Doan, Hudelot, Ouerdane, Tami

MODÈLES DE RECHERCHE : MODÈLE PROBABILISTE

Exercices

Exercice 1

On considère la fonction h qui intervient dans le calcul du score OKAPI BM25 (cf. support de cours 2) :

$$\begin{aligned} h : \mathbb{R}^+ &\to \mathbb{R} \\ x &\mapsto \ln \frac{\left(\alpha\lambda^x e^{-\lambda} + (1-\alpha)\,\mu^x e^{-\mu}\right)\left(\beta e^{-\lambda} + (1-\beta)\,e^{-\mu}\right)}{\left(\beta\lambda^x e^{-\lambda} + (1-\beta)\,\mu^x e^{-\mu}\right)\left(\alpha e^{-\lambda} + (1-\alpha)\,e^{-\mu}\right)} \end{aligned}$$

où $\alpha \in]0,1[$, $\beta \in]0,1[$ et $\mu < \lambda$.

1. Quelles sont les caractéristiques de la fonction h (description) ?
2. Quelle est la limite $\lim_{x\to+\infty} \ln h(x)$?

Exercice 2

On considère une collection de documents $\mathcal{C} = \{d_1, \cdots, d_i, \cdots, d_N\}$ et un ensemble de requêtes $\mathcal{Q} = \{q_1, \cdots, q_l, \cdots, q_L\}$ données, où pour chaque couple $(d_i, q_l) \in \mathcal{C} \times \mathcal{Q}$ on dispose d'un jugement de pertinence binaire R. On suppose de plus que chaque document $d_i \in \mathcal{C}$ est représenté par un vecteur binaire de dimension V, $\mathbf{d}_i = (t_{1,i}, \cdots, t_{j,i}, \cdots t_{V,i})$. Rappelons la probabilité $p_j := P\left(t_{j,i} = 1 | R = 1, q\right)$ (resp. $s_j := P\left(t_{j,i} = 1 | R = 0, q\right)$) que le terme d'indice j du vocabulaire apparaisse dans un document pertinent (resp. non pertinent) vis-à-vis de la requête q.

1. Pour une requête fixe q, quelles sont les lois de probabilité suivies par le $j$ème terme du vocabulaire, si ce dernier apparaît dans un document d_i pertinent ou non pertinent vis-à-vis de cette requête ? (Nous notons $t_{j,i}$ ce $j$ème terme).
2. Soit d_i (resp. $d_{i'}$) un document jugé pertinent (resp. non pertinent) pour une requête de $\mathcal{Q}$. Montrer que $P\left(t_{j,i} | R = 1, q\right) = p_j^{t_j} (1-p_j)^{(1-t_j)}, \forall t_{j,i} \in \mathbf{d}_i$ et $\forall t_{j,i'} \in \mathbf{d}_{i'}$ $P\left(t_{j,i'} | R = 0, q\right) = s_j^{t_{j,i'}} (1-s_j)^{(1-t_{j,i'})}$.
   Nous noterons par la suite $P\left(t_{j,i} | R = 1, q\right) = P\left(t_{j,i} | p_j\right)$ et $P\left(t_{j,i} | R = 0, q\right) = P\left(t_{j,i} | s_j\right)$ où p_j et s_j jouent chacun le rôle de paramètre.
3. On note $\mathcal{R}$ (resp. $\bar{\mathcal{R}}$) le sous-ensemble des documents de $\mathcal{C}$ jugés pertinents au moins une fois (respectivement jamais jugés pertinents), par rapport à une requête de $\mathcal{Q}$ (i.e. $\mathcal{C} = \mathcal{R} \cup \bar{\mathcal{R}}$). On suppose de plus que les termes apparaissant dans n'importe quel document de $\mathcal{C}$ sont indépendants les uns des autres.
   Donner l'expression de $P(\mathbf{d}_i | \mathbf{p})$ pour $d_i \in \mathcal{R}$ et $\mathbf{p} = (p_1, \cdots, p_j, \cdots p_V)$ puis donner l'expression de $P(\mathbf{d}_{i'} | \mathbf{s})$ pour $\mathbf{d}_{i'} \in \bar{\mathcal{R}}$ et $\mathbf{s} = (s_1, \cdots, s_j, \cdots, s_V)$.

	Documents	Pertinent $\mathcal{R}$	Non-Pertinent $\bar{\mathcal{R}}$	Total
Terme présent	$\{t_j = 1\}$	r	$df_{t_j} - r$	df_{t_j}
Terme absent	$\{t_j = 0\}$	$\lvert\mathcal{R}\rvert - r$	$N - df_{t_j} - \lvert\mathcal{R}\rvert + r$	$N - df_{t_j}$
	Total	$\lvert\mathcal{R}\rvert$	$N - \lvert\mathcal{R}\rvert$	N

4. Il existe différentes méthodes statistiques pour estimer les paramètres $\mathbf{p} = (p_1, \cdots, p_j, \cdots p_V)$ et $\mathbf{s} = (s_1, \cdots, s_j, \cdots, s_V)$, parmi lesquelles la méthode du maximum de vraisemblance (MV) qui est la plus utilisée dans la littérature. Nous allons estimer les paramètres (vecteurs) **p** et **s** respectivement sur les sous ensembles $\mathcal{R}$ et $\bar{\mathcal{R}}$. Pour une collection de documents $\mathcal{X} = \left\{d_1, \cdots, d_{|\mathcal{X}|}\right\}$ ($\mathcal{X}$ étant $\mathcal{R}$ ou $\bar{\mathcal{R}}$), la méthode du MV consiste à trouver l'ensemble des paramètres $\boldsymbol{\lambda}^{MV}$ ($\mathbf{p}^{MV}$ ou $\mathbf{s}^{MV}$) qui maximise la vraisemblance des données $P(\mathcal{X}|\boldsymbol{\lambda})$. Dans le cas où on suppose que les documents sont tous indépendamment distribués, donner l'expression de $P(\mathcal{X}|\boldsymbol{\lambda})$.
5. Dire pourquoi l'estimateur du maximum de vraisemblance $\boldsymbol{\lambda^{MV}}$ peut s'obtenir grâce à l'équation :

$$\boldsymbol{\lambda^{MV}} = argmax_{\boldsymbol{\lambda}} \ln\left(P(\mathcal{X}|\boldsymbol{\lambda})\right)$$

Soit le tableau de contingence suivant, comptabilisant la présence et l'absence du terme d'indice j du vocabulaire dans les sous-ensembles $\mathcal{R}$ et $\bar{\mathcal{R}}$.

6. Montrer que, $\forall j \in \{1, \cdots, V\}$, $p_j^{MV} = \frac{r}{|\mathcal{R}|}$, $s_j^{MV} = \frac{df_{t_j} - r}{N - |\mathcal{R}|}$