Modèles de Markov Cachés et approches discriminantes

Exercice 1 – Synthèse sur les MMC

Q 1.1 Formalisation de l'apprentissage d'un MMC.
Généralement on apprend un MMC à partir d'une base de données d'apprentissage non étiquetée, c'est-à-dire constituée d'un ensemble de séquences d'observations, mais sans les séquences d'états associées. On commence par se placer dans ce cadre.

Q 1.1.1 On suppose que l'on dispose d'une base d'apprentissage d'une seule séquence d'observations $X = \{\mathbf{x}^{(1)}\}$. Quelle propriété satisfait le modèle λ qui maximise la vraisemblance des données d'apprentissage ? Quel algorithme utiliser pour faire l'apprentissage ?

Q 1.1.2 On suppose que l'on dispose d'une base d'apprentissage de N séquences $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)} ..., \mathbf{x}^{(N)}\}$. Quelle propriété satisfait le modèle λ qui maximise la vraisemblance des données d'apprentissage ?

Q 1.1.3 On considère maintenant le cas d'une base de données d'apprentissage étiquetée, c'est-à-dire constituée d'un ensemble de couples (séquence d'observations, séquence d'états). On suppose que l'on dispose d'une base d'apprentissage étiquetée de N séquences $XS = \{(\mathbf{x}^{(1)}, \mathbf{s}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{s}^{(2)}), ..., (\mathbf{x}^{(N)}, \mathbf{s}^{(N)})\}$. Quelle propriété satisfait le modèle λ qui maximise la vraisemblance des données d'apprentissage ?

Q 1.2 Difficulté de l'apprentissage d'un MMC.
On considère le cas d'une base de données d'apprentissage non étiquetée. On vous fournit la séquence d'observations $\mathbf{x} = (1, 2, 1, 1, 3, 2)$ produite par un modèle Markovien, mais on ne vous dit pas par quel type de modèle (nombre d'états etc) cette séquence a été produite, ni la séquence d'états correspondante.

Q 1.2.1 Quel modèle Markovien maximise la vraisemblance de la séquence $\mathbf{x}$ (nombre d'états, lois de probabilité de transitions et d'émission) ? Quel est son pouvoir de généralisation ?

Q 1.2.2 En supposant que la séquence a été générée par un modèle MMC à 1 état, quels sont les paramètres de ce modèle ?

Q 1.2.3 On suppose que cette séquence a été générée par un MMC à deux états. Proposez des paramètres pour ce modèle. Pouvez-vous prouver que votre modèle est localement optimal ? Vous commencerez par définir ce que signifie localement optimal.

Q 1.3 Apprentissage en présence de données étiquetées.
On change de cadre maintenant et on suppose que l'on vous fournit comme corpus d'apprentissage des données étiquetées, c'est-à-dire un ensemble de couples (séquence d'observations, séquence d'états). On considère une base d'apprentissage constituée d'une séquence $XS = \{(\mathbf{x} = (1, 2, 1, 1, 3, 2), \mathbf{s} = (1, 1, 1, 2, 2, 2))\}$ et on vous demande le MMC qui maximise la vraisemblance de cette base d'apprentissage.

Q 1.3.1 Quel est le nombre d'états du MMC ?

Q 1.3.2 Quels sont les paramètres du modèle optimal ? Pouvez-vous démontrer son optimalité ?
On généralise maintenant en considérant une base d'apprentissage étiquetée $XS = \{(\mathbf{x}^1, \mathbf{s}^1), \ldots, (\mathbf{x}^N, \mathbf{s}^N)\}$.

Q 1.3.3 Comment trouve-t-on le nombre d'états du modèle optimal ?

Q 1.4 Dans les stratégies d'apprentissage des MMC, quelle est la différence entre l'algorithme Baum-Welch simplifié et la version complète ? Sur quel variable intermédiaire repose la version complète ?

Exercice 2 – Approche discriminante : régression logistique

Jusqu'ici, nous avons toujours travaillé sur le critère de la vraisemblance selon le schéma :

1. Modélisation probabiliste d'une situation = 1 classe de données (chiffres manuscrits, mouvements du stylo sur des lettres...), paramètre θ

2. Optimisation des θ = trouver $\theta^\star$ maximisant la vraisemblance

Pourtant, ce type d'approche présente une faiblesse évidente dans les problèmes de classification : les classes sont apprises de manière isolées et on ne peut pas se focaliser sur l'information discriminante (ce qui distingue une classe d'une autre). Une autre classe de modèles permet de palier cette faiblesse : sur les données multi-variées, il s'agit de la régression logistique (ou classifieur de maximum d'entropie), sur les séquences (et les graphes) des CRF (Conditional Random Fields). Nous allons travailler sur le premier modèle pour vous donner une idée de ce qui est faisable.

Régression logistique (problèmes de classification supervisé). On pose le problème de la classification binaire (2 classes). On suppose que l'espace auquel appartiennent les données est $\mathbb{R}^d$ et que l'espace des étiquettes (probabilistes) est $\mathcal{Y} = \{0, 1\}$.

1. En régression logistique, la classification d'un exemple $\mathbf{x}_i \in \mathbb{R}^d$ se fait selon la valeur d'une fonction f dépendant d'un vecteur $\mathbf{w} \in \mathbb{R}^d$ et d'un réel b qui est défini par :

$$f(\mathbf{x}_i) = \frac{1}{1 + \exp(-(\mathbf{x}_i\mathbf{w} + b))}, \quad \mathbf{x}_i,\ \mathbf{w} \text{ respectivement en ligne et colonne}$$

   (a) Donner les valeurs minimale et maximale de f. Tracer f en fonction de $\alpha = \mathbf{x}\mathbf{w} + b$. Pourquoi f est-elle intéressante pour un problème de classification binaire ? En déduire une règle d'affectation à une classe utilisant $f(\mathbf{x})$ pour un exemple $\mathbf{x}$.

   (b) Quelles sont les limites de la classification par régression logistique (donner un exemple de problème qui ne peut être appris). Par exemple, dans le cas où $d = 2$, $\mathbf{w} = [-2\ 1]$ et $b = 1$, représenter graphiquement la surface de décision correspondant à la règle d'affectation proposée précédemment (les limites du modèles apparaissent alors clairement).

2. Soit un ensemble d'apprentissage $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ et une distribution inconnue $p(X, Y)$ sur $\mathbb{R}^d \times \mathcal{Y}$. Rappeler les hypothèses usuellement faites en classification supervisée.
   Puis, écrire la vraisemblance jointe des données de S en fonction des $P(y_i|\mathbf{x}_i)$.

3. Dans le cadre de la régression logistique, on construit $f(\mathbf{x})$ pour estimer la probabilité $P(Y = 1|X = x)$ qu'un exemple $\mathbf{x}$ soit de la classe $+1$. En utilisant ce modèle statistique pour les données, écrire la log-vraisemblance discriminante de l'échantillon S en fonction des y_i et des $f(\mathbf{x}_i)$.
   NB : l'expression de la vraisemblance est très particulière dans la regression logistique, on la nomme *discriminante* pour éviter la confusion

4. Quels sont les paramètres à estimer ?

5. Donner l'expression de $\frac{\partial L_{\log}}{\partial w_j}$, pour $j = 1, \ldots, d$ ainsi que l'expression de $\frac{\partial L}{\partial b}$.

6. Le gradient de $L_{\log}$ peut-il s'annuler directement ? Proposer une équation de mise à jour (type gradient) permettant de produire une suite de paramètres menant à un maximum de vraisemblance.

7. En considérant l'ensemble d'apprentissage $S = \left\{ \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, 0 \right) \right\}$ les valeurs initiales $w^0 = [0\ 1]^T$ et $b^0 = -1$ et un pas d'apprentissage fixe $\varepsilon = 0.3$, faire deux itérations des algorithmes d'apprentissage proposés.