Accès à l'information - Cours sur les modèles conditionnels

François Yvon

4 janvier 2017

Des petits calculs

Dans le cas du modèle MaxEnt, nous avons montré que le gradient de $\log Z(x)$ était égal à $\mathrm{P}(y \mid x; \boldsymbol{\theta}) F(x, y)$. Montrer que le Hessien (matrice des dérivées seconde) de $\log Z_{\boldsymbol{\theta}(x))}$ est égale à la covariance des caractéristiques, soit :

$$\frac{\delta \log Z_{\boldsymbol{\theta}(x)}}{\delta \theta_i \delta \theta_j} = \mathbb{E}_{\mathrm{P}(y \mid x; \boldsymbol{\theta})}(F_i(x, y) F_j(x, y)).$$

Encore les modèles de Markov

Étant donnés deux ensembles de prénoms masculins M = { arnaud, bruno, manuel, vincent, yves } et féminins F = {anne, eva, julie, luce, marie }, on considère le mécanisme de génération suivant :

1. avec probabilité 0.2, choisir un caractère dans $[a - z]$ uniformément au hasard ; avec probabilité 0.8 choisir un prénom masculin uniformément au hasard ;
2. avec probabilité 0.2, choisir un caractère dans $[a - z]$ uniformément au hasard ; avec probabilité 0.8 choisir un prénom féminin uniformément au hasard ;
3. revenir en 1 avec probabilité 0.5, ou stopper.

Ce mécanisme produit des séquences telles que :

$$S = fdbrunopjhannedkivincentlannepxmanuelzcluce.$$

1. Quelle est la probabilité de S ?

Ce mécanisme est supposé être bruité uniformément - chaque fois qu'un caractère (à l'intérieur ou à l'extérieur d'un prénom) est écrit, il est susceptible d'être remplacé (avec probabilité 0.3) par un caractère tiré uniformément au hasard dans $[a - z]$.

1. Observant une séquence T bruitée, expliquez comment vous feriez pour trouver la séquence de prénoms la plus probable.

Regarder la télé

Au programme de la semaine : Le Fernando Pereira (`http://research.google.com/pubs/author1092.html`) parle des problèmes structurés et de l'extraction d'information : `http://videolectures.net/iiia06_pereira_slm/`. Enjoy.

Jouer avec Wapiti

Wapiti est un logiciel libre implantant fidèlement le modèle des CRF. Il est téléchargeable à l'adresse `https://github.com/Jekub/Wapiti`. La seule complexité (et richesse) de Wapiti est sa capacité à gérer des ensembles de descripteurs très riches. Les descripteurs sont introduits dans le fichier de configuration par le truchement de motifs (patterns), dont le fonctionnement est explicité ici : `https://wapiti.limsi.fr/manual.html#description`.

Vous trouvez des données libres de droit pour le problèle de la reconnaissance des NE en espagnol et en hollandais à l'adresse : `http://www.cnts.ua.ac.be/conll2002/ner/`. Sans parler un seul mot de ces langues, il devrait vous falloir moins d'une demi-heure pour construire un système de reconnaissance des NE qui fait quelque chose. Pas mal !

Vous pourrez également

1. utiliser les scripts d'évaluation des résultats pour comparer l'efficacité de jeux de caractéristiques différents et mesurer l'intérêt d'introduire des dépendances (bigrammes) entre étiquettes.
2. regarder l'effet sur la taille du modèle de différents choix pour la pénalisation