INF-6422: Rapport laboratoire 2

Rendu le Mardi, 16 février 2016

à *François Labrèche*

Thomas Luinaud, Paul Berthier

Table des matières

1. Mise en contexte

1.1

Le spam, ou encore courriel indésirable, peut être défini comme étant une communication électronique non sollicitée. Selon le rapport Q1-2014 de Kaspersky, le spam représenterait aujourd'hui plus de la moitié du trafic électronique. Souvent utilisé pour des fins commerciales, le spam peut aussi être utilisé pour fins d'escroquerie ou prendre la forme d'hameçonnage (phishing en anglais) afin de tromper le destinataire dans le but d'obtenir des informations personnelles.

1.2

Bien que le coût d'envoi d'un message électronique puisse être négligeable pour les spammeurs, celui associé à sa réception peut causer des coûts non négligeables tant aux destinataires qu'aux prestataires de services compte tenu du volume élevé d'envoi. Plusieurs méthodes de détection ont par conséquent été développées afin de filtrer les messages indésirables. Une première catégorie de filtres consiste à bloquer les messages sur la base d'une liste. Cette méthode peut elle-même être divisée en plusieurs techniques, soit le recours aux listes noires (blacklisting), aux listes blanches (whitelisting), aux listes grises (greylisting), etc. Une autre approche qui a démontré son efficacité consiste à filtrer les messages sur la base de leur contenu en utilisant des méthodes d'apprentissage automatique. Le présent travail pratique vous permettra de vous familiariser avec certaines méthodes d'apprentissage automatique et d'en évaluer la performance dans un contexte de détection de spam.

2. Méthode statistique

Une approche souvent utilisée revient à filtrer les messages électroniques sur la base de leur contenu (content-based filtering). Un exemple classique consiste à filtrer les messages en fonction de la fréquence d'apparition de certains mots. Utilisez le fichier spambase afin d'appliquer une méthode statistique qui vous permettra de classifier les messages en deux catégories, soit spam (1) ou non spam (0). Compte tenu de la nature binomiale (0/1) de la variable dépendante, la régression logistique peut être utilisée comme méthode de classification. Votre variable dépendante (0/1) représente la catégorie associée (spam ou non spam) et les variables indépendantes représentent la fréquence d'apparition de certains mots.

2.1

Effectuez une régression logistique en utilisant l'ensemble des variables (57) contenues dans le fichier spambase. Divisez les données afin d'en utiliser 66phase d'apprentissage de votre modèle. À quoi servira l'autre 33Donnez pour chaque variable indépendante son coefficient ainsi que son « odd ratio » associé. Quelle est la signification de ces deux valeurs ?

2.2

Évaluez et discutez des performances de votre modèle en termes de : taux de vrai positif, taux de faux positif, précision, sensibilité (recall), « F-measure » et l'aire sous la courbe (ROC area). Expliquez la signification de chacune de ces mesures. Donnez la matrice de confusion de votre modèle et expliquez ce qu'elle représente.

2.3

Donnez un exemple de contre-mesure de type Tokenization attack qu'un spammeur pourrait facilement utiliser afin de contourner un filtre basé uniquement sur la fréquence d'apparition de certains mots. Votre méthode ne doit pas modifier la signification du message et ne doit pas ajouter de nouveaux mots. Appliquez votre méthode au message suivant :

DEAR RECEIVER,

You have just received a Taliban virus. Since we are not so technologically advanced in Afghanistan, this is a MANUAL virus. Please click on this link (http ://clickme.com) to delete all the files on your hard disk yourself and send this mail to everyone you know. Thank you very much for helping us.

-Taliban hacker.

3. Apprentissage automatique

Une autre catégorie de méthodes qui a fait ses preuves en détection de spam consiste à s'inspirer de l'intelligence artificielle et de recourir à des techniques dites d'apprentissage automatique (machine learning). Ces méthodes ont l'avantage de s'adapter et d'apprendre pour continuellement améliorer leur performance.

3.1

Les méthodes d'apprentissage automatique peuvent être divisées en plusieurs catégories. Parmi les plus courantes, nous retrouvons l'apprentissage non-supervisé, semi-supervisé, et supervisé. Expliquez les caractéristiques de chacune de ces méthodes.

3.2

La classification naïve bayésienne (naive bayes classifier) est un exemple de méthode qui peut être utilisée afin de résoudre des problèmes de classification par apprentissage supervisé. Appliquez cette méthode au fichier spambase afin de filtrer les messages en fonction des 57 variables continues. Utilisez 66d'apprentissage.
Évaluez et discutez des performances de votre modèle en termes de : taux de vrai positif, taux de faux positif, précision, sensibilité (recall), « F-measure » et l'aire sous la courbe (ROC area). Donnez la matrice de confusion.

3.3

Une autre méthode de classification, les forêts d'arbres décisionnels (random forest), consiste à effectuer un apprentissage sur plusieurs arbres de décisions. Appliquez cette méthode au fichier spambase et utilisez les 57 variables continues. Utilisez 66% des données pour la phase d'apprentissage. Évaluez et discutez des performances de votre modèle en termes de : taux de vrai positif, taux de faux positif, précision, sensibilité (recall), « F-measure » et l'aire sous la courbe (ROC area). Donnez la matrice de confusion.

3.4

Donnez un exemple de contre-mesure de type Statistical attack qu'un spammeur pourrait utiliser afin d'échapper à un filtre basé sur la fréquence des mots en utilisant une méthode d'apprentissage automatique. Votre exemple de contre-mesure peut impliquer de modifier le contenu du message. Proposez une solution qui permettrait de contrecarrer cette contre-mesure.

4. Performance et optimisation

2.1

Comparez et discutez, en termes de performance, les résultats que vous avez obtenus pour les différentes méthodes utilisées (régression logistique, classification naïve bayésienne, forêts d'arbres décisionnels). Selon vos résultats, quelles méthodes semblent donner les meilleures performances pour le jeu de données spambase ?

2.2

Nommez un avantage et un inconvénient d'utiliser des filtres basés sur l'apprentissage machine supervisé. Comparez avec l'apprentissage machine non supervisé en donnant un avantage et un inconvénient dans un contexte de détection de spam. Selon vous, est-ce qu'une méthode semble plus appropriée ? Justifiez votre réponse et n'oubliez pas de donner vos références.

2.3

Le jeu de données spambase est principalement basé sur la fréquence d'apparition de certains mots et sur la présence de lettres majuscule. Donnez au moins deux autres exemples de caractéristiques (features) qui pourraient être prises en compte afin d'améliorer la performance des modèles de classification que vous avez développés. Donnez vos références.