ESG Finance
Machine Learning
TD2
Classification de critiques de films

Octobre 2020

1 Introduction

On souhaite apprendre un modèle de régression logistique pour répondre à une problématique de *sentiment analysis* sur des critiques de films. Il s'agit donc d'une problématique de *classification binaire* sur des données *textuelles*. Les données pour entraîner et tester notre modèle sont issues du site IMDB [1] et peuvent être récupérées sous la forme de fichiers textes depuis le site de Stanford [2].

2 Modèle

2.1 Représentation en sac de mots

On utilisera une représentation en sac de mots (bag of words) des critiques. Chaque critique sera représentée par un vecteur dont la dimension est égale à la taille d'un vocabulaire choisi. Formellement, étant donné un vocabulaire $V = \{w_1, ..., w_{|V|}\}$ (ici, chaque w_i est donc un mot, e.g., $w_{32} =$ "$lumineux$") on représentera une critique i par un vecteur $x_i \in \mathbb{R}^{|V|}$ de dimension la taille du vocabulaire définie par :

$$x_i^k = \begin{cases} 1 \text{ si le mot } w_k \text{ apparaît dans la critique } i \\ 0 \text{ sinon} \end{cases} \quad (1)$$

1. https ://www.imdb.com/
2. http ://ai.stanford.edu/ amaas/data/sentiment/

Un ensemble de N critiques peut alors être représenté par une matrix X de taille $N \times |V|$.

2.2 Modèle

On modélise la probabilité $p(y|x)$ qu'une critique soit positive (critique avec au moins 5 étoiles sur 10) avec une régression logistique. Un modèle de régression logistique prend la forme suivante :

$$p(y_i = 1|x_i) = \sigma(f(x_i)) \tag{2}$$

où $\sigma : \mathbb{R} \to [0, 1]$ est la fonction logistique :

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

et $f(x_i) = x_i\theta^T$ une fonction linéaire de vecteur de paramètres $\theta \in \mathbb{R}^{|V|}$.

2.3 Estimation des paramètres du modèle

Etant fixé le jeux de données d'entrainement D, on cherche le vecteur de paramètre θ qui maximise la vraisemblance des données $L(D, \theta)$:

$$\theta^* = \arg\max_{\theta} L(D, \theta) \tag{4}$$

On suppose les observations indépendantes et identiquement distribuées. On peut donc poser :

$$L(D, \theta) = \prod_{i=1}^{n} p(y_i|x_i) = \prod_{i=1}^{n} p(y_i = 1|x_i)^{y_i}(1 - p(y_i = 1|x_i))^{1-y_i} \tag{5}$$

On minimisera en pratique l'opposé de la log-vraisemblance $l(D, \theta)$ (negative log likelyhood) :

$$-l(D, \theta) = \sum_{i=1}^{n} y_i \log p(y_i = 1|x_i) + (1 - y_i) \log(1 - p(y_i = 1|x_i)) \tag{6}$$

La log-vraisemblance pour notre modèle de régression logistique peut s'exprimer de la façon suivante :

$$l(D, \theta) = -\sum_{i=1}^{n} y_i \log \sigma(f(x_i)) + (1 - y_i) \log(1 - \sigma(f(x_i))) \tag{7}$$

$$= \sum_{i=1}^{n} y_i x_i \theta^T - \log(1 + e^{x_i\theta^T}) \tag{8}$$

On donne la dérivée de la log-vraisemblance :

$$\frac{\partial l(D,\theta)}{\partial \theta} = \mathbf{X}^T(\mathbf{y} - \mathbf{p}) \tag{9}$$

où $\mathbf{X}$ est la matrice de taille $n \times d$ encodant les observations, $\mathbf{y} = (y_1, y_2, ..., y_n)$ le vecteur de classes encodant la vérité terrain et $\mathbf{p}$ le vecteur de probabilité renvoyé par le modèle, i.e., $\mathbf{p}_i = \sigma(f(x_i))$. La maximisation de log-vraisemblance (la minimisation de l'opposé de la log-vraisemblance) sera réalisée avec une méthode de descente de gradient.

3 Pratique

Vous apprenez un modèle de régression logistique sur un sous ensemble des données IMDB en prenant soin de séprarer données de TRAIN, de VAL et TEST. Quelle est la taille de votre vocabulaire ? Pourquoi ? Quelle métriques vous semblent pertinentes pour évaluer votre modèle ? Quelle est alors sa performance ? Quels sont les top-10 mots qui ont le plus de poids pour classifier une critique "POSITIVE" ? "NEGATIVE" ? Comment améliorer le modèle ?