IFT 6390 Fundamentals of Machine Learning
Ioannis Mitliagkas

Devoir 2 - Partie pratique réalisé par

Clément DETRY
Hamed Nazim MAMACHE

- Ce devoir peut être être fait en équipes de maximum 2. Vous pouvez discuter avec des étudiant.e.s d'autres équipes mais les réponses et le code que vous soumettez doivent être les vôtres. A noter que nous utiliserons l'outil de détection de plagiat de Gradescope. Tous les cas suspectés de plagiat seront enregistrés et transmis à l'Université pour vérification.
- La partie pratique doit être codée en python (avec les librairies numpy et matplotlib), et envoyée sur Gradescope sous fichier python. Pour permettre l'évaluation automatique, vous devez travailler directement sur le modèle donné dans le répertoire de ce devoir. Ne modifiez pas le nom du fichier ou aucune des fonctions signatures, sinon l'évaluation automatique ne fonctionnera pas. Vous pouvez bien sur ajouter de nouvelles fonctions et importations python
- Les figures, courbes et parties pratiques du rapport doivent être envoyées au format pdf sur Gradescope. Pour le rapport il est recommandé d'utiliser un Jupyter notebook, en écrivant les formules mathématiques avec MathJax et en exportant vers pdf. Vous pouvez aussi écrire votre rapport en LaTeX; LyX; Word. Dans tout les cas, exportez votre rapport vers un fichier pdf que vous enverrez. Vous êtes bien sur encouragés à vous inspirer de ce qui a été fait en TP.
- Vous devez soumettre vos solutions sur Gradescope en utilisant le devoir intitulé `Devoir 2 / Homework 2 - Pratique/Practical - 6390A/B` pour le code et `Devoir 2 / Homework 2 - Pratique/Practical - Rapport/Report - 6390A/B` pour le rapport.

Vous devez travailler sur le modèle `solution.py` du répertoire et compléter les fonctions basiques suivantes en utilisant numpy et python

Un-contre-tous, Perte L2 SVM

Cette partie consiste à implémenter le SVM un-contre-tous avec pénalité L2, qui est couramment utilisé pour faire de la classification multi-classe. La fonction de perte d'un SVM avec pénalité L2 est différentiable et impose une plus grande pénalité sur les points qui sont à l'intérieur de la marge maximale. Dans l'approche un-contre-tous (*one-versus-all* ou *OVA* en anglais), nous entraînons m classificateurs binaires, soit un pour chaque classe, et lors de la prédiction, nous sélectionnons la classe qui maximise la marge pour un point test.

Considérant un jeu d'entraînement $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, où $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{1, .., m\}$, p est le nombre de traits (ou attributs) et m est le nombre de classes, nous voulons minimiser la fonction objectif suivante:

$$\frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in S} \sum_{j'=1}^{m} \mathcal{L}(\mathbf{w}^{j'}; (\mathbf{x}_i, y_i)) + \frac{C}{2} \sum_{j'=1}^{m} \|\mathbf{w}^{j'}\|_2^2$$

où

$$\mathcal{L}(\mathbf{w}^{j'}; (\mathbf{x}_i, y_i)) = \Big(\max\{0, 2 - (\langle \mathbf{w}^{j'}, \mathbf{x}_i \rangle) \mathbb{1}\{y_i = j'\}\}\Big)^2$$

et

$$\mathbb{1}\{y_i = j'\} = \begin{cases} 1 & \text{if } y_i = j' \\ -1 & \text{if } y_i \neq j' \end{cases}$$

Afin de mettre à jour les paramètres $\mathbf{w}$ de notre fonction objectif, nous utiliserons des techniques de descente de gradient. (Note: Vous remarquerez que dans le jeu de données fourni pour cette partie, le dernier élément de chaque ligne est un 1. Cette notation permet de ne pas avoir de paramètre de biais b séparé, mais plutôt de l'inclure implicitement dans $\mathbf{w}$ comme étant un autre poids.)

Le jeu de données est constitué de données de capteurs de smartphones et de labels de classification pour la direction du mouvement. Il peut être téléchargé ici.

Le fichier solution contient une fonction appelée `load_data()` pour lire et effectuer quelques transformations sur les données telle que la normalization. Pour que `load_data()` fonctionne bien, vous devez dézipper le jeu de données téléchargé et le placer dans le même dossier que votre fichier `solution.py`. En résumé, vous devez avoir la structure de projet suivante:

```
[dossier du devoir]/
├── solution.py
├── Smartphone Sensor Data/
└── ...
```

1. [5 pts] Quelle est la dérivée du terme de régularisation de la fonction de perte

$$\frac{C}{2}\sum_{j'=1}^{m}\|\mathbf{w}^{j'}\|_2^2$$

par rapport à w_k^j? (le $k^{ième}$ poids du vecteur de poids pour la $j^{ième}$ classe)? Écrivez tous les étapes et mettez la réponse dans votre fichier PDF.

Réponse:
Calcul de $\frac{\partial}{\partial w_k^j}\frac{C}{2}\sum_{j'=1}^{m}\|\mathbf{w}^{j'}\|_2^2$:

$$\begin{aligned}
\frac{\partial}{\partial w_k^j}\frac{C}{2}\sum_{j'=1}^{m}\|\mathbf{w}^{j'}\|_2^2 &= \frac{C}{2}\sum_{j'=1}^{m}\frac{\partial}{\partial w_k^j}\|\mathbf{w}^{j'}\|_2^2 \\
&= \frac{C}{2}\sum_{j'=1}^{m}\frac{\partial}{\partial w_k^j}(\sum_{i=1}^{p} w_i^{j'2}) \\
&= \frac{C}{2}\sum_{j'=1}^{m}\sum_{i=1}^{p}\frac{\partial}{\partial w_k^j} w_i^{j'2} \\
&= C w_k^j
\end{aligned}$$

Donc :

$$\boxed{\frac{\partial}{\partial w_k^j}\frac{C}{2}\sum_{j'=1}^{m}\|\mathbf{w}^{j'}\|_2^2 = C w_k^j}$$

2. [10 pts] Quelle est la dérivée du terme appelé *hinge loss*

$$\frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\sum_{j'=1}^{m}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))$$

de la fonction de perte par rapport à w_k^j?

Exprimez votre réponse en termes de $\mathbf{x}_{i,k}$ (la $k^{ième}$ entrée du $i^{ième}$ exemple $\mathbf{x}_i$).

Assumez que

$$\frac{\partial}{\partial a}\max\{0,a\} = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases}$$

(Cette dernière affirmation n'est pas exactement vraie: à a=0, la dérivée n'est pas définie. Cependant, pour ce problème, nous allons assumer qu'elle est correcte.)

Réponse:
Calcul de $\frac{\partial}{\partial w_k^j}\frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\sum_{j'=1}^{m}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))$:

$$\frac{\partial}{\partial w_k^j}\frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\sum_{j'=1}^{m}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = \frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\sum_{j'=1}^{m}\frac{\partial}{\partial w_k^j}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))$$

On pose : $f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = \max\{0, 2-(\langle \mathbf{w}^{j'},\mathbf{x}_i\rangle)\mathbb{1}\{y_i = j'\}\}$.
On a alors :

$$\begin{aligned}\frac{\partial}{\partial w_k^j}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) &= \frac{\partial}{\partial w_k^j}f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))^2 \\ &= 2f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))\frac{\partial}{\partial w_k^j}f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))\end{aligned}$$

On pose $g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = 2-(\langle \mathbf{w}^{j'},\mathbf{x}_i\rangle)\mathbb{1}\{y_i = j'\}$, tel que :
$f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = \max\{0, g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))$.
On a alors :

$$\begin{aligned}\frac{\partial}{\partial w_k^j}f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) &= \frac{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))}{\partial w_k^j}\frac{\partial}{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))}f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) \\ &= \frac{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))}{\partial w_k^j}\frac{\partial}{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))}\max\{0, g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))\}\end{aligned}$$

et :

$$\frac{\partial}{\partial w_k^j}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = 2f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))\frac{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))}{\partial w_k^j}\frac{\partial}{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))}\max\{0, g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))\}$$

Or :

$$\begin{aligned}\frac{\partial}{\partial w_k^j}g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) &= \frac{\partial}{\partial w_k^j}(2-(\langle \mathbf{w}^{j'},\mathbf{x}_i\rangle)\mathbb{1}\{y_i = j'\}) \\ &= \begin{cases} -x_{i,k}\mathbb{1}\{y_i = j'\} & \text{if } j' = j \\ 0 & \text{if } j' \neq j \end{cases}\end{aligned}$$

D'où :

$$\frac{\partial}{\partial w_k^j}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = \begin{cases} -2f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))x_{i,k}\mathbb{1}\{y_i=j'\}\frac{\partial\ \max\{0,g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))\}}{\partial g(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))} & \text{if } j'=j \\ 0 & \text{if } j'\neq j \end{cases}$$

Donc :

$$\frac{\partial}{\partial w_k^j}\frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\sum_{j'=1}^{m}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = \frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}2f(\mathbf{w}^{j'};(\mathbf{x}_i,y_i))(-x_{i,k}\mathbb{1}\{y_i=j'\})\frac{\partial\ \max\{0,g)\}}{\partial g}|_j$$

En assumant que :

$$\frac{\partial}{\partial a}\max\{0,a\} = \begin{cases} 1 & \text{if } a>0 \\ 0 & \text{if } a\leq 0 \end{cases}$$

On a finalement :

$$\boxed{\frac{\partial}{\partial w_k^j}\frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\sum_{j'=1}^{m}\mathcal{L}(\mathbf{w}^{j'};(\mathbf{x}_i,y_i)) = \frac{2}{n}\sum_{(\mathbf{x}_i,y_i)\in S}\max\{0,2-(\langle\mathbf{w}^j,\mathbf{x}_i\rangle)\mathbb{1}\{y_i=j\}\}\ (-x_{i,k})\mathbb{1}\{y_i=j\}}$$

3. [30 pts] Complétez les méthodes suivantes dans le code:

   (a) [5 pts] SVM.make_one_versus_all_labels: Étant donné un tableau d'étiquettes qui sont des entiers et le nombre de classes m, cette fonction devrait retourner un tableau 2-d qui correspond au terme $\mathbb{1}\{y_i = j'\}$ défini plus haut. Dans ce tableau, chaque ligne contient des -1 à l'exception de l'élément qui correspond à la bonne classe et qui devrait être un 1. Par exemple, si le tableau que l'on donne en entrée est $[1, 0, 2]$ et que $m = 4$, la fonction retournera le tableau suivant:$[[-1, 1, -1, -1], [1, -1, -1, -1], [-1, -1, 1, -1]]$. Les entrées de la fonction sont y (un tableau numpy de dimension (nombre de classes,)) et m (un entier représentant le nombre de classes), et la sortie devrait être un tableau numpy de dimension (nombre d'exemples, m). Pour ce devoir, m sera égal à 6, mais vous devriez implémenter cette fonction pour qu'elle puisse fonctionner avec n'importe quel $m > 2$.

   (b) [5 pts] SVM.compute_loss : Étant donné un minibatch d'exemples, cette fonction devrait calculer la perte. Les entrées de la fonction

sont x (un tableau numpy de dimension (minibatch size, 562)) et y (un tableau numpy de dimension (minibatch size, 6)) et la sortie devrait être la perte calculée, un scalaire.

(c) [10 pts] SVM.compute_gradient: Considérant un minibatch d'exemples, cette fonction devrait calculer le gradient de la fonction de perte par rapport au paramètre **w**. Les entrées de la fonction sont X (un tableau numpy de dimension (minibatch size, 562)) et y (un tableau numpy de dimension (minibatch size, 6)) et la sortie devrait être le gradient calculé, un tableau numpy de dimension $(562, 6)$, soit la même dimension que celle du paramètre **w**. (Indice: utilisez les expressions que vous avez dérivées précédemment.)

(d) [5 pts] SVM.infer: Étant donné un minibatch d'exemples, cette fonction devrait prédire la classe de chaque exemple, c'est-à-dire la classe qui a le plus haut score. L'entrée de la fonction est X (un tableau numpy de dimension (minibatch size, 562)) et la sortie est *y_inferred* (un tableau numpy de dimension (minibatch size, 6)). La sortie devrait être en format un-contre-tous, c'est-à-dire -1 pour les classes qui ne sont pas prédites et +1 pour la classe prédite.

(e) [5 pts] SVM.compute_accuracy: Étant donné un tableau de classes prédites et un tableau des vraies classes, cette fonction devrait retourner la proportion de classifications correctes, soit un scalaire entre 0 et 1. Les entrées de cette fonction sont *y_inferred* (un tableau numpy de dimension (minibatch size, 6)) et y (un tableau numpy de dimension (minibatch size, 6)) et la sortie est un scalaire.

4. [5 pts] La méthode SVM.fit utilise le code que vous avez écris ci-dessus pour entraîner le SVM. Après chaque époque (après avoir passer à travers tous les exemples du jeu de données), SVM.fit calcule la perte et l'exactitude des points d'entraînement et la perte et l'exactitude des points tests.

   Faites le graphique de ces quatre quantités en fonction du nombre d'époques, pour $C = 1, 5, 10$. Utilisez comme hyperparamètres 200 époques, un taux d'apprentissage de 0.0001 et une longueur de minibatch de 100.

Vous devriez avoir 4 graphiques, soit un graphique pour chaque quantité, incluant les courbes pour les 3 valeurs de C. Ajoutez ces 4 graphiques dans votre rapport.

Sur la base de ces graphiques, le surapprentissage semble-t-il être un problème pour ce jeu de données et cet algorithme? Expliquez brièvement.

Réponse:

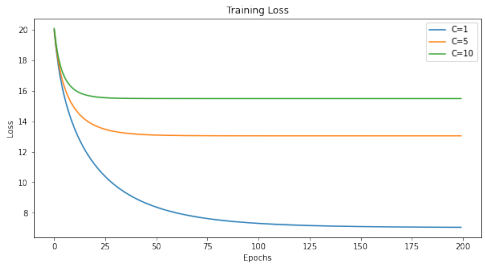

Figure 1: Evolution de la Train loss

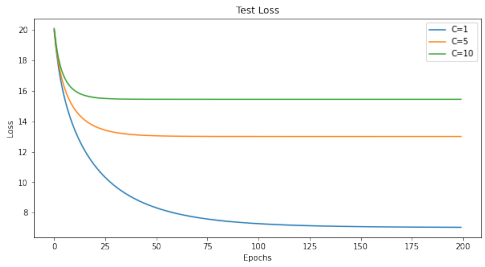

Figure 2: Evolution de la Test loss

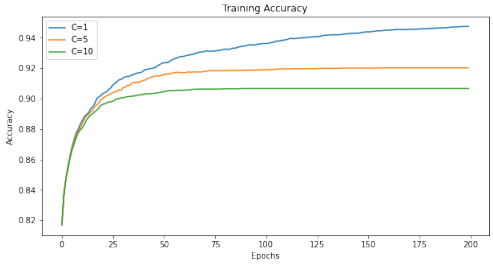

Figure 3: Evolution de la Train accuracy

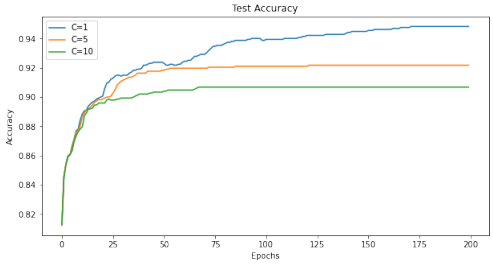

Figure 4: Evolution de la Test accuracy

Observations: Sur la base de ces graphiques, le surapprentissage ne semble pas être un problème pour ce jeu de données et pour cet algorithme. En effet, sur les graphiques correspondant au jeu de test, on n'observe ni une augmentation de la fonction de perte ni une diminution de l'accuracy trahissant tous deux un overfitting du modèle. En revanche, on observe qu'en utilsant un facteur de pénalité plus grand (C), la perte est plus élevée et l'accuracy est moins bonne. Ainsi, avec une importante pénalité, le modèle sous-performe et se retrouve à la limite du sous-apprentissage.