# Modèles Stochastiques

## Rapport TP 4 : Champs aléatoires conditionnels (CRF)


Nom et Prénom : **DJAGHLOUL Ayoub**
Filiére : IA

# Table des matières

# 1 Utilisation de Wapiti

Wapiti est un outil qui met en œuvre les CRF en proposant plusieurs méthodes d'optimisation et de régularisation. Les CRF sont des modèles graphiques discriminants qui calculent la probabilité $P(y|x)$ d'un ensemble d'étiquettes $y = (y_1, \ldots, y_T)$ à prédire, étant donné une séquence d'observations $x = (x_1, \ldots, x_T)$, sous la forme d'un modèle exponentiel :

$$P(y|x) = \frac{\exp\left(\sum_{k=1}^{K} \theta_k F_k(x, y)\right)}{Z(\theta, x)}$$

Wapiti distingue deux types de fonctions caractéristiques $F_k(x, y)$ :

— Les traits unigrammes $f_{y,x}(y_{t-1}, y_t, x_t) = \mathbf{1}(y_t = y, x_t = x)$, notés $u$ dans les fichiers de paramètres.

— Les traits bigrammes $f_{y',y,x}(y_{t-1}, y_t, x_t) = \mathbf{1}(y_{t-1} = y', y_t = y, x_t = x)$, notés $b$ dans les fichiers de paramètres.

Il est également possible de définir des traits n-grammes au niveau des observations $x$ dans un fichier de paramètres.

# 2 Reconnaissance optique de caractères manuscrits

Pour la tâche de reconnaissance optique de caractères (OCR), les modèles CRF prennent en entrée une image en noir et blanc et produisent en sortie la suite des caractères présents sur l'image.

Le premier modèle obtenu avec seulement les quatre premiers pixels de l'image a donné des performances très faibles, avec un taux d'erreur de 81.22%. Après avoir construit un nouveau fichier de paramètres permettant d'utiliser tous les pixels de l'image, nous avons obtenu les résultats suivants :

```
Errors: 878/1081 (81.22%)
ERRORS BY TAG
a : 109 (100.00%) d : 60 (068.97%) e : 61 (039.10%)
h : 58 (100.00%) i : 91 (100.00%) n : 109 (100.00%)
o : 82 (100.00%) r : 121 (100.00%) ... t : 86 (056.58%)
```

Le taux d'erreur global sur les données de test était de 81.22%.

### 2.1 Utilisation de tous les pixels de l'image

Le modèle initial utilisant seulement les 4 premiers pixels de l'image a obtenu de très faibles performances. Nous avons donc construit un nouveau fichier de paramètres permettant d'utiliser tous les 320 pixels de l'image :

```
u0:%x[0,0]
u1:%x[0,1]
...
u319:%x[0,319]
```

Avec ce nouveau modèle, nous avons obtenu les résultats suivants :

```
Errors: 54/1081 (5.00%)
ERRORS BY TAG
a : 6 (005.50%) d : 10 (011.49%) e : 2 (001.28%)
h : 12 (020.69%) ... t : 2 (001.32%)
```

### 2.2 Prise en compte des lettres voisines

Le modèle précédent ne prenait pas en compte les lettres voisines. Nous avons donc modifié le fichier de paramètres pour intégrer une fonction caractéristique de type bigramme :

```
u0:%x[-1,0]/%x[0,0]
u1:%x[0,0]/%x[1,0]
...
u319:%x[0,319]

b
```

Avec ce modèle amélioré, nous avons obtenu un taux d'erreur sur les données de test de 4.26%.

## 3 Étiquetage par les parties du discours

Pour l'étiquetage par les parties du discours (PoS), nous avons repris les mêmes données que dans le TP précédent sur les HMM.

### 3.1 Modèle CRF avec les mêmes relations de dépendance que le HMM

Nous avons construit un premier modèle CRF prenant les mêmes relations de dépendance que le HMM du TP précédent, mais avec des relations non orientées :

```
u0:%x[0,0]

avec CRF :
Errors: 9024/81569 (11.06%)
-------------------------
|     ERRORS BY TAG     |
| ADJ  : 1496 (029.66%) |
| ADP  :  313 (002.71%) |
| ADV  :  383 (010.19%) |
| AUX  :   78 (001.93%) |
| CCONJ:   36 (002.41%) |
| DET  :  261 (002.09%) |
| INTJ :   23 (007.01%) |
| NOUN : 2005 (013.38%) |
| NUM  :  130 (010.73%) |
| PART :   20 (007.52%) |
| PRON :  670 (012.41%) |
| PROPN: 1692 (036.77%) |
| PUNCT:    1 (000.01%) |
| SCONJ:   37 (003.93%) |
| SYM  :    3 (003.95%) |
| VERB : 1621 (022.67%) |
| X    :  255 (072.86%) |
-------------------------

avec HMM:
Errors: 18553/81569 (22.75%)
-------------------------
|     ERRORS BY STATE    |
| ADJ  : 2413 (038.14%) |
| ADP  : 3314 (022.86%) |
| ADV  :  791 (019.49%) |
| AUX  : 1310 (024.74%) |
| CCONJ:   52 (003.45%) |
```

```
| DET  :  888 (006.78%) |
| INTJ :   75 (019.48%) |
| NOUN : 1321 (008.84%) |
| NUM  :  308 (022.70%) |
| PART :   65 (020.70%) |
| PRON :  524 (030.43%) |
| PROPN: 4523 (055.98%) |
| PUNCT:    0 (000.00%) |
| SCONJ:  795 (047.07%) |
| SYM  :  516 (088.66%) |
| VERB :  857 (012.98%) |
| X    :  801 (077.77%) |
-------------------------
```

Ce modèle CRF a obtenu un taux d'erreur de 11.03% sur les données de test, contre 22.75% avec le HMM.

### 3.2 Faire évoluer le fichier de paramètres pour intégrer les mots dans le voisinage immédiat et aprés une fenêtre de cinq mots

Après avoir exploré le modèle initial, nous avons évolué vers une approche qui intègre les mots dans le voisinage immédiat et une fenêtre de cinq mots dans notre modèle CRF. Voici le fichier de paramètres correspondant :

Pour le voisinage immédiat (-1 et +1) :

```
u00:%x[-1,0]
u01:%x[0,0]
u02:%x[1,0]
u03:%x[-1,0]/%x[0,0]/%x[1,0]

b

Errors: 5800/81569 (7.11%)
-------------------------
|     ERRORS BY TAG     |
| ADJ  : 1119 (022.19%) |
| ADP  :  349 (003.02%) |
| ADV  :  409 (010.88%) |
| AUX  :  108 (002.67%) |
```

```
| CCONJ:   37 (002.48%) |
| DET  :  260 (002.08%) |
| INTJ :   13 (003.96%) |
| NOUN : 1077 (007.18%) |
| NUM  :  148 (012.21%) |
| PART :   20 (007.52%) |
| PRON :  297 (005.50%) |
| PROPN:  991 (021.53%) |
| PUNCT:    1 (000.01%) |
| SCONJ:   76 (008.08%) |
| SYM  :   19 (025.00%) |
| VERB :  704 (009.84%) |
| X    :  172 (049.14%) |
-------------------------
```

Pour la fenêtre de cinq mots (-2 à +2) :

```
u00:%x[-2,0]
u01:%x[-1,0]
u02:%x[0,0]
u03:%x[1,0]
u04:%x[2,0]
u05:%x[-2,0]/%x[-1,0]/%x[0,0]
u06:%x[-1,0]/%x[0,0]/%x[1,0]
u07:%x[0,0]/%x[1,0]/%x[2,0]

b
```

En utilisant ces nouveaux paramètres, nous avons obtenu les métriques suivantes :

```
Errors: 5645/81569 (6.92%)
-------------------------
|     ERRORS BY TAG     |
| ADJ  :  942 (018.68%) |
| ADP  :  341 (002.95%) |
| ADV  :  506 (013.46%) |
| AUX  :   84 (002.07%) |
| CCONJ:   34 (002.28%) |
| DET  :  254 (002.03%) |
| INTJ :   15 (004.57%) |
| NOUN : 1025 (006.84%) |
```

```
| NUM  :  140 (011.55%) |
| PART :   19 (007.14%) |
| PRON :  278 (005.15%) |
| PROPN: 1080 (023.47%) |
| PUNCT:    3 (000.04%) |
| SCONJ:   68 (007.23%) |
| SYM  :   16 (021.05%) |
| VERB :  695 (009.72%) |
| X    :  145 (041.43%) |
-------------------------
```

### 3.3 Amélioration des performances avec un nouveau modèle

Nous avons exploré différentes techniques pour améliorer les performances de nos modèles CRF. Après plusieurs expériences (sgd-l1 , rprop ,...), nous avons identifié une combinaison de paramètres et de techniques qui a donné les meilleurs résultats sur les données de test.

Le fichier de paramètres correspondant à notre meilleur modèle est le suivant :

```
u00:%x[-2,0]
u01:%x[-1,0]
u02:%x[0,0]
u03:%x[1,0]
u04:%x[2,0]
u05:%x[-2,0]/%x[-1,0]/%x[0,0]
u06:%x[-1,0]/%x[0,0]/%x[1,0]
u07:%x[0,0]/%x[1,0]/%x[2,0]

b
```

La commande utilisée pour entraîner le modèle est la suivante :

```
wapiti-1.5.0/wapiti train -a rprop -p pos.model -d POS/fr-dev.word-pos
POS/fr-train.word-pos model23.out

wapiti-1.5.0/wapiti label -m model23.out POS/fr-test.word fr-test-voisinage4.word

 python3 compErrors.py POS/fr-test.word-pos fr-test-voisinage4.word
```

Enfin, nous avons testé d'autres algorithmes d'apprentissage, intégré des expressions régulières et des fonctions bigrammes. Les meilleures performances obtenues sont de 6.55% sur les données de test, avec le fichier de paramètres ci-dessus et en utilisant l'algorithme rprop pour l'apprentissage **-a rprop**

En utilisant ce modèle, nous avons obtenu les performances suivantes sur les données de test :

```
Errors: 5340/81569 (6.55%)
-------------------------
|     ERRORS BY TAG     |
| ADJ  :  921 (018.26%) |
| ADP  :  326 (002.82%) |
| ADV  :  418 (011.12%) |
| AUX  :  101 (002.49%) |
| CCONJ:   34 (002.28%) |
| DET  :  197 (001.58%) |
| INTJ :   21 (006.40%) |
| NOUN :  842 (005.62%) |
| NUM  :  163 (013.45%) |
| PART :   19 (007.14%) |
| PRON :  277 (005.13%) |
| PROPN: 1055 (022.92%) |
| PUNCT:    1 (000.01%) |
| SCONJ:   74 (007.86%) |
| SYM  :   48 (063.16%) |
| VERB :  653 (009.13%) |
| X    :  190 (054.29%) |
-------------------------
```

Nous avons constaté une amélioration significative par rapport aux modèles précédents, ce qui démontre l'efficacité de notre approche.