IFT 725 : Devoir 2
Travail individuel

Remise : 12 octobre 2012, 8h30 (**au plus tard**)

Dans ce devoir, vous devez implémenter en Python un champs Markovien aléatoire conditionnel (CRF) basé sur une structure en chaîne, pour la classification séquentielle.

L'implémentation du CRF doit être contenue dans une classe `LinearChainCRF` qui hérite de la classe `Learner` de la librairie MLPython. La définition de la classe doit être mise dans un fichier nommé `crf.py`. La classe supporte l'utilisation des hyper-paramètres suivants :

- `lr` : le taux d'apprentissage de la descente de gradient stochastique (`float`)
- `dc` : la constante de décroissement du taux d'apprentissage
- `L2` : le taux de régularisation L2 sur les matrices de poids du CRF (`float`)
- `L1` : le taux de régularisation L1 sur les matrices de poids du CRF (`float`)
- `n_epochs` : nombre d'itérations d'entraînement à effectuer (`int`)

Un squelette de la classe `LinearChainCRF` est fourni dans le fichier `crf.py` disponible sur le site web du cours. Le squelette spécifie également la signature de toutes les méthodes que vous devez implémenter. **Il est important d'utiliser la librairie Numpy dans votre implémentation, afin qu'elle soit efficace**.

La méthode `verify_gradients` est déjà implémentée. Elle compare le calcul des gradients avec une approximation par différence finie. Il est important de l'utiliser afin de vous assurer que votre implémentation de l'inférence et des gradients sont correctes. Le script `run_verify_gradients.py` vous est fourni afin de procéder à cette vérification, pour différentes configurations des hyper-paramètres. Les différences rapportées entre votre implémentation et l'approximation par différences finies devraient être plus petite que 10^{-10}.

De plus, un script `run_crf.py` est également mis à votre disposition afin d'entraîner votre CRF selon la méthode de l'arrêt prématuré (*early stopping*), sur le jeu de données *OCR Letters*. Contrairement au devoir 1, les examples sont organisés en différentes séquences, où chaque séquence correspond à une suite de caractères écrits, formant ainsi un mot. Le script nécessite comme arguments la valeur de chaque hyper-paramètre, comme suit :

```
Usage: python run_crf.py lr dc L2 L1

Ex.: python run_crf.py 0.1 0 0 0
```

Le script donnera les erreurs moyennes d'entraînement et de validation après chaque époque. De plus, il enregistrera dans un fichier nommé `results_crf_ocr_letters_sequential.txt` le résultat final de l'apprentissage, c'est-à-dire les erreurs moyennes d'entraînement, de validation et de test, et ce pour le nombre d'époques minimisant l'erreur de validation. L'écart type de la moyenne (i.e. la "*standard error*") est également donné pour chaque erreur moyenne (à utiliser pour calculer les intervalles de confiance). Les erreurs qui doivent être calculées par votre implémentation sont l'erreur de classification et la log-vraisemblance négative régularisée. Chaque nouvelle exécution du script ajoutera une ligne au fichier `results_crf_ocr_letters_sequential.txt` avec le résultat associé. L'arrêt prématuré utilise l'erreur de classification sur l'ensemble de validation pour déterminer quand arrêter et utilise un horizon (*"look ahead"*) de 5 époques.

Pour que le script puisse fonctionner correctement, vous devrez préalablement télécharger le jeu de données *OCR Letters* pour la classification séquentielle à l'aide du script `download_ocr_letters_sequential.py`

disponible sur le site web du cours. Pour cela, vous devez préalablement définir la variable d'environnement `MLPYTHON_DATASET_REPO` puis lancer le script comme suit :

```
python download_ocr_letters_sequential.py
```

Une fois votre implémentation complète, vous devez générer des résultats sur le jeu de données *OCR Letters* pour la classification séquentielle, permettant d'évaluer la performance de votre implémentation. Spécifiquement, on vous demande :

- de rapporter les taux d'erreur de classification sur l'ensemble d'entraînement et de validation pour **au moins 15 choix différents des hyper-paramètres** ;
- d'illustrer **la progression du taux d'erreur de classification en entraînement et en validation**, pour une configuration des hyper-paramètres de votre choix ;
- d'illustrer également **la progression de la log-vraisemblance négative moyenne régularisée en entraînement et en validation**, pour une configuration des hyper-paramètres de votre choix ;
- de rapporter le taux d'erreur de test **seulement pour le choix d'hyper-paramètres ayant la meilleure performance de validation** ;
- de spécifier également un **intervalle de confiance à 95%** de l'erreur de test.

Ces résultats doivent être rapportés dans un rapport suivant le format d'un article scientifique de la conférence NIPS [1]. Le document doit être en format PDF et doit avoir été généré à l'aide du langage LaTeX [2].

Le rapport doit contenir les sections suivantes :

- **Introduction** : donne un résumé de l'objectif du devoir et du contenu du rapport ($^1/_2$ page).
- **Description de l'approche** : décrit mathématiquement la méthode (au plus 4 pages), incluant
  - une description du CRF avec chaîne linéaire utilisé ;
  - une description de l'inférence dans un CRF avec chaîne linéaire ;
  - une description de l'objectif optimisé par l'apprentissage ;
  - une description du calcul des gradients ;
  - une description de tous les hyper-paramètres.
- **Expériences** : présentation des résultats de vos expériences, c'est-à-dire :
  - un **tableau** avec les erreurs d'entraînement et de validation pour au moins 15 choix d'hyper-paramètres ;
  - un **graphique** illustrant la progression du taux d'erreur de classification en entraînement et en validation, pour une configuration des hyper-paramètres de votre choix
  - un **graphique** illustrant la progression de la log-vraisemblance négative régularisée moyenne en entraînement et en validation, pour une configuration des hyper-paramètres de votre choix
  - le résultat sur l'ensemble de test **seulement** pour le choix d'hyper-paramètres ayant la meilleure performance sur l'ensemble de validation ;
  - l'intervalle de confiance de 95% pour le taux d'erreur de classification de test.

La répartition des points suivra la barème suivant :

- **10 points** pour la justesse de l'implémentation, ainsi que sa qualité (e.g. bonne utilisation de Numpy) ;
- **4 points** pour la qualité et justesse des sections **Introduction** et **Description de l'approche** du rapport ;
- **4 points** pour l'inclusion de tous les résultats attendus dans la section **Expériences** et pour leur validité.

1. Voir `http://nips.cc/PaperInformation/StyleFiles` pour obtenir les fichiers permettant de suivre ce format. Vous devrez ajouter la ligne `\nipsfinalcopy` après la commande `\author` pour que l'information soit inscrite.
2. Voir `http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/GSWLaTeX.pdf` pour une introduction et `http://www.andy-roberts.net/writing/latex/pdfs` pour savoir comment générer un document PDF à partir d'un fichier LaTeX

Le rapport peut être remis en anglais ou en français. Tout le travail de ce devoir doit être fait individuellement. Aucun code, texte du rapport ou résultats ne doivent être partagés ou transmis entre les étudiants. Par contre, les étudiants sont encouragés à discuter certains éléments de solution du devoir de vive voix.

Veuillez soumettre votre implémentation et votre rapport à l'aide de l'outil **turnin** :

```
turnin -c ift725 -p devoir_2 crf.py rapport.pdf
```

Bon travail !