Master 2 IAAA - Machine Learning ECM 3A
Cours de Data Science
TD 6 - 2019-2020

T. Artières and QARMA

Octobre 2019

I. Maximum de vraisemblance et densité Gaussienne

On souhaite prédire estimenr la densité sous jacente à un ensemble de données observées $D = \{x^1, ..., x^N, \forall i \in [\![1; N]\!] \, x^i \in \mathbb{R}^d\}$. On suppose les données i.i.d. et la densité Gaussienne de paramètres μ et σ, $\mathcal{N}_{\mu,\Sigma}$.

(1) Exprimer la log-vraisemblance des données.

(2) Quelle est la solution au sens du maximum de vraisemblance de l'estimation de μ et de Σ ?

II. EM et mélange Gaussien

On considère un ensemble de données observées $D = \{x^1, ..., x^N, \forall i \in [\![1; N]\!], x^i \in \mathbb{R}^d\}$. On suppose que la densité sous jacente aux données est une loi de type mélange Gaussien à K composantes et on souhaite estimer ses paramètres que l'on rassemble sous une variable notée θ.

(1) Exprimer la densité $p(x)$ en introduisant les notations nécessaires. Commentez sur la nature de chacune des quantités introduites.

(2) On considère qu'à l'itération t, on a obtenu les paramètres θ_t. On note h_i la variable aléatoire indiquant la composante ($h^i \in [\![1; N]\!]$) par laquelle a été produite l'observation i. L'étape E consiste à calculer les quantités $\alpha_j^i = q(h^i|x^i) = p(h^i = j|x^i, \theta_t)$ (en omettant l'index t sur les α). Donnez l'expression de ces quantités en faisant apparaître les paramètres des différentes composantes Gaussiennes (moyenne, covariance etc)

(3) Calculez le gradient de la fonction auxiliaire $l(\theta|\theta_t)$ par rapport à la moyenne d'une des lois composantes. Commentez le résultat obtenu.

(4) Calculez le gradient de la fonction auxiliaire $l(\theta|\theta_t)$ par rapport à une probabilité a priori d'une loi composante. Il faut, pour prendre en compte la contrainte de distribution, utiliser le Lagrangien.

III. EM pour PLSA

On considère le problème de l'apprentissage d'un modèle de type PLSA (Probabilistic Latent Semantic Analysis) avec l'algorithme EM. Le modèle PLSA a été proposé pour fouiller des collections de données textuelles et peut être appliqué à d'autres types de données pourvu qu'elles soient discrètes.

On considère ici que l'on dispose d'un corpus de document $D = \{d_i, i = 1, ..., N\}$. Tous les mots de tous les documents forment un vocabulaire V de taille M, $V = \{w_1, ..., w_M\}$. Un document d_i est constitué d'un ensemble de mots (on ne tient pas compte du caractère séquentiel des mots d'un document dans la suite). Seuls la présence et le nombre d'occurrences de chaque mot du vocabulaire dans un document nous intéresse. On note $n(d_i, w_j)$ le nombre d'occurrences du mot w_j dans le document d_i.

Les documents du corpus abordent globalement un ensemble de sujets (par ex. sport, actualités, économie...), chaque document peut aborder plusieurs sujets (par ex. un texte parlant de l'économie du sport). On s'intéresse à découvrir automatiquement l'ensemble des sujets abordés dans le corpus D et les sujets abordés dans chacun des documents. On considérera dans la suite le nombre de sujets fixé $Z = \{z_1, ...z_K\}$, et chaque sujet z_k est vu comme une distribution de probabilité sur V. On note $p(w_j|z_k)$ la probabilité du mot w_j conditionnellement au sujet z_k.

Dans la modélisation PLSA chaque document est vu comme une distribution de probabilité sur les sujets : $(p(z_k|d_i), k = 1..K)$. Tous les mots d'un document sont supposés être générés indépendamment les uns des autres. La probabilité d'un mot w_j dans un document d_i s'écrit: $p(w_j|d_i) = \sum_k p(w|z = k)p(z = k|d^i)$. La log-vraisemblance de l'ensemble (noté W_i) des mots d'un documents d_i s'écrit:

$$\begin{aligned}
\log p(W_i|\ d_i) &= \sum_{j=1}^{M} n(d_i, w_j) \log p(w_j|d_i) \\
&= \sum_{j=1}^{M} n(d_i, w_j) \log \sum_z p(w_j, z|d_i) \\
&= \sum_{j=1}^{M} n(d_i, w_j) \log \sum_z p(w_j|z)p(z|d_i)
\end{aligned}$$

La log-vraisembleance d'un corpus entier est donnée par :

$$ll(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \sum_z p(w_j|z)p(z|d_i)$$

(1) On souhaite apprendre le modèle PLSA à partir du corpus D avec l'algorithme EM. Quels sont les paramètres du modèle ? Quel en est le nombre ?

(2) Exprimez la fonction auxiliaire Q.

(3) En formant le Lagrangien à partir des contraintes sur les paramètres (ce sont des paramètres de distributions), en déduire les formules de ré-estimation pour les paramètres du modèle.

IV. Rappels mathématiques

(1) Dérivation de formes algébriques
Soit $A \in \mathbb{R}^{k \times k}$, $v \in \mathbb{R}^k$. Alors:

$$\frac{\partial v^T a}{\partial v} = \frac{\partial a^T v}{\partial v} = a$$
$$\frac{\partial v^T A v}{\partial v} = (A + A^T)v$$
$$\Rightarrow \text{ Si } A \text{ est symétrique } \frac{\partial v^T A v}{\partial v} = 2Av$$
$$\frac{\partial \log(det(M))}{\partial M} = M^{-1}$$
$$\frac{\partial (Av)^T (Av)}{\partial v} = 2A^T Av$$

(2) Rappel d'optimisation Lagrangienne avec contrainte d'égalité
Pour optimiser (e.g. minimiser) une fonction $C(w)$ sous la constrainte que $g(w) = 0$ on forme le Lagrangien $L(w, \beta) = C(w) + \beta f(w)$ que l'on minimise par rapport à w et que l'on maximise par rapport à β. On a :

$$\min_w C(w)$$
$$\text{s.t. } g(w) = 0$$
$$\Leftrightarrow$$
$$\min_w \max_\beta C(w) + \beta g(w)$$

(3) Rappel d'optimisation Lagrangienne avec contrainte d'inégalité
Pour optimiser (e.g. minimiser) une fonction $C(w)$ sous la constrainte que $f(w) \leq 0$ on forme le Lagrangien $L(w, \beta) = C(w) - \beta(f(w))$ que l'on minimise par rapport à w et que l'on maximise par rapport à β. On a :

$$\min_w C(w)$$
$$\text{s.t. } f(w) \geq 0$$
$$\Leftrightarrow$$
$$\min_w \max_\beta C(w) - \beta f(w)$$
$$\text{s.t. } \beta \geq 0$$

(4) Optimisation Lagrangienne avec contraintes multiples
Pour optimiser (e.g. minimiser) une fonction $C(w)$ sous des constraintes d'inégalité $\{f_i(w) \leq 0\}_i$ et des contraintes d'égalité $\{g_j(w) = 0\}_j$ on forme le Lagrangien

$L(w, (\beta_i)_i, (\gamma_j)_j) = C(w) - \sum_i \beta_i f_i(w) + \sum_j \gamma_j g_j(w)$ que l'on minimise par rapport à w et que l'on maximise par rapport aux βet aux γ. On a :

$$\begin{aligned}
&\min_w C(w) \\
&\quad \text{s.t. } \forall i, f_i(w) \geq 0 \text{ Et } \forall j, g_j(w) = 0 \\
&\quad \Leftrightarrow \\
&\min_w \max_{\beta,\gamma} C(w) - \sum_i \beta_i f_i(w) + \sum_j \gamma_j g_j(w) \\
&\quad \text{s.t. } \beta \geq 0
\end{aligned}$$