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.

Analyse Convexe

1. QCM. Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction dérivable telle que ∇f est L-Lipschitzien avec $L > 0$ et soient x, y deux vecteurs de $\mathbb{R}^n$.

Affirmation	Vrai	Faux
$\lvert f(x) - f(y)\rvert \leq L\lvert x - y\rvert$		x
$\lVert\nabla f(x) - \nabla f(y)\rVert \leq L\lVert x - y\rVert$	x	
Si $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ alors $f(x_{k+1}) > f(x_k)$		x
$f(y) \leq f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}\lVert y - x\rVert^2$	x	

2. Donner la formule de la sous-différentielle de la valeur absolue.

Solution :

$$\partial|\cdot|(x) = \begin{cases} \{-1\} & \text{si } x < 0 \\ [-1, 1] & \text{si } x = 0 \\ \{1\} & \text{si } x > 0 \end{cases}$$

Deep Learning

1. Quelle est la formule mathématique de la fonction d'activation ReLu ? $Relu(X) = \max(0, X)$

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 ?

   — Développement limité en série de Taylor à l'ordre 1 pour l'approximation locale.
   — Dérivation des fonctions composées pour le calcul de proche en proche.

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 ?

C1 : 32*(3*3*3+1)= 896 C2 : 16*(32*3*3+1) = 4624 C3 : 10*(16*25*25+1) = 100010
Total : 105530 paramètres

(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 ?

Shape = (batchsize, 50, 50, 32)

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	x		
Early stopping	x		
Batch normalization		x	
Adagrad		x	
LSTM			x
Dropout	x		
Regularization	x		
Adam		x	
'Xavier' initialization		x	
Auto-encoder			x

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

Classification : estimation de la probabilité d'appartenance à chaque classe.
Régression : calcul de la position des bounding box

Modèle de Markov

On dispose d'une matrice de transitions A entre éléments grammaticaux dans une phrase. Ces éléments grammaticaux peuvent être considérés comme des états d'une chaîne de Markov. Ils sont appelés tags en traitement du langage naturel. La matrice comprend 6 états : 1 : état initial de phrase, 2 : déterminant, 3 : Nom commun, 4 : Nom propre, 5 : verbe, 6 : état final de phrase.

$$A = \begin{bmatrix} 0 & 0.5 & 0 & 0.4 & 0.1 & 0 \\ 0 & 0 & 0.95 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0.9 & 0.1 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.1 \\ 0 & 0.25 & 0 & 0.25 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

1. La matrice est-elle gauche-droite ou ergodique ?
2. Quelle est la probabilité de la suite de tags : Etat initial-Déterminant-Nom commun-Verbe-Nom Propre-Etat final ?
3. On suppose que la matrice A correspond à la matrice de transitions d'un modèle de Markov caché. Les observations possibles sont les mots en français (appelés tokens). Enoncer une séquence d'observations correspondant à la suite de tags précédente.

corrigé

1) modèle ergodique
2)

$$0.5 \times 0.95 \times 0.9 \times 0.25 \times 0.1 = 0.0106875$$

3) Les touristes aiment Paris

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 $\mathcal{S} \subset \mathbb{R}^d \times \mathbb{R}^d$ et $\mathcal{D} \subset \mathbb{R}^d \times \mathbb{R}^d$ deux ensembles finis de paires d'observations. On considère le problème d'apprentissage de métrique suivant :

$$\min_{\boldsymbol{M} \in \mathbb{S}_+^d} \left[\sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} \max\left(0, d^2_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) - 1\right) \quad + \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} \max(0, 1 - d^2_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)) \right]$$

1. Quelle est l'interprétation de la fonction objective ci-dessus ?
2. Pourquoi est-il intéressant d'apprendre une matrice $\boldsymbol{M} \in \mathbb{S}_+^d$ de rang inférieur à d ? Donner une manière d'inciter les solutions du problème à être de rang faible.

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

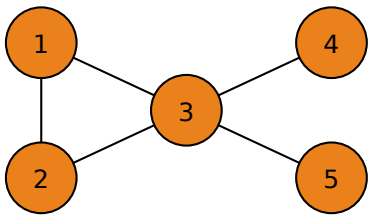

Figure 1 – Un graphe non dirigé.

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 phénomène de l'attachement préférentiel caractérisant certains réseaux ? Donner une manière empirique d'identifier qu'un réseau est de ce type.

Réponses

Apprentissage de métrique

1. La fonction objective cherche à imposer des distances inférieures ou égales à 1 pour les paires de $\mathcal{S}$, et supérieures ou égales à 1 pour les paires de $\mathcal{D}$. Si ce n'est pas le cas, on subit une perte égale à l'écart entre la valeur réalisée et la valeur souhaitée (c'est-à-dire 1).
2. Si $\boldsymbol{M}$ est de rang $k \leq d$, $D_{\boldsymbol{M}}$ est équivalente à une distance Euclidienne après projection linéaire des données dans un espace à k dimensions. On effectue donc une réduction de dimension quand $k < d$.

   Pour inciter la solution à être de rang faible, deux manières possibles :
   - Utiliser la norme trace (aussi appelée norme nucléaire) comme régularisation dans la fonction objective. Le problème reste alors convexe.
   - Opérer un changement de variable : on optimise une matrice de transformation $\boldsymbol{L} \in \mathbb{R}^{k \times d}$ et on remplace $\boldsymbol{M}$ par $\boldsymbol{L}^T\boldsymbol{L}$. Le problème devient alors non convexe.

Passage à l'échelle des méthodes à noyaux

1. Quand n est grand, il est moins coûteux en temps de calcul de résoudre le problème dans le primal que dans le dual. En effet, résoudre le problème dans le dual nécessite de construire une matrice de Gram à $O(n^2)$ entrées. À l'inverse, la fonction objective du primal est une somme de n termes que l'on peut minimiser efficacement même quand n est grand (par exemple avec l'algorithme du gradient stochastique).

Graphes et apprentissage

1. On a
$$\boldsymbol{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$
2. Dans un processus d'attachement préférentiel, les noeuds qui rejoignent le réseau se lient avec un noeud déjà présent avec une probabilité proportionnelle au degré de ce noeud. En pratique, on peut identifier un tel réseau à la distribution des degrés dans le graphe, qui suit une loi de puissance.