Ecole Nationale Supérieur Paris-Saclay
Master Mathématiques Vision Apprentissage

Homework 1

DUCHEMIN Quentin & OREISTEIN Pierre

Cours: Probabilistic Graphical Models

October 2018

Resultats des questions théoriques

Pour plus de détails sur le calcul des résultats ci-dessous, on pourra se reporter à l'annexe.

Exercice 1 : Learning in discrete graphical models

Les estimateurs de maximum des vraisemblances sont:

π_m	θ_{km}
$\sum_{i=1}^{n} \frac{e_i^m}{n}$	$\frac{\sum_{i=1}^{n} e_i^m * d_i^k}{\sum_{i=1}^{n} e_i^m}$

Où les e_i, d_i sont les variables "one-hot encoding":

- $e_i = (0, \cdots, 0, \underbrace{1}_{\text{m ième position}}, 0, \cdots, 0), \quad si \;\; z_i = m.$
- $d_i = (0, \cdots, 0, \underbrace{1}_{\text{k ième position}}, 0, \cdots, 0), \quad si \;\; x_i = k$

Exercice 2.1(a) : Estimateurs pour le modèle LDA

Les estimateurs de maximum des vraisemblance des différents paramètres sont :

π	μ_0	μ_1	Σ
$\frac{\sum_{i=1}^{n} y_i}{n}$	$\frac{\sum_{i=1}^{n}(1-y_i)x_i}{\sum_{i=1}^{n}(1-y_i)}$	$\frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i}$	$\frac{1}{n}\sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T + \frac{1}{n}\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T$

Calculons la quantité $p(y=1|x)$:

$$\begin{aligned} p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)} \\ &= \frac{1}{1+\exp\left(x^T \underbrace{\Sigma^{-1}(\mu_0-\mu_1)}_{=c} \underbrace{-\frac{1}{2}\mu_0\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1\Sigma^{-1}\mu_1 + \log\left(\frac{1-\tilde{\pi}}{\tilde{\pi}}\right)}_{=d}\right)} \\ &= \frac{1}{1+\exp\left(x^T c + d\right)} \end{aligned}$$

Le calcul précédent met en évidence le fait que la probabilité $p(y=1|x)$ que l'observation x a été tirée selon la gaussienne 1 s'écrit comme composée d'une fonction affine et d'une sigmoïde. On retrouve donc la formule utilisée lors d'une régression logistique. L'intérêt du modèle LDA par rapport à une régression logistique dans la situation étudiée est que les estimateurs du maximum de vraisemblance ont une forme close que nous venons d'identifier (ce qui n'est pas le cas pour la régression logistique). Cependant, le modèle LDA nécessite l'estimation d'un plus grand nombre de paramètres ce qui peut s'avérer problématique en grande dimension. De même le modèle LDA impose plus de contraintes. Ceci peut amener à plus de plus grandes erreurs sur les données de test si le modèle est très éloigné de la réalité.

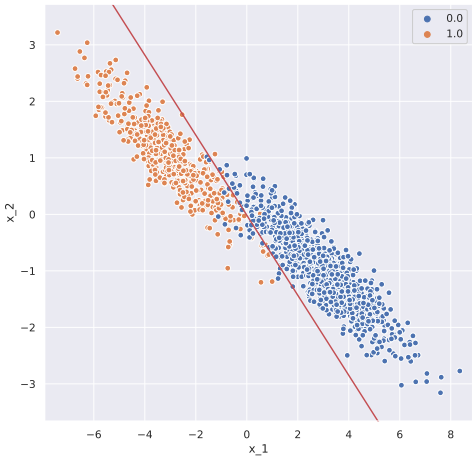
Exercice 2.5(a) : Estimateurs pour le modèle QDA

π	μ_0	μ_1	Σ_0	Σ_1
$\frac{\sum_{i=1}^{n} y_i}{n}$	$\frac{\sum_{i=1}^{n}(1-y_i)x_i}{\sum_{i=1}^{n}(1-y_i)}$	$\frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i}$	$\frac{\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T}{\sum_{j=1}^{n}(1-y_j)}$	$\frac{\sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T}{\sum_{j=1}^{n} y_j}$

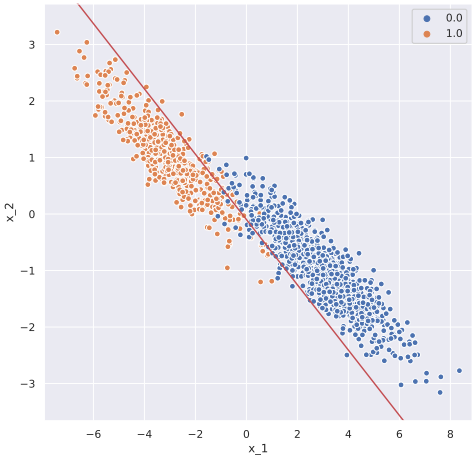
DATASET A

Les graphes présentent les données de test et la courbe définie par l'équation $p(y = 1|x) = 0.5$.

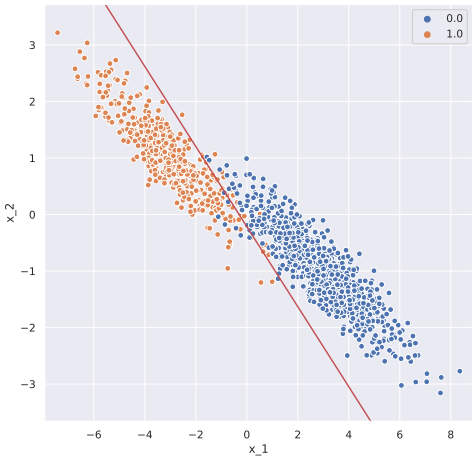
LDA



Régression Logistique



Moindres carrés



QDA

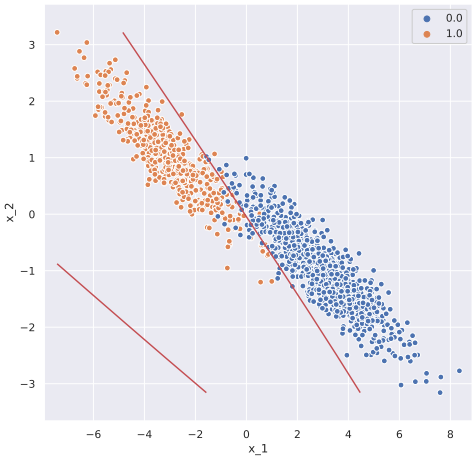


Tableau des erreurs

(% de classifications incorrectes)

	Train	Test
LDA	1.3	2
Logistique	0	3.5
Moindres carrés	2	2.6
QDA	1.3	2.1

Commentaires

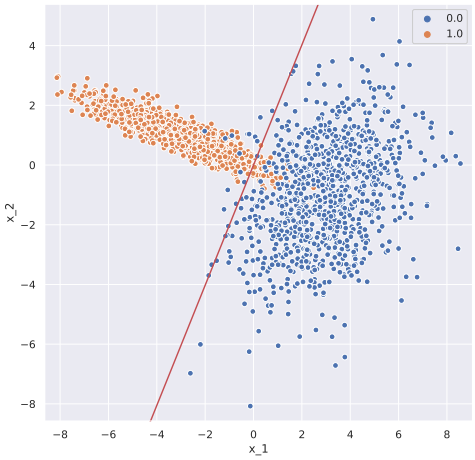
• Tout d'abord, on peut remarquer que la régression logistique a un score idéal sur la base d'entraînement. Ceci est lié au fait que les deux classes du dataset d'entraînement sont séparables. Automatiquement, cette méthode souffre donc d'over-fitting. Son score est donc naturellement en retrait vis à vis des autres méthodes sur les données de test.

• Ensuite, si on observe la répartition des données, celles-ci semblent être générées par deux gaussiennes de même matrice de covariance; seule leur moyenne semblent différentes. De ce fait, ceci pourrait expliquer pourquoi la LDA performe le mieux. En effet, même si la QDA a un score proche en données de test, si les matrices de covariances sont égales, la LDA peut s'approcher plus facilement de la solution; la LDA correspondant alors aux modéles exactes des données.

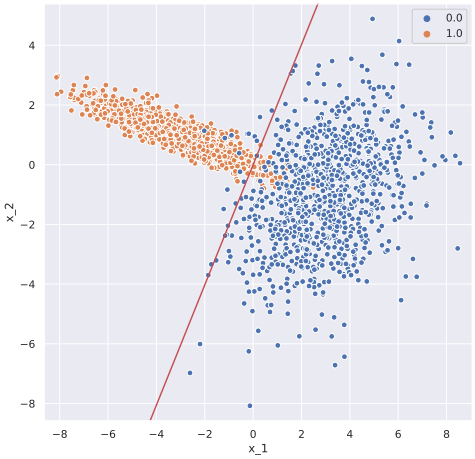
DATASET B

Les graphes présentent les données de test et la courbe définie par l'équation $p(y = 1|x) = 0.5$.

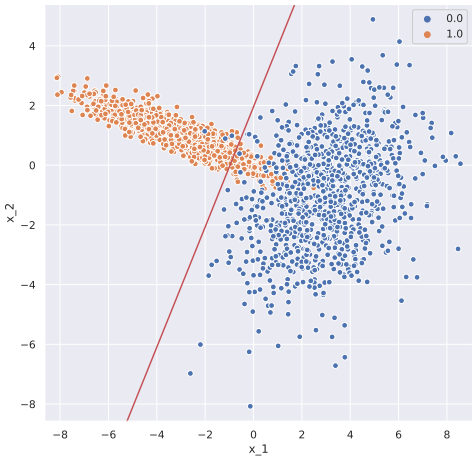
LDA



Régression Logistique



Moindres carrés



QDA

Tableau des erreurs

(% de classifications incorrectes)

	Train	Test
LDA	3	4.1
Logistique	0.3	4.1
Moindres carrés	6.7	7.3
QDA	2.3	2.9

• Le modèle QDA fournit les meilleurs scores sur les bases d'entraînement et de test. Ce constat peut être expliqué par la structure des données. En effet, ici, il semble que les données proviennent de deux gaussiennes avec des matrices de covariance différentes. Ceci correspond donc parfaitement au modèle de la QDA. Il est donc logique que cette méthode performe le mieux.

Commentaires

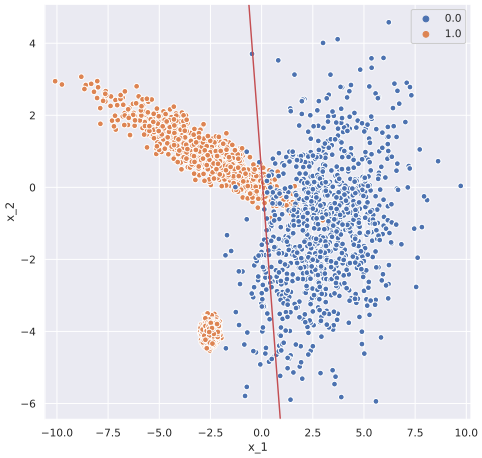
• Les résultats obtenus pour le régression logistique et le modèle LDA sont identiques. Ceci est surement lié à la remarque de 2.1(a) : la quantité $p(y = 1|x)$ dans le cadre du modèle LDA peut s'écrire sous la forme d'une composée d'une fonction affine par une sigmoïde tout comme dans la régression logistique. Ainsi, comme cela semble être le cas ici, l'optimum des paramètres de la régression logistique peut correspondre précisément à ceux de la LDA.

• Enfin, la méthode des moindres carrés fournit des scores sensiblement moins bons que les autres méthodes. En faîte, le pouvoir explicatif de cette méthode est limité (au sens de l'espace des prédicateurs accessibles; des droites) comparé à la QDA par exemple (des coniques). De plus, le modèle supposé par la régression linéaire est assez éloigné des données observées : dans ce modèle on suppose que y prend des valeurs réelles selon une combinaison linéaires de x_1 et x_2, ce qui n'est visiblement pas le cas.

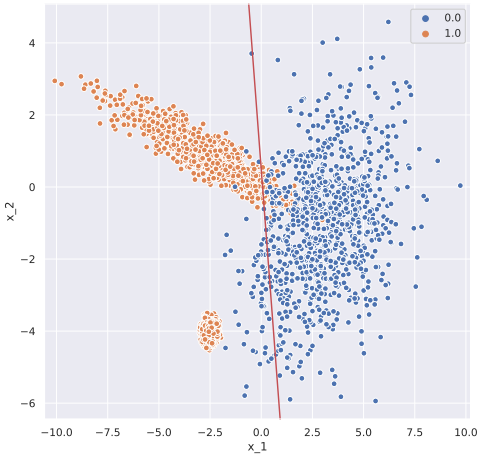
DATASET C

Les graphes présentent les données de test et la courbe définie par l'équation $p(y = 1|x) = 0.5$.

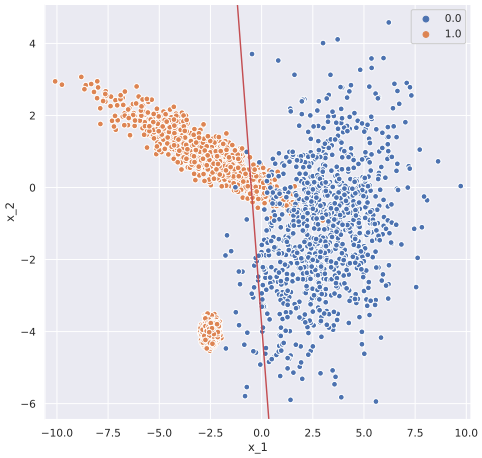
LDA



Régression Logistique



Moindres carrés



QDA

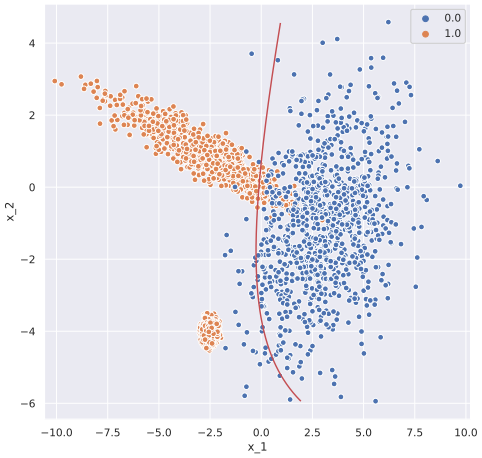


Tableau des erreurs

(% de classifications incorrectes)

	Train	Test
LDA	5.5	4.2
Logistique	5.5	4.2
Moindres carrés	5.7	4.5
QDA	4.5	3.6

• Sur ce jeu de données, nous constatons de meilleurs performances sur la base de test. Une explication possible serait que la base d'entraînement représente fidèlement à une petite échelle la répartition des données réelles. Notons que la taille de la base d'entraînement (400) est supérieure aux deux premiers dataset et peut permettre un moins grand over-fitting.

• Nous remarquons comme sur le dataset précédent que les résultats pour la méthode LDA et logistique sont identiques. Ce résultat expérimental vient donc encore appuyer la comparaison entre les deux modèles menée en 2.1(a).

Commentaires

• Sur les trois datasets étudiés, on peut constater que la méthode QDA fournit globalement les meilleurs résultats sur les données de test. Afin de justifier ce constat, remarquons que l'espace des prédicateurs de la méthode QDA contient celui la méthode LDA (car on relâche simplement la contrainte d'égalité des matrices de covariance).

• Fort de cette étude, nous comprenons que seul un phénomène de surapprentissage pourrait permettre que les scores de la méthode LDA soit meilleur (en test) que ceux de la méthode QDA.

• Notons finalement que les performances du modèle QDA sont ici moins bonnes que sur les deux précédents datasets. Ce phénomène peut être expliqué par une étude des données : la classe rouge présente deux clusters ce qui laisse penser que les données n'ont pas été générées à partir du modèle QDA.

Annexes

1 Annexes Exercice 1: Learning in discrete graphical models

1.1 Maximisation de la vraisemblance

Le modèle étant définit, on cherche donc à estimer les paramètres $\pi = (\pi_m)_{m=1..M}$ et $\theta = (\theta_{mk})_{m=1..M,k=1..K}$. On fait le choix d'estimer ces paramètres via la maximisation de la log-vraisemblance; la vraisemblance correspondant à la probabilité d'observer les observations selon le modèle défini. On a donc effectué le calcul suivant:

$$\max_{\pi,\theta} \; log\left(p(\prod_{i=1}^{n}(x_i,z_i);\pi,\theta)\right) = \max_{\pi,\theta} \; log\left(\prod_{i=1}^{n} p((x_i,z_i);\pi,\theta)\right) \quad \text{car les variables } ((x_i,z_i))_{i=1..n} \text{ sont indépendantes}$$

$$= \max_{\pi,\theta} \sum_{i=1}^{n} \; log\left(p((x_i,z_i)|\pi,\theta)\right)$$

$$= \max_{\pi,\theta} \sum_{i=1}^{n} \; log\left(p(x_i|z_i;\pi,\theta) \; * \; p(z_i;\pi,\theta)\right)$$

Or la loi de x sachant z est parfaitement déterminé par θ
et de même la loi de z est parfaitement déterminé par π

$$= \max_{\pi,\theta} \sum_{i=1}^{n} \; log\left(p(x_i|z_i;\theta) \; * \; p(z_i;\pi)\right)$$

$$= \max_{\pi,\theta} \sum_{i=1}^{n} \; \left(log\left(p(x_i|z_i;\theta)\right) \; + \; log\left(p(z_i;\pi)\right)\right)$$

$$= \max_{\theta} \sum_{i=1}^{n} \; log\left(p(x_i|z_i;\theta)\right) \; + \; \max_{\pi} \sum_{i=1}^{n} \; log\left(p(z_i;\pi)\right)$$

On constate donc que le calcul des estimateurs π et θ est découplé. On va donc pouvoir calculer les estimateurs séparément. Commençons donc par π.

1.2 Calcul de l'estimateur de π pour le maximum de vraisemblance

De façon à simplifier les calculs, on utilisera dans la suite les variables "one-hot encoding":

$$e_i = (0,\cdots,0,\underbrace{1}_{\text{m ième position}},0,\cdots,0), \quad si \;\; z_i = m$$

Pour l'estimation de π on calcule donc:

$$\max_{\pi} \sum_{i=1}^{n} \; log\left(p(z_i;\pi)\right) = \max_{\pi} \sum_{i=1}^{n} \; log\left(\prod_{m=1}^{M} \pi_m^{e_i^m}\right) \quad \text{avec } e_i^m \text{ la m ème composante de la variable "one-hot encoding" } e_i.$$

$$= \max_{\pi} \sum_{i=1}^{n} \sum_{m=1}^{M} \; log\left(\pi_m^{e_i^m}\right)$$

$$= \max_{\pi} \sum_{i=1}^{n} \sum_{m=1}^{M} e_i^m \; log\left(\pi_m\right)$$

1.3 Formulation du problème d'optimisation

De plus, du fait du modèle choisi, on veut obtenir une loi de probabilité pour π. Ainsi, la maximisation de la vraisemblance peut se réécrire comme le problème d'optimisation suivant:

$$\begin{aligned} \min_{\pi} \ & f(\pi) \\ \text{s.c} \quad & h_j(\pi) \leq 0 \quad , \forall j \in [\![1, 2*M]\!] \\ & g(\pi) = 0 \\ & \pi \in \mathbb{R}^M \end{aligned} \tag{1}$$

Avec:

- $f(\pi) = -\sum_{i=1}^{n} \sum_{m=1}^{M} e_i^m \ log\left(\pi_m\right)$
- $h_j(\pi) = \begin{cases} -\pi_m & , \forall j \in [\![1, M]\!], \\ \pi_m - 1 & , \forall j \in [\![M+1, 2*M]\!] \end{cases}$
- $g(\pi) = \sum_{m=1}^{M} \pi_m - 1$

Avant de continuer, on peut d'ores et déjà faire des remarques importantes. Comme 0 n'appartient pas au domaine de f, on en déduit que π a nécessairement des composantes non nulles ($\pi_m \neq 0$ nécessairement). Ainsi, si notre problème admet une solution, ses composantes sont forcément non nulles.
De même, si on suppose $\exists m \in [\![1, M]\!], \pi_m = 1$ alors nécessairement toutes les autres composantes de π sont nulles au vu de la dernière contrainte de notre problème d'optimisation. Or du fait de la remarque précédente, ceci ne peut être une solution de notre problème. On en déduit donc que nécessairement $\forall m \in [\![1, M]\!], 0 < \pi_m < 1$. Les contraintes h_j sont donc nécessairement non actives ($\forall j, \quad h_j(\pi) < 0$).

On constate alors notre problème présente plusieurs propriétés intéressantes:

- f est une fonction convexe différentiable.
- les fonctions h_j sont toutes convexes et différentiables.
- g est une fonction affine.

Enfin, on peut remarquer que les conditions de Slater sont vérifiées. En effet, on peut remarquer que le vecteur $\pi = \left(\frac{1}{M}, \cdots, \frac{1}{M}\right)$ appartient à l'intérieur de l'espace des contraintes.

1.4 Formulation du Lagrangien

Les remarques précédentes nous poussent donc à formuler le lagrangien de ce problème:

$$\mathcal{L}(\pi, \lambda, \mu) = f(\pi) + \lambda * g(\pi) + \sum_{j=1}^{2*M} \mu_j * h_j(\pi)$$

Avec:

- $\lambda \in \mathbb{R}$,
- $\mu_j \in \mathbb{R}^+ \quad , \forall j \in [\![1, 2*M]\!]$,
- $\sum_{j=1}^{2*M} \mu_j * h_j(\pi) = 0$

De même, vu que les $\forall m \in [\![1, 2*M]\!], \ h_j(\pi) < 0$, on en déduit que nécessairement $\forall j \in [\![1, 2*M]\!], \ \mu_j = 0$.

Ainsi, si on parvient à trouver un point selle $(\hat{\pi}, \hat{\lambda})$ du langragien, on aura prouvé que $\hat{\pi}$ est une solution optimale de notre problème d'après le théorème de KKT.

1.5 Résolution

Pour trouver ce point selle, on peut chercher les points critiques du lagrangien. Ainsi:
Dérivons par rapport aux π_m:

$$\begin{aligned}\frac{\partial \mathcal{L}(\pi,\lambda)}{\partial \pi_m} = 0 &\iff -\lambda + \sum_{i=1}^{n} \frac{e_i^m}{\pi_m} = 0 \\ &\iff \frac{1}{\pi_m}\sum_{i=1}^{n} e_i^m = \lambda \\ &\iff \pi_m = \frac{1}{\lambda}\sum_{i=1}^{n} e_i^m\end{aligned}$$

De même dérivons par rapport à λ

$$\frac{\partial \mathcal{L}(\pi,\lambda)}{\partial \lambda} = 0 \iff \sum_{m=1}^{M} \pi_m - 1 = 0$$

Or on cherche un point critique, on va donc chercher les points à l'intersection de ces deux droites. On injecte le résultat précédent:

$$\begin{aligned}\frac{\partial \mathcal{L}(\hat{\pi},\hat{\lambda})}{\partial \lambda} = 0 &\iff \sum_{m=1}^{M} \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} e_i^m = 1 \\ &\iff \frac{1}{\hat{\lambda}} \underbrace{\sum_{i=1}^{n} \underbrace{\sum_{m=1}^{M} e_i^m}_{=1}}_{=n} = 1 \\ &\iff \hat{\lambda} = n\end{aligned}$$

De la même manière, on injecte le résultat précédent dans la première équation:

$$\frac{\partial \mathcal{L}(\hat{\pi},\hat{\lambda})}{\partial \pi_m} = 0 \iff \hat{\pi}_m = \sum_{i=1}^{n} \frac{e_i^m}{n}$$

Il est facile de se persuader que le point trouver est un point selle du fait de la convexité du lagrangien. On a donc trouvé un point selle du lagrangien. On a donc trouver une solution optimal à notre problème d'optimisation. On a donc réussi à calculer l'estimateur du maximum de vraisemblance pour π.

1.6 Calcul de l'estimateur de θ pour le maximum de vraisemblance

De la même manière que précédemment, on va utiliser les variables "one-hot encoding":

$$d_i = (0, \cdots, 0, \underbrace{1}_{\text{k ième position}}, 0, \cdots, 0), \quad si \;\; x_i = k$$

Calculons maintenant l'estimateur du maximum de vraisemblance pour θ:

$$\begin{aligned}
\max_{\theta} \sum_{i=1}^{n} log\left(p(x_i|z_i;\theta)\right) &= \max_{\theta} \sum_{i=1}^{n} log\left(\prod_{m=1}^{M} \left(p(x_i|z_i=m;\theta)\right)^{e_i^m}\right) \\
&= \max_{\theta} \sum_{i=1}^{n}\sum_{m=1}^{M} log\left(p(x_i|z_i=m;\theta)^{e_i^m}\right) \\
&= \max_{\theta} \sum_{i=1}^{n}\sum_{m=1}^{M} log\left(\prod_{k=1}^{K} \left(p(x_i=k|z_i=m;\theta)\right)^{d_i^k * e_i^k}\right) \\
&= \max_{\theta} \sum_{i=1}^{n}\sum_{m=1}^{M}\sum_{k=1}^{K} d_i^k * e_i^m \ * \ log\left(p(x_i=k|z_i=m;\theta)\right) \\
&= \max_{\theta} \sum_{i=1}^{n}\sum_{m=1}^{M}\sum_{k=1}^{K} d_i^k * e_i^m \ * \ log\left(\theta_{m,k}\right)
\end{aligned}$$

On s'aperçoit alors que les $\theta_{m,.}$ sont découplés. On va donc chercher leurs estimateurs séparément. Dans la suite, on considère donc m fixé.

1.7 Formulation du problème d'optimisation

Du fait du modèle choisit, on veut obtenir pour tout m, une loi de probabilité pour $\theta_{m,.}$. Ainsi, pour tout m, la maximisation de la vraisemblance peut se réécrire comme le problème d'optimisation suivant:

$$\begin{aligned}
&\min_{\pi} \ f(\theta_{m,.}) \\
&\text{s.c} \quad h_j(\theta_{m,.}) \leq 0 \quad ,\forall j \in [\![1, 2*K]\!] \\
&\qquad g(\theta_{m,.}) = 0 \\
&\qquad \theta_{m,.} \in \mathbb{R}^K
\end{aligned} \tag{2}$$

Avec:

- $f(\theta_{m,.}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \ d_i^k * e_i^m \ * \ log\left(\theta_{m,k}\right)$
- $h_j(\theta_{m,.}) = \begin{cases} -\theta_{m,k} & ,\forall j \in [\![1, K]\!], \\ \theta_{m,k} - 1 & ,\forall j \in [\![K+1, 2*K]\!] \end{cases}$
- $g(\theta_{m,.}) = \sum_{k=1}^{K} \theta_{m,k} - 1$

Avant de continuer, on peut d'ores et déjà faire les mêmes remarques que dans la section précédentes. Du fait des différentes contraintes et du domaine de f, les contraintes h_j sont donc nécessairement non actives ($\forall j, \quad h_j(\theta_{m,.}) < 0$).

De plus et de la même manière, on constate alors que notre problème présente plusieurs propriétés intéressantes:

- f est une fonction convexe différentiable en $\theta_{m,.}$.
- les fonctions h_j sont toutes convexes et différentiables.
- g est une fonction affine.

Enfin, on peut remarquer que les conditions de Slater sont là aussi vérifiées. En effet, on peut remarquer que le vecteur $\theta_{m,.} = \left(\frac{1}{K}, \cdots, \frac{1}{K}\right)$ appartient à l'intérieur de l'espace des contraintes.

1.8 Formulation du Lagrangien

Les remarques précédentes nous poussent donc à formuler le lagrangien de ce problème:

$$\mathcal{L}(\theta_{m,.}, \lambda, \mu) = f(\theta_{m,.}) + \lambda * g(\theta_{m,.}) + \sum_{j=1}^{2*K} \mu_j * h_j(\theta_{m,.})$$

Avec:

- $\lambda \in \mathbb{R}$,
- $\mu_j \in \mathbb{R}^+ \quad , \forall j \in [\![1, 2 * K]\!]$,
- $\sum_{j=1}^{2*K} \mu_j * h_j(\theta_{m,.}) = 0$

De même, vu que les $\forall m \in [\![1, 2 * K]\!], \quad h_j(\theta_{m,.}) < 0$, on en déduit que nécessairement $\forall j \in [\![1, 2 * K]\!], \quad \mu_j = 0$.

Ainsi, si on parvient à trouver un point selle $(\hat{\theta}_{m,.}, \hat{\lambda})$ du langragien, on aura prouvé que $\hat{\theta}_{m,.}$ est une solution optimale de notre problème d'après le théorème de KKT.

1.9 Résolution

Pour trouver ce point selle, on peut chercher les points critiques du lagrangien. Ainsi:
Dérivons par rapport aux $\theta_{m,.}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta_{m,.}, \lambda)}{\partial \theta_{m,k}} = 0 &\iff -\lambda + \sum_{i=1}^{n} \frac{e_i^m * d_i^k}{\theta_{m,k}} = 0 \\
&\iff \frac{1}{\theta_{m,k}} \sum_{i=1}^{n} e_i^m * d_i^k = \lambda \\
&\iff \theta_{m,k} = \frac{1}{\lambda} \sum_{i=1}^{n} e_i^m * d_i^k
\end{aligned}$$

De même dérivons par rapport à λ

$$\frac{\partial \mathcal{L}(\theta_{m,.}, \lambda)}{\partial \lambda} = 0 \iff \sum_{k=1}^{K} \theta_{m,k} - 1 = 0$$

Or on cherche un point critique, on va donc chercher les points à l'intersection de ces deux droites. On injecte le résultat précédent:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\hat{\theta}_{m,.}, \hat{\lambda})}{\partial \lambda} = 0 &\iff \sum_{k=1}^{K} \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} e_i^m * d_i^k = 1 \\
&\iff \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} \sum_{k=1}^{K} * d_i^k = 1 \\
&\iff \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} e_i^m \underbrace{\sum_{k=1}^{K} d_i^k}_{=1} = 1 \\
&\iff \hat{\lambda} = \sum_{i=1}^{n} e_i^m
\end{aligned}$$

De la même manière, on injecte le résultat précédent dans la première équation:

$$\frac{\partial \mathcal{L}(\hat{\theta}_{m,.}, \hat{\lambda})}{\partial \theta_{m,k}} = 0 \Longleftrightarrow \hat{\theta}_{m,k} = \frac{\sum_{i=1}^{n} e_i^m * d_i^k}{\sum_{i=1}^{n} e_i^m}$$

Il est facile de se persuader que le point trouver est un point selle du fait de la convexité du lagrangien. On a donc trouvé un point selle du lagrangien. On a fait nos calculs à m fixé. Cependant, ils restent valables quelque soit m. On a donc trouver une solution optimale à notre problème d'optimisation. Ainsi, on a trouvé notre estimateur de maximum de vraisemblance θ.

2 Annexes Exercice 2.1a: Modèle LDA

Afin d'éviter toute confusion de notation, nous considérons que $y \sim Bernoulli(\tilde{\pi})$.
La log-vraisemblance des observations s'écrit alors :

$$\begin{aligned}
\log(VS) &= \log\left(\prod_{i=1}^{n} p((x_i, y_i))\right) \\
&= \log\left(\prod_{i=1}^{n} [p(x_i|y_i=1)p(y_i=1)]^{y_i} \times [p(x_i|y_i=0)p(y_i=0)]^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \left[\log(p(x_i|y_i=1) + \log(p(y_i=1))\right] + (1-y_i)\left[\log(p(x_i|y_i=0)) + \log(p(y_i=0))\right] \\
&= \sum_{i=1}^{n} y_i \left[\log(p(x_i|y_i=1) + \log(\tilde{\pi})\right] + (1-y_i)\left[\log(p(x_i|y_i=0)) + \log(1-\tilde{\pi})\right]
\end{aligned}$$

Comme par hypothèse $x|\{y=j\} \sim \mathcal{N}(\mu_j, \Sigma)$, on a pour tout $i \in [\![1, n]\!]$ et tout $j \in \{0, 1\}$:

$$p(x_i|y_i=j) = \frac{1}{2\pi|\Sigma|} \exp\left(-\frac{1}{2}(x_i-\mu_j)^T\Sigma^{-1}(x_i-\mu_j)\right)$$

Ainsi :

$$\begin{aligned}
\log(VS) = &\sum_{i=1}^{n} y_i \left[-\log(2\pi|\Sigma|) - \frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) + \log(\tilde{\pi})\right] \\
&+ (1-y_i)\left[-\log(2\pi|\Sigma|) - \frac{1}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0) + \log(1-\tilde{\pi})\right] \\
= &\sum_{i=1}^{n} y_i \left[\log(2\pi|\Lambda|) - \frac{1}{2}(x_i-\mu_1)^T\Lambda(x_i-\mu_1) + \log(\tilde{\pi})\right] \\
&+ (1-y_i)\left[\log(2\pi|\Lambda|) - \frac{1}{2}(x_i-\mu_0)^T\Lambda(x_i-\mu_0) + \log(1-\tilde{\pi})\right]
\end{aligned}$$

où nous avons effectué le changement de variable $\Sigma^{-1} = \Lambda$. Ce changement de variable permet de passer d'une fonction non concave en Σ à une fonction concave en Λ.

Notant $\theta = (\tilde{\pi}, \mu_0, \mu_1, \Lambda)$ les paramètres de notre modèle, l'estimateur du maximum de vraisemblance (noté θ^{MLE}) est défini par :

$$\begin{aligned}
(MLE) \quad \theta^{MLE} \in \quad & \underset{\theta}{\text{argmax}} \ \log(VS) \\
\text{s.c} \quad & 0 \leq \tilde{\pi} \leq 1 \\
& \Lambda \in S_2^{++}(\mathbb{R})
\end{aligned}$$

Nous cherchons à utiliser le théorème de Karush-Kuhn et Tucker (KKT). Nous vérifions les conditions d'application du théroème :

- Nous maximisons une fonction concave sur un domaine convexe défini par les contraintes (noté $\mathcal{D}$), donc le problème d'optimisation est convexe.
- De plus, il est immédiat de voir qu'il existe un point réalisable contenu strictement dans $\mathcal{D}$. La condition de qualification de Slater est donc vérifiée.

Ainsi, θ est une solution optimale au problème (MLE) si et seulement s'il existe $(\mu, \lambda) \in (\mathbb{R}_+)^2$ vérifiant les conditions de KKT.
La condition des écarts complémentaires donne : $\tilde{\pi}\mu = \tilde{\pi}\lambda = 0$. Comme les valeurs $\tilde{\pi} = 0$ et $\tilde{\pi} = 1$ n'appartiennent pas au domaine de f, on en déduit que la valeur optimale du paramètre $\tilde{\pi}$ appartient à $]0, 1[$. Nous en déduisons que $\mu = \lambda = 0$.
La condition d'annulation du gradient du lagrangien par rapport à θ s'écrit :

- pour $\tilde{\pi}$:

$$\frac{\partial \log(VS)}{\partial \tilde{\pi}} = \frac{\sum_{i=1}^n y_i}{\tilde{\pi}} - \frac{\sum_{i=1}^n 1 - y_i}{1 - \tilde{\pi}} \underbrace{-\mu + \lambda}_{=0} = 0$$

  Ce qui amène à la valeur : $\tilde{\pi} = \frac{\sum_{i=1}^n y_i}{n}$.

- pour μ_1:

$$\frac{\partial \log(VS)}{\partial \mu_1} = -2 \sum_{i=1}^n y_i \Sigma^{-1}(x_i - \mu_1) = 0$$

  Ce qui amène à la valeur : $\mu_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n y_i}$.

- pour μ_0:

  Un calcul strictement analogue au précédent permet d'obtenir l'estimateur de maximum de vraisemblance pour μ_0 sous la forme : $\mu_0 = \frac{\sum_{i=1}^n (1 - y_i) x_i}{\sum_{i=1}^n (1 - y_i)}$.

- pour Σ:

  L'estimateur du maximum de vraisemblance de Σ est donnée par :

$$\begin{aligned} \underset{\Lambda \in S_2^{++}(\mathbb{R})}{\operatorname{argmax}} & \sum_{i=1}^n y_i \left[\log(det(\Lambda)) + \frac{1}{2} Tr\left(\Lambda (x_i - \mu_1)(x_i - \mu_1)^T \right) \right] \\ & + \sum_{i=1}^n (1 - y_i) \left[\log(det(\Lambda)) - \frac{1}{2} Tr\left(\Lambda (x_i - \mu_0)(x_i - \mu_0)^T \right) \right] \end{aligned}$$

Nous cherchons donc à maximiser une fonction concave différentiable sur un domaine convexe, ouvert et non vide. Une condition nécessaire et suffisante d'optimalité est l'annulation du gradient de cette fonction de Λ :

$$\sum_{i=1}^n \frac{1}{2} y_i \Lambda^{-1} + \sum_{i=1}^n \frac{1}{2} (1 - y_i) \Lambda^{-1} - \sum_{i=1}^n \frac{1}{2} y_i (x_i - \mu_1)(x_i - \mu_1)^T - \sum_{i=1}^n \frac{1}{2} y_i (x_i - \mu_0)(x_i - \mu_0)^T = 0$$

Finalement, l'estimateur du maximum de vraisemblance de Σ est :

$$\Sigma^{MLE} = (\Lambda^{MLE})^{-1} = \frac{1}{n} \sum_{i=1}^n y_i (x_i - \mu_1)(x_i - \mu_1)^T + \frac{1}{n} \sum_{i=1}^n (1 - y_i)(x_i - \mu_0)(x_i - \mu_0)^T$$

Calculons la quantité $p(y = 1|x)$:

$$
\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\
&= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)} \\
&= \frac{1}{1+\frac{p(x|y=0)}{p(x|y=1)}\frac{p(y=0)}{p(y=1)}} \\
&= \frac{1}{1+\frac{1-\tilde{\pi}}{\tilde{\pi}}\exp\left(-\frac{1}{2}(x-\mu_0)\Sigma^{-1}(x-\mu_0)^T+\frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)^T\right)} \\
&= \frac{1}{1+\exp\left(x^T\underbrace{\Sigma^{-1}(\mu_0-\mu_1)}_{=c}\underbrace{-\frac{1}{2}\mu_0\Sigma^{-1}\mu_0+\frac{1}{2}\mu_1\Sigma^{-1}\mu_1+\log\left(\frac{1-\tilde{\pi}}{\tilde{\pi}}\right)}_{=d}\right)} \\
&= \frac{1}{1+\exp\left(x^Tc+d\right)}
\end{aligned}
$$

Le calcul précédent met en évidence le fait que la probabilité $p(y=1|x)$ que l'observation x a été tirée selon la gaussienne 1 s'écrit comme composée d'une fonction affine et d'une sigmoïde. On retrouve donc la formule utilisée lors d'une régression logistique. L'intérêt du modèle LDA par rapport à une régression logistique dans la situation étudiée est que les estimateurs du maximum de vraisemblance ont une forme close que nous venons d'identifier (ce qui n'est pas le cas pour la régression logistique). Cependant, le modèle LDA nécessite l'estimation d'un plus grand nombre de paramètres ce qui peut s'avérer problématique en grande dimension.

3 Exercice 2.5a: Modèle QDA

Nous reprenons le modèle de la question 2.1(a) en retirant la contrainte d'égalité entre les matrices de covariances des deux gaussiennes. La log-vraisemblance possède alors la forme suivante :

$$
\begin{aligned}
\log(VS) &= \sum_{i=1}^{n} y_i\left[-\log(2\pi|\Sigma_1|)-\frac{1}{2}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)+\log(\tilde{\pi})\right] \\
&\quad +(1-y_i)\left[-\log(2\pi|\Sigma_0|)-\frac{1}{2}(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)+\log(1-\tilde{\pi})\right] \\
&= \sum_{i=1}^{n} y_i\left[\log(2\pi|\Lambda_1|)-\frac{1}{2}(x_i-\mu_1)^T\Lambda_1(x_i-\mu_1)+\log(\tilde{\pi})\right] \\
&\quad +(1-y_i)\left[\log(2\pi|\Lambda_0|)-\frac{1}{2}(x_i-\mu_0)^T\Lambda_0(x_i-\mu_0)+\log(1-\tilde{\pi})\right]
\end{aligned}
$$

où nous avons effectué les changements de variable $\Sigma_1^{-1}=\Lambda_1$ et $\Sigma_0^{-1}=\Lambda_0$. Ce passage nous permet d'obtenir une fonction concave en Λ_1 et en Λ_0.
Un raisonnement strictement analogue à la question 2.1a est alors applicable. Les estimateurs de π, μ_0 etμ_1 restent identiques au cas précédent. Leur formes sont données en page 11.

L'estimateur du maximum de vraisemblance de Λ_1 est donnée par :

$$
\underset{\Lambda\in S_2^{++}(\mathbb{R})}{\operatorname{argmax}} \sum_{i=1}^{n} y_i\left[\log(det(\Lambda_1))+\frac{1}{2}Tr\left(\Lambda_1(x_i-\mu_1)(x_i-\mu_1)^T\right)\right]
$$

Nous cherchons donc à maximiser une fonction concave différentiable sur un domaine convexe, ouvert et non vide. Une condition nécessaire et suffisante d'optimalité est l'annulation du gradient de cette fonction de Λ_1 :

$$\sum_{i=1}^{n} \frac{1}{2} y_i \Lambda_1^{-1} - \sum_{i=1}^{n} \frac{1}{2} y_i (x_i - \mu_1)(x_i - \mu_1)^T = 0$$

Finalement, l'estimateur du maximum de vraisemblance de Σ_1 est :

$$\Sigma_1^{MLE} = (\Lambda_1^{MLE})^{-1} = \frac{\sum_{i=1}^{n} y_i (x_i - \mu_1)(x_i - \mu_1)^T}{\sum_{j=1}^{n} y_j}$$

Un raisonnement strictement analogue permet d'obtenir l'estimateur du maximum de vraisemblance pour Σ_0:

$$\Sigma_0^{MLE} = (\Lambda_0^{MLE})^{-1} = \frac{\sum_{i=1}^{n} (1 - y_i)(x_i - \mu_0)(x_i - \mu_0)^T}{\sum_{j=1}^{n} (1 - y_j)}$$