APPRENTISSAGE STATISTIQUE AVANCÉ

EXAMEN - CONTRÔLE DE CONNAISSANCES (DURÉE 1 HEURE 30)

Les notes de cours ne sont pas autorisées, l'usage d'ordinateurs ou tablettes est prohibé.

OPTIMISATION

1. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction convexe dérivable et $A : \mathbb{R}^m \to \mathbb{R}^n$ une fonction affine. Pour chaque assertion, dire si elle est vraie ou fausse.

	VRAI	FAUX
$\forall x \in \mathbb{R}^n, \forall y \in \mathbb{R}^n, f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle$		
$\nabla(f \circ A)(x) = \nabla f(A(x))$		
$\nabla f(x) = 0 \Leftrightarrow x \in \arg\min_{z \in \mathbb{R}^n} f(z)$		

2. On s'intéresse à la résolution du problème

$$\min_{w \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^{N} f(x_i, w)$$

où x_i est le vecteur de paramètres de la donnée i.
On propose les deux algorithmes suivants :

Algorithme 1
Choisir $w_0 \in \mathbb{R}^p$ et $(\alpha_k)_k$ où $\alpha_k \geq 0, \forall k$
pour $k \geq 0$:
$i_{k+1} \sim \mathcal{U}(\{1, \ldots, N\})$
$w_{k+1} = w_k - \alpha_k \nabla f(x_{i_{k+1}}, w_k)$

Algorithme 2
Choisir $w_0 \in \mathbb{R}^p$ et $(\alpha_k)_k$ où $\alpha_k \geq 0, \forall k$
pour $k \geq 0$:
$w_{k+1} = w_k - \alpha_k \frac{1}{N} \sum_{i=1}^{N} \nabla f(x_i, w_k)$

Comment s'appellent chacun de ces algorithmes ? Discuter brièvement de leurs avantages respectifs ?

3. Soit $A \in \mathbb{R}^{d \times d}$ une matrice symétrique définie positive et soit $b \in \mathbb{R}^d$. Pout tous $x, y \in \mathbb{R}^d$, on pose $\langle x, y \rangle_A := \langle Ax, y \rangle$. Soit $f(x) := \frac{1}{2} x^\top A x - b^\top x$. Pour $d \in \mathbb{R}^d$, on définit

$$s^* = \arg\min_{s \in \mathbb{R}} f(x + sd).$$

Pour chaque assertion, dire si elle est vraie ou fausse.

(a) $\nabla f(x) = Ax - b$.

(b) $\nabla^2 f(x) = \frac{A + A^\top}{2}$

(c) $s^* = -\dfrac{\langle d, \nabla f(x) \rangle}{\langle d, \nabla^2 f(x) d \rangle}$

(d) $s^* = \dfrac{\langle d, b - Ax \rangle}{\langle d, d \rangle_A}$

(e) Soit $d' = -r + \alpha d$. Si $\langle d', d \rangle_A = 0$, alors $\alpha = \dfrac{\langle r, d \rangle_A}{\langle d, d \rangle_A}$.

DEEP LEARNING

1. Quelle est la formule mathématique de la fonction d'activation ReLu ?
2. Pour apprendre les paramètres d'un réseau de neurones, on minimise une fonction de coût C, par descente de gradient. La formule analytique de C en fonction des paramètres est très complexe. Quelles sont les deux propriétés mathématiques que l'on utilise pour modifier les paramètres à chaque itération ?
3. On veut prédire 10 valeurs à partir d'une image couleur 100×100 pixels. Pour cela on utilise un CNN à 3 couches, dont deux couches de convolution standard (avec padding, stride 2) et ayant respectivement 32 puis 16 filtres 3×3. Les 3 couches ont des biais. On a un pooling après chaque couche de convolution.
   (a) Quelle est le nombre de paramètres à apprendre ?
   (b) on utilise tensorflow pour apprendre ce réseau et la fonction tf.nn.conv2d (T, W, strides, padding) pour les convolutions. Quelles sont les dimensions du tensor T en input de l'appel de tf.nn.conv2d pour la 2ieme couche de convolution ?
4. Pour chacune des techniques suivantes indiquez son principal intérêt. (1 réponse)

	Limiter le surapprentissage	Augmenter la vitesse de convergence	traiter une tâche spécifique
Data augmentation			
Early stopping			
Batch normalization			
Adagrad			
LSTM			
Dropout			
Regularization			
Adam			
'Xavier' initialization			
Auto-encoder			

5. Quelles sont les deux tâches effectuées simultanément par les réseaux YOLO ou SSD ?

RNN

1. Donnez la formulation mathématique d'un RNN (RNN simple) d'entrée $x^{<t>}$, d'état caché $a^{<t>}$ et de sortie $\hat{y}^{<t>}$.
2. Nous notons T_x la longeur de la séquence $x^{<t>}$ d'entrée d'un RNN et T_y celle de la sortie $y^{<t>}$. En fonction de T_x et T_y quelles sont les 4 architectures principales de modèles séquentiels ? Donnez pour chacun un exemple d'utilisation.
3. Quel problème une cellule de type LSTM permet-elle de résoudre ? Comment le résout-elle ?

APPRENTISSAGE DE MÉTRIQUE

Soit $\mathbb{S}_+^d$ le cône des matrices $d \times d$ symétriques semi-définies positives. La distance de Mahalanobis $D_{\boldsymbol{M}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ associée à $\boldsymbol{M} \in \mathbb{S}_+^d$ est définie par $D_{\boldsymbol{M}}(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^T \boldsymbol{M} (\boldsymbol{x} - \boldsymbol{x}')}$. Soit $\mathcal{R} \subset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ un ensemble fini de triplets d'observations. On considère le problème d'apprentissage de métrique suivant :

$$\begin{aligned} \min_{\boldsymbol{M} \in \mathbb{S}_+^d, \boldsymbol{\xi} \geq 0} \quad & \sum_{i,j,k} \xi_{ijk} \\ \text{s.t.} \quad & D_{\boldsymbol{M}}^2(\boldsymbol{x}_i, \boldsymbol{x}_k) - D_{\boldsymbol{M}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 1 - \xi_{ijk} \quad \forall (\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) \in \mathcal{R} \end{aligned}$$

1. Quelle est l'interprétation des contraintes de ce problème ? Proposer une fonction de perte permettant d'obtenir un problème d'optimisation équivalent sans faire appel à des contraintes. Indice : il faut utiliser une perte hinge, c'est-à-dire de la forme $[a]_+ = \max(0, 1 - a)$.
2. Pourquoi est-il intéressant d'apprendre une matrice $\boldsymbol{M} \in \mathbb{S}_+^d$ dont le rang est inférieur à d ? Modifier le problème d'optimisation ci-dessus de manière à inciter les solutions à être de rang faible tout en gardant le problème convexe.

PASSAGE À L'ÉCHELLE DES MÉTHODES À NOYAUX

1. Quand le nombre n d'exemples d'apprentissage est grand, est-il plus efficace de résoudre le problème SVM dans sa forme primale ou dans sa forme duale ? Justifier.

GRAPHES ET APPRENTISSAGE

1. Calculer la matrice Laplacienne $\boldsymbol{L}$ associée au graphe représenté en Figure 1. Rappel : on a $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, où $\boldsymbol{D}$ est une matrice diagonale contenant le degré des noeuds et $\boldsymbol{A}$ est la matrice d'adjacence du graphe.
2. Qu'est-ce que le mécanisme de l'attachement préférentiel caractérisant les réseaux dits scale-free ? Donner une manière empirique d'identifier qu'un réseau est de ce type.

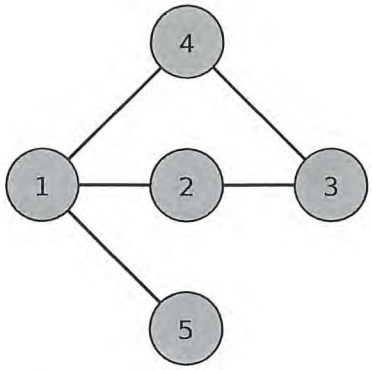

FIGURE 1 – Un graphe non dirigé.

CONDITIONAL RANDOM FIELDS (CRF)

On rappelle qu'étant données :
— une séquence $\underline{\mathbf{x}} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$;
— la séquence d'étiquettes correspondantes : $\underline{y} = (y_1, \cdots, y_n)$;
un modèle CRF définit la probabilité a posteriori des sorties conditionnellement aux entrées suivant :

$$p(\underline{y}|\underline{\mathbf{x}};\boldsymbol{\theta}) \triangleq \frac{1}{Z(\underline{\mathbf{x}},\boldsymbol{\theta})} \exp \sum_{j=1}^{D} \theta_j F_j(\underline{\mathbf{x}},\underline{y}) \triangleq \frac{1}{Z(\underline{\mathbf{x}},\boldsymbol{\theta})} \Psi(\underline{\mathbf{x}},\underline{y};\boldsymbol{\theta}); \quad \boldsymbol{\theta} = \{\theta_1, \cdots, \theta_D\};$$

où :
— $Z(\underline{\mathbf{x}},\boldsymbol{\theta}) = \sum_{\underline{y}} \exp \sum_j \theta_j F_j(\underline{\mathbf{x}},\underline{y})$ à la fonction partition (servant de normalisation) ;
— les $F_j(\underline{\mathbf{x}},\underline{y}); 1 \leq j \leq D$ sont des fonctions de caractéristiques (*feature functions*).

1. Expliquer la différence entre caractéristiques (*features*) et fonctions de caractéristiques (*feature functions*).

2. Comment obtient-on un CRF en chaine linéaire (*linear-chain CRF*) ?

3. A quoi sert l'algorithme de Viterbi ?

4. Posons $M_i(y_{i-1}, y_i, \underline{\mathbf{x}}) \triangleq \exp\left(\sum_{j=1}^{D} \theta_j f_j(y_{i-1}, y_i, \underline{\mathbf{x}}, i)\right)$; de sorte que :

$$p(\underline{y}|\underline{\mathbf{x}};\boldsymbol{\theta}) = \frac{1}{Z(\underline{\mathbf{x}},\boldsymbol{\theta})} \prod_{i=1}^{n} M_i(y_{i-1}, y_i, \underline{\mathbf{x}}).$$

On définit les "scores" *forward* et *backward*, respectivement, comme :
— $\boldsymbol{\alpha}_m(y_m) = \sum_{y_{m-1}} M_m(y_{m-1}, y_m)\boldsymbol{\alpha}_{m-1}(y_{m-1})$; $2 \leq m \leq n$;
— $\boldsymbol{\beta}_m(y_m) = \sum_{y_{m+1}} M_{m+1}(y_m, y_{m+1})\boldsymbol{\beta}_{m+1}(y_{m+1})$; $1 \leq m \leq n-1$ et $\boldsymbol{\beta}_n(y_n) = 1$.

a. Exprimer les probabilités marginales $p(y_{m-1}, y_m|\underline{\mathbf{x}}) \triangleq \sum_{\underline{y}\setminus\{y_{m-1},y_m\}} p(\underline{y}|\underline{\mathbf{x}})$ en fonction de $\boldsymbol{\alpha}_{m-1}(y_{m-1})$ et $\boldsymbol{\beta}_m(y_m)$.
b. En déduire $p(y_m|\underline{\mathbf{x}})$. Cette expression permet d'envisager une autre forme de décodage dans laquelle on garde à chaque position m l'étiquette la plus probable (indépendamment des autres positions).

Q-LEARNING

On considère un processus de décision Markovien (MDP) avec un facteur d'actualisation $\gamma \leq 1$ sur les récompenses et un horizon de temps infini.

1. Rappeler la définition de la fonction objectif R, puis de la fonction valeur V associée à une politique π.
2. Décrire la technique d'itération par valeur (Value Iteration) permettant de trouver une solution optimale. Sous quelle condition la convergence est-elle garantie ?
3. Dans quel(s) cas cette technique s'applique-t-elle ? Justifier la réponse.
   Sinon, quels algorithmes peut-on utiliser ?

APPRENTISSAGE SEMI-SUPERVISÉ

1. Définir une contrainte sur le modèle en fonction des données étiquetées et non étiquetées qui permette d'effectuer de l'apprentissage semi-supervisé
2. Prendre l'exemple de la fonction de coût quadratique et de la régression ridge pour appliquer la contrainte choisie avec un modèle à noyau.
3. Au delà des modèles à noyaux pour quel type de modèles cette approche peut-elle s'appliquer ?

APPRENTISSAGE STRUCTURÉ

1. Proposer une approche de prédiction structurée pour traiter un problème de classification multi-classe.
2. Traiter le cas où les classes appartiennent à une hiérarchie.

HMM

Soit la matrice ci dessous qui inclut les transitions entre caractères (bi-grams de caractères) pour la langue française. Seulement certaines de ces transitions sont affichées ici pour des questions de lisibilité. L'état start correspond à un espace en début de mot, l'état stop à un espace de fin de mot.

	start	*a*	*b*	...	*n*	*o*	...	*stop*
start	0	0.0762	0.0117	...	0.0655	0.0186	...	0
a	0	0	0.0034	...	0.1775	0	...	0.1571
b	0	0.0310	0.0010	...	0.0010	0.0310	...	0.0010
...	.	..	.	...	...	...	...	...
n	0	0.0339	0	...	0.0290	0.1138	...	0.2058
o	0	0	0.0082	...	0.3005	0	...	0.0027
stop	0	0	0	...	0	0	...	1

1. Expliquer les éléments nuls dans la matrice.

2. Quelle est la probabilité de la séquence de caractères correspondant au mot : " bon ", incluant les espaces de début et fin de mot ?