ ne dépend de μ_1 que par $l_1 = \sum_{n=1}^{N} -\frac{y_n}{2}(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1)$ donc la valeur de μ_1 pour laquelle cette quantité est maximale est également celle qui maximise $l(\Sigma, \mu_0, \mu_1)$ (par rapport à μ_1).
Or $\nabla_{\mu_1} l_1 = 2\sum_{n=1}^{N} y_n(x_n+\mu_1)^T\Sigma^{-1}$ donc ce gradient s'annule pour $\mu_1 = \frac{\sum n=1^N y_n x_n}{\sum_{n=1}^{N} y_n}$.
Ainsi l'estimateur du maximum de vraisemblance de μ_1 est

$$\hat{\mu_1} = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} y_n}$$

et de même

$$\hat{\mu_0} = \frac{\sum_{n=1}^{N}(1-y_n)x_n}{\sum_{n=1}^{N} 1-y_n}.$$

2.1.2 Estimation de Σ

Il faut maximiser $l(\Sigma, \hat{\mu_0}, \hat{\mu_1})$ par rapport à Σ. On pose $\Gamma = \Sigma^{-1}$ et on maximise $l(\Gamma, \hat{\mu_0}, \hat{\mu_1})$ par rapport à Γ.
Le gradient de l en Γ est

$$\nabla_\Gamma l = \frac{1}{2}\sum_{n=1}^{N}(\Gamma^{-1} - y_n(x_n-\hat{\mu_1})(x_n-\hat{\mu_1})^T - (1-y_n)(x_n-\hat{\mu_0})(x_n-\hat{\mu_0})^T)$$

donc ce gradient s'annule pour

$$\Gamma^{-1} = \frac{\sum_{n=1}^{N}(y_n(x_n-\hat{\mu_1})(x_n-\hat{\mu_1})^T + (1-y_n)(x_n-\hat{\mu_0})(x_n-\hat{\mu_0})^T)}{N}.$$

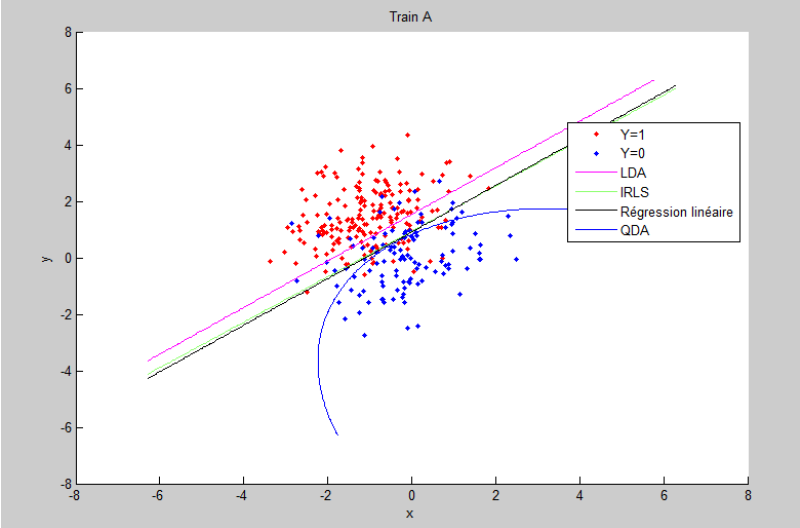
Ainsi l'estimateur du maximum de vraisemblance de Σ est

$$\hat{\Sigma} = \frac{\sum_{n=1}^{N}(y_n(x_n - \hat{\mu_1})(x_n - \hat{\mu_1})^T + (1 - y_n)(x_n - \hat{\mu_0})(x_n - \hat{\mu_0})^T)}{N}.$$

2.2 Comparaison des trois modèles

	LDA	Régression Logistique	Régression Linéaire
A	0.216667	**0.16**	**0.16**
B	0.126667	0.1	**0.08333**
C	0.155	**0.12**	0.16

Figure 1: Training data A

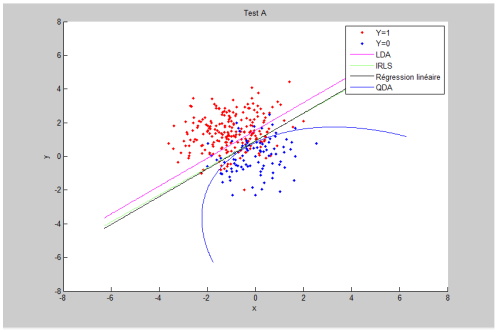


Jeu de données A Le barycentre des données est équidistant des nuages de points $Y = 0$ et $Y = 1$, ce qui explique la bonne performance de la régression linéaire. En effet, plus un point est éloigné du barycentre, plus il a de poids. Ici, les points de classe $Y = 0$ ont donc à peu près le même poids que ceux de classe $Y = 0$. Bien que les deux ensembles de points ne soient pas très bien séparés, la régression linéaire maximise facilement la bonne classification des données de tests.

Entre LDA et IRLS, la pente des droites est très proche, mais la constante varie. IRLS est une méthode d'approximation itérative, qui suit le même principe que LMS (least mean square), qui converge très bien pour la régression linéaire. Comme cette dernière est très efficace ici, il n'est donc pas surprenant qu'IRLS marche si bien sur ces données.

On peut supposer que si il y a un biais dans les données, il sera plus facilement corrigé par un algorithme itératif (IRLS) que par un calcul direct (LDA). Cela peut expliquer le plus faible taux d'erreur pour IRLS que pour LDA.

Figure 2: Test data A



Jeu de tests B La régression linéaire fonctionne très bien sur ce jeu de données, car les points tels que $Y = 1$ sont visuellement proches d'une droite, qui est elle même orthogonal à la moyenne des vecteurs pour $Y = 0$. Ainsi il semble réaliste de trouver un vecteur θ tel que pour le point X tel que $Y = 1$, $\theta X = 1$, et dans l'autre cas $\theta X = 0$

On remarque que les deux autres méthodes sont aussi relativement bonnes. Ce n'est pas surprenant, les données étant visuellement séparées.

Figure 3: Training data B

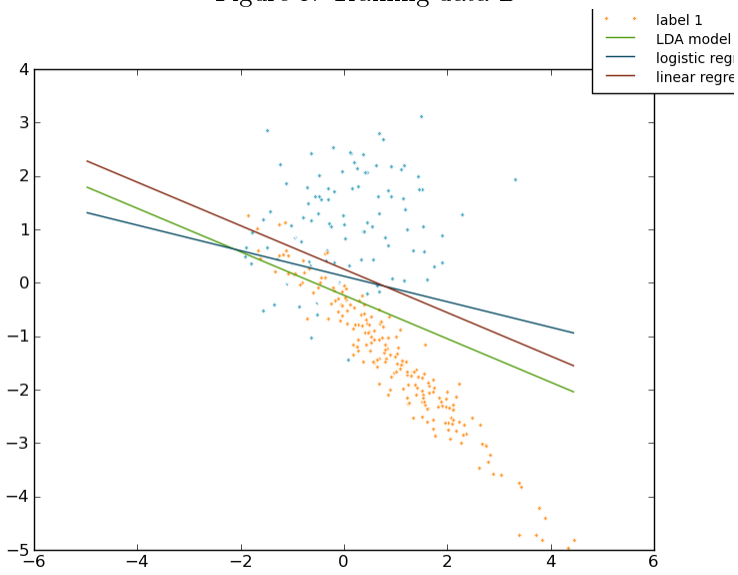


Figure 4: Test data B

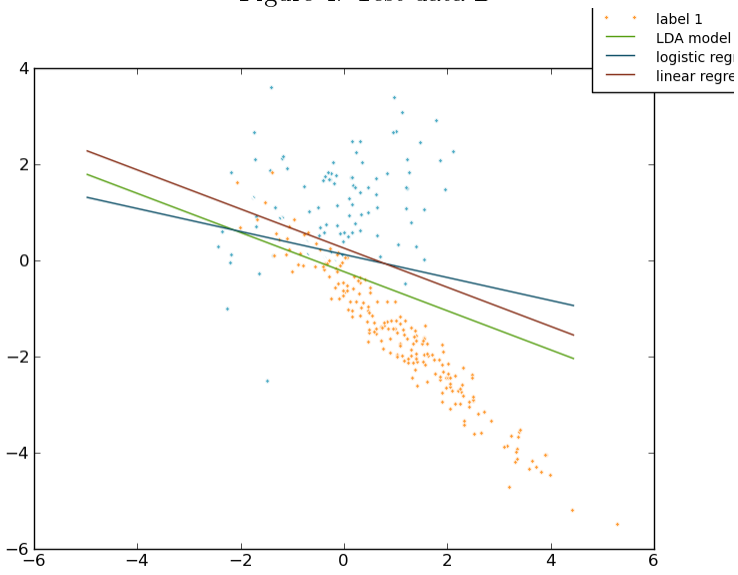
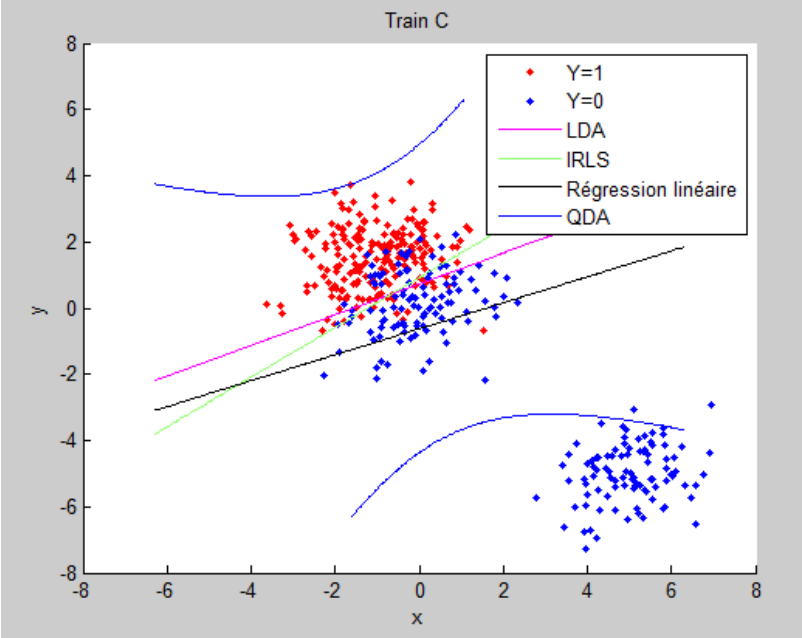


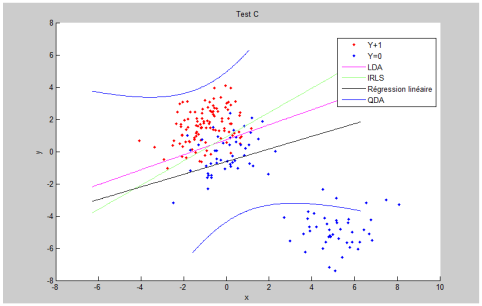
Figure 5: Test data B



Jeu de tests C Les points sont visuellement proche d'une même droite. Si un vecteur θ est tel que $\theta X = 1$ pour $Y = 1$, alors pour une grande partie des points tels que $Y = 0$, on aura aussi $\theta X = 1$. Pour la régression linéaire, la distance au barycentre des points joue énormement (mauvais traitement des données bimodales). Même les points exceptionnellement éloignés du barycentre ont une importance sur le calcul de theta, alors qu'ils représentent plus du bruit. La régression linéaire est très peu robuste.

De même que pour le jeu de données B, IRLS et LDA donnent des résultats tout à fait satisfaisants.

Figure 6: Test data B



2.3 5.Modèle génératif (QDA)

2.3.1 Maximum de vraisemblance

On suppose que l'on observe N valeurs (x_i, y_i) iid. Calculons le maximum de vraisemblance de $\theta = (\pi, \Sigma_0, \Sigma_1, \mu_0, \mu_1)$.

La log-vraisemblance est par définition:

$l(\theta) = \log(\prod_{n=1}^{N} p(y_n|\pi)p(x_n|y_n, \theta))$

$l(\theta) = \sum_{n=1}^{N} \log(p(y_n|\pi)) + \sum_{n=1}^{N} \log(p(x_n|y_n, \Sigma_0, \Sigma_1, \mu_0, \mu_1))$

On peut donc maximiser le premier terme indépendamment du deuxième terme.

D'après l'exercice 1, l'estimateur du maximum de vraisemblance de π est

$$\hat{\pi} = \frac{\sum_{n=1}^{N} y_n}{N}.$$

Le deuxième terme s'écrit:

$l(\Sigma_0, \Sigma_1, \mu_0, \mu_1) = \sum_{n=1}^{N} \log(\frac{1}{2\pi} \frac{1}{|\Sigma_1|^{\frac{y_n}{2}}} \frac{1}{|\Sigma_0|^{\frac{1-y_n}{2}}} (\exp(-\frac{1}{2}(x_n-\mu_1)^T\Sigma_1^{-1}(x_n-\mu_1)))^{y_n}(\exp(-\frac{1}{2}(x_n-\mu_0)^T\Sigma_0^{-1}(x_n-\mu_0)))^{1-y_n}$

$l(\Sigma_0, \Sigma_1, \mu_0, \mu_1) = \sum_{n=1}^{N}(\log(\frac{1}{2\pi}) - \frac{y_n}{2}\log(|\Sigma_1|) - \frac{1-y_n}{2}\log(|\Sigma_0|) - \frac{y_n}{2}(x_n-\mu_1)^T\Sigma_1^{-1}(x_n-\mu_1) - \frac{1-y_n}{2}(x_n-\mu_0)^T\Sigma - 0^{-1}(x_n-\mu_0))$.
Les estimateurs de μ_0 et μ_1 se calculent de la même manière qu'à la question 1.

$l(\Sigma_0, \Sigma_1, \mu_0, \mu_1)$ ne dépend de Σ_1 que par $l_2 = \sum_{n=1}^{N}(-\frac{y_n}{2}\log(|\Sigma_1|)) - \frac{y_n}{2}(x_n-\hat{\mu}_1)^T\Sigma_1^{-1}(x_n-\hat{\mu}_1)$.
On pose $\Gamma_1 = \Sigma_1$ et on maximise l_2 en fonction de Γ_1.
$\nabla_{\Gamma_1} l_2 = \frac{1}{2}(\sum_{n=1}^{N} y_n)\Gamma^{-1} - \frac{1}{2}\sum_{n=1}^{N}(y_n(x_n-\mu_1)(x_n-\mu_1)^T)$
donc ce gradient s'annule pour $\Gamma^{-1} = \frac{\sum_{n=1}^{N}(y_n(x_n-\mu_1)(x_n-\mu_1)^T)}{\sum_{n=1}^{N} y_n}$.
Ainsi l'estimateur du maximum de vraisemblance de Σ_1 est

$$\hat{\Sigma}_1 = \frac{\sum_{n=1}^{N}(y_n(x_n-\mu_1)(x_n-\mu_1)^T)}{\sum_{n=1}^{N} y_n}.$$

De même, on montre que

$$\hat{\Sigma}_0 = \frac{\sum_{n=1}^{N}((1-y_n)(x_n-\mu_0)(x_n-\mu_0)^T)}{\sum_{n=1}^{N} 1-y_n}.$$

2.3.2 Conique d'équation $p(y=1|x) = 0.5$

D'après la formule de Bayes:
$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$

$p(y=1|x) = \frac{\frac{\pi}{|\Sigma_1|}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{\frac{\pi}{|\Sigma_1|}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)) + \frac{\pi}{|\Sigma_0|}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0))}$

$p(y=1|x) = \frac{1}{1+\exp(\log(\frac{1-\pi}{\pi}) + \frac{1}{2}((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}$

Ainsi $p(y=1|x) = 0.5$ *ssi* $\log(\frac{1-\pi}{\pi}) + \frac{1}{2}((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) = 0$ ce qui définit bien une conique.