Cours

Charles Vin

Date

1 Généralité

Métrique

- Taux de reconnaissance : $\frac{bonnepred}{totalpred}$
- Précision pour une classe c : $\frac{N^C_{correct}}{N^C_{predits}}$
- Rappel / Recall pour une classe c : $\frac{N^C_{correct}}{N^C_{tot}}$
- F1 : need plus de vulgarisation
- ROC : Faux positif VS Vrais positif
- AUC :

Problème d'équilibre des classes

- Accuracy poubelle $\rightarrow$ Utiliser les autres metrics : AUC / ROC
- Ré-équilibrer le jeux de données : supprimer des données dans la classe majoritaire
- Fonction de coût : pénaliser plus les erreurs dans la classe minoritaire (cours 1 diapo 51)

Problème de dimension Ajouter un terme sur la fonction coût (ou vraisemblance) pour pénaliser le nombre (ou le poids) des coefficients utilisés pour la décision. (Cours 1 diapo 50)

TF-IDF encoding if word k is in most documents, it is probably useless $\rightarrow$ TF-Idf encoding : Donne plus de poids au keyword et un petit peu moins au stopword. A combiner avec blacklist

2 Bag of Word

2.1 Stengths and drawbacks

Avantage :

- Easy light fast
- Opportunity to enrich (Context encoding, Part Of Speech)
- Efficient implementation
- Still very effective with classification

Limite :

- Loose document / sentence structure : mitigated with N-gram
- Several task missing : POS tagging, text generation
- Semantic gap : On peut pas utiliser la distance euclidienne pour mesurer la différence sémantique

2.2 Classification

Naive Bayes Rapide, interprétable, naturellement multiclasse. Perf à améliorer. Bien filtrer les stop word. Extention par robustesse (?)

Classifieur linéaire Scalable, attention sensible au dimension