Partie 3/3 : Apprentissage actif pour les données textuelles

François HU - Data scientist au DataLab de la Société Générale Assurances - *19/11/19* -
https://nbviewer.jupyter.org/github/curiousML/DSA/tree/master/text_mining/ (https://nbviewer.jupyter.org/github/curiousML/DSA/tree/master/text_mining/)

Out[3]: cacher / afficher code

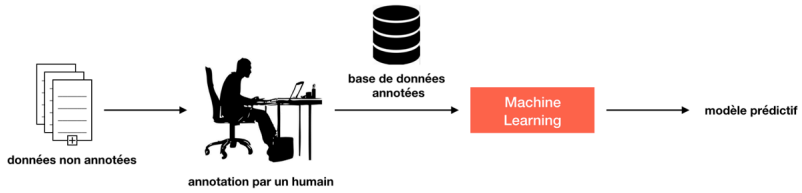
Out[1]:

1. Contexte et objectifs

- **Classifier automatiquement des données textuelles volumineuses** : ces données sont labellisées par des experts humain (dans la littérature : "**oracles**")

Processus classique d'apprentissage statistique passif (**passive machine learning**) :

Out[5]:



- **Problèmes** : processus d'annotation peut être trop **long**, trop **difficile** et/ou trop **coûteux**

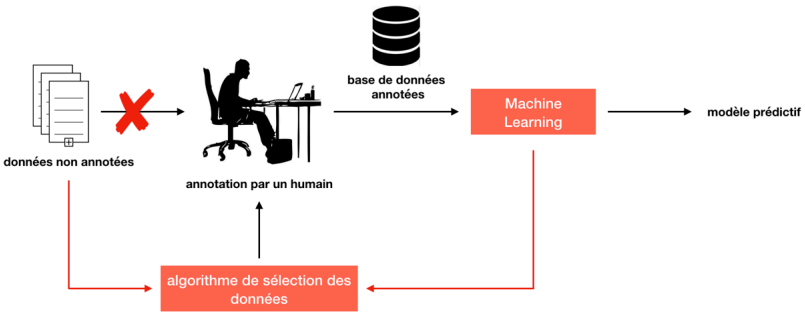
- **Question** : pouvons-nous entraîner les modèles avec moins de données annotées et avec moins d'intervention humaine ?

2. Active Learning

- Active Learning : Modèle plus performant avec moins d'entraînement si nous avons la possibilité de choisir les données à entraîner

Processus d'apprentissage statistique actif (**active machine learning**) :

Out[11]:



- Pour l'apprenant actif (active learner), il existe différentes **stratégies de requête** pour décider la donnée la plus informative (nous voulons éviter de requêter sur des données redondantes ou non-informatives)

3. Stratégies d'échantillonnage

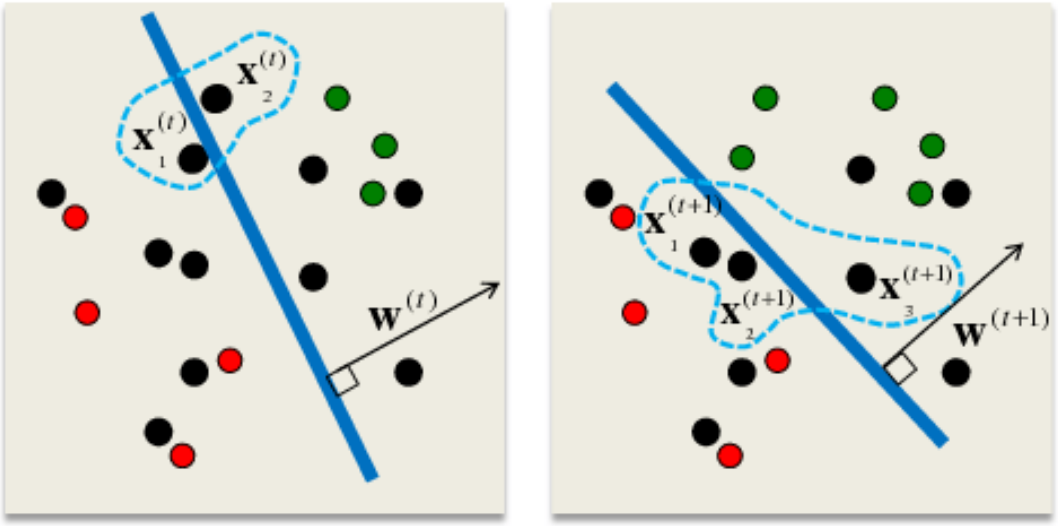
Notations (supposons que nous avons un petit ensemble de données labélisées) :

- x_A^* l'instance la plus informative (i.e. la meilleure requête) selon une stratégie de séléction A ;
- $\mathcal{L} = \{x_i^*, y_i\}_{i=1,\cdots,|\mathcal{L}|}$, l'ensemble des données labellisées ;
- $\mathcal{U} = \{x_i\}_{i=1,\cdots,|\mathcal{U}|}$, l'ensemble des données non labellisées ;
- θ, le modèle actuel ;
- C, l'ensemble des labels. Dans le cas binaire, les classes 0 et 1 (ou -1 et 1) sont souvent utilisées

nous allons voir **deux stratégies classiques** : échantillonnage **basé sur l'incertitude du modèle** et échantillonnage **basé sur le désaccord des modèles**

Echantillonnage basé sur l'incertitude (Uncertainty-based sampling)

Out[5]:



(Prateek Jain, Sudheendra Vijayanarasimhan et Kristen Grauman)

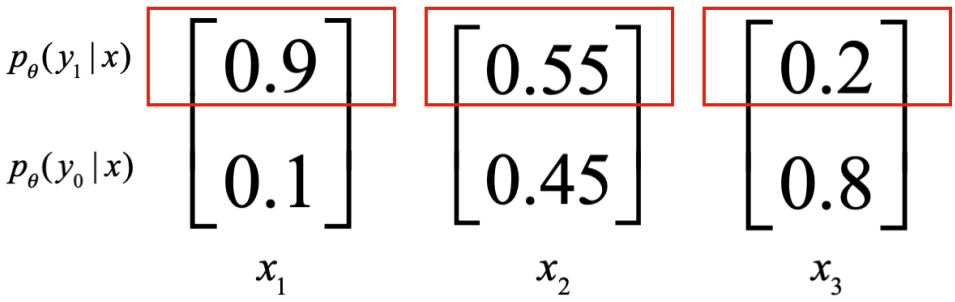
Echantillonnage "*least confident*"

- L'apprenant actif requête les instances pour lesquelles il n'est pas certain comment labelliser (Lewis and Gale, 1994).
- Dans le cas d'un problème de :
  - **classification binaire**, l'échantillonnage incertain requête les instances où leur probabilité à posteriori d'être positif est proche de 0.5
  - **classification multi-classe**, l'échantillonnage incertain requête les instances où leur prédiction est la moins confiante

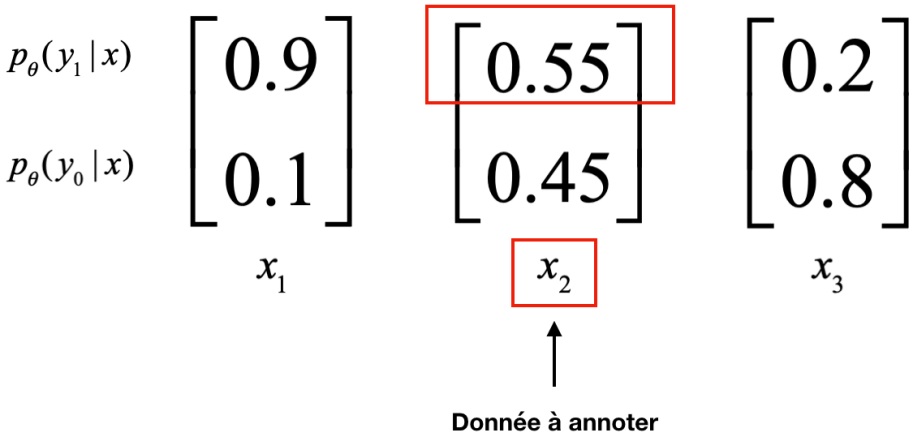
$$x_{LC}^* = \arg\max_{x \in \mathcal{U}} \left\{1 - P_\theta(\hat{y} \mid x)\right\}$$

où $\hat{y} = \arg\max_y \{P_\theta(y \mid x)\}$ est la classe avec la plus grande probabilité à postériori sous le modèle θ.

Out[6]:

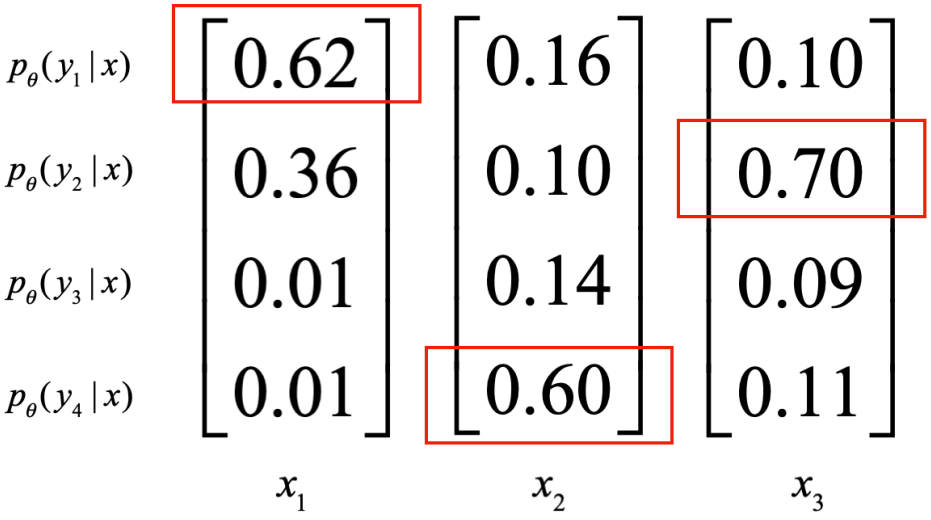


Out[7]:

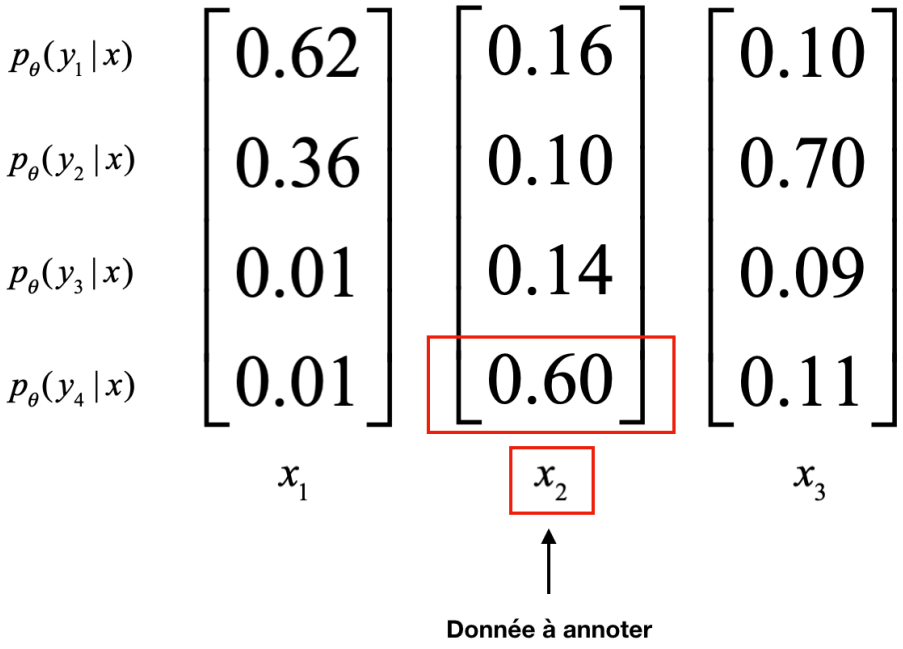


Cette stratégie considère seulement l'information sur le label le plus probable mais ne tient pas en compte les autres labels

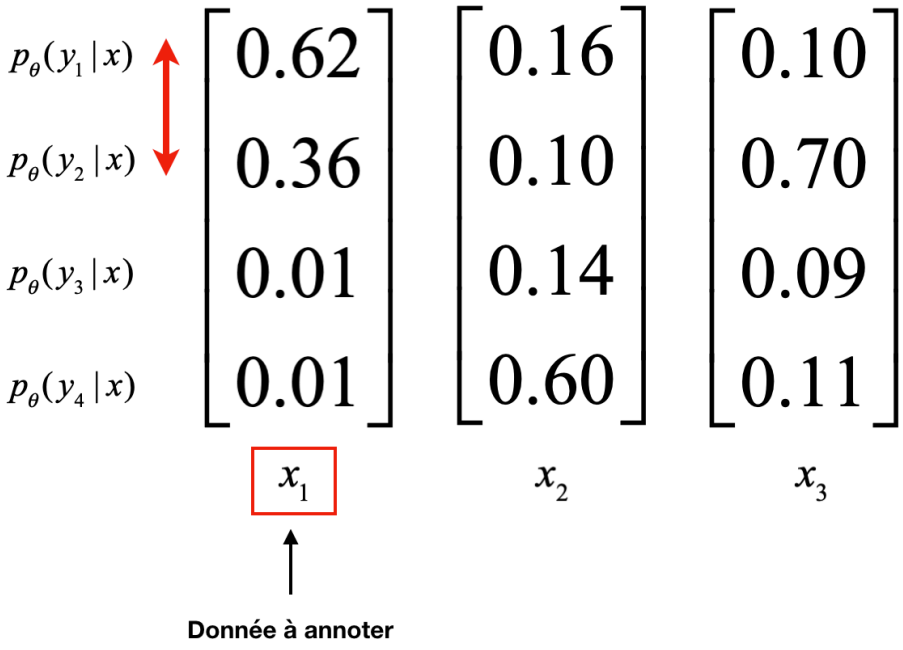
Out[8]:



Out[9]:



Out[10]:



Echantillonnage avec marge

- (Scheffer et al, 2001) propose une stratégie pour remédier ce problème : l'**échantillonnage avec marge** (**margin sampling**)

$$x_M^* = \arg\max_{x \in \mathcal{U}} \left\{ P_\theta(\hat{y_1} \mid x) - P_\theta(\hat{y_2} \mid x) \right\}$$

  où $\hat{y_i}$ est la i-ème plus probable classe à postériori sous le modèle θ.

- L'apprenant est donc :
  - **plus certain** pour les données avec de **grandes marges** car il y a une grande différence de probabilité
  - **moins certain** sur les **petites marges**
- Cette stratégie aide l'apprenant à discriminer entre deux classes les plus probables mais cette approche continue d'ignorer la distribution des sorties pour les classes restantes (et cela peut poser problème si nous avons un très grand ensemble de labels)

Echantillonnage par entropie

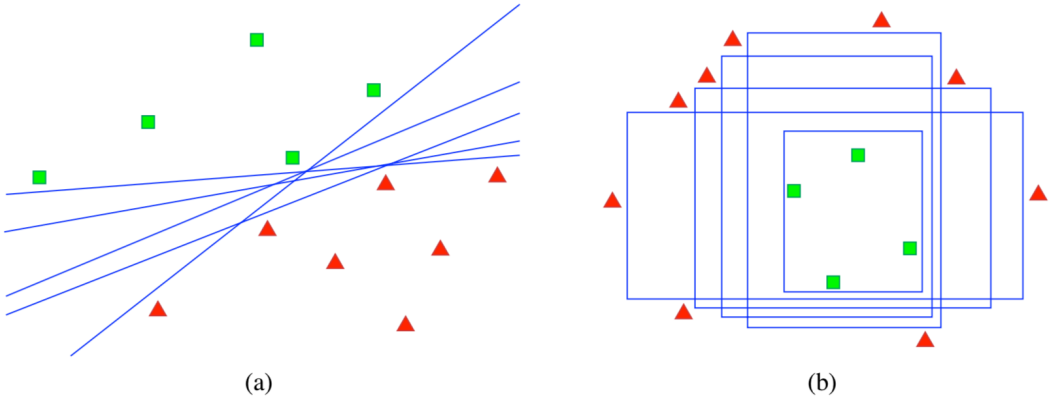
- Une stratégie d'échantillonnage incertain plus générale (Shannon, 1948) utilise l'**entropie** comme mesure d'incertitude

$$x_H^* = \arg\max_{x \in U} \left\{ -\sum_{i=1}^{|C|} P_\theta(y_i \mid x) \log P_\theta(y_i \mid x) \right\}$$

- Une entropie est une mesure d'information théorique qui représente la quantité d'information nécessaire pour "encoder" une distribution
- Pour une classification binaire, les trois stratégies ci-dessus sont équivalentes
- Pour une classification multi-classe, l'approche entropique se généralise bien (voir Setteles et Craven (2008), pour les travaux sur les séquences)
- Intuitivement :
  - l'entropie semble être plus appropriée si la fonction objectif est de minimiser la **perte logistique** (**log-loss**)
  - l'échantillonnage incertain et l'échantillonnage de marge semblent plus appropriées si nous voulons réduire l'**erreur de classification** (**accuracy**).

Echantillonnage basé sur le désaccord (Query by committee)

Out[11]:

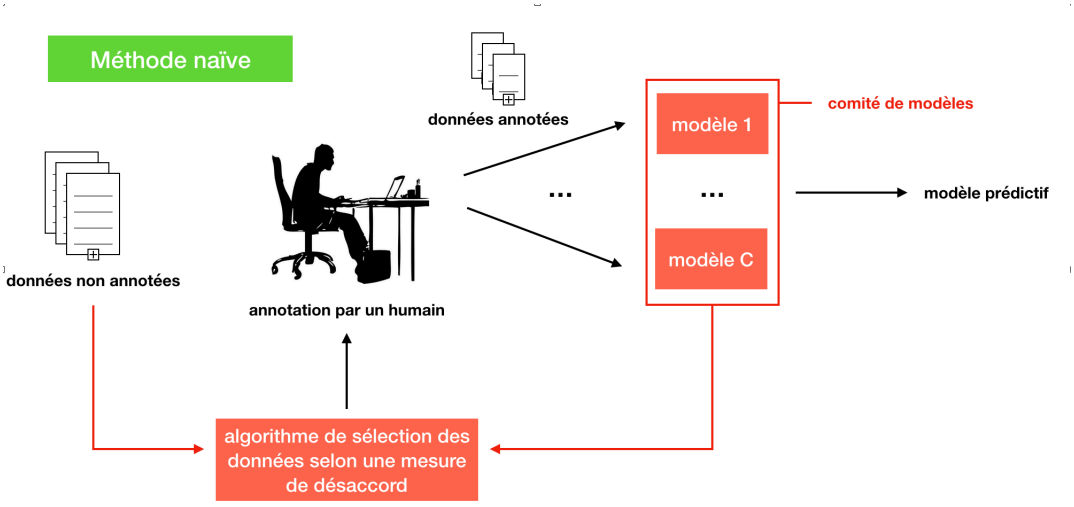


(Settles, 2010)

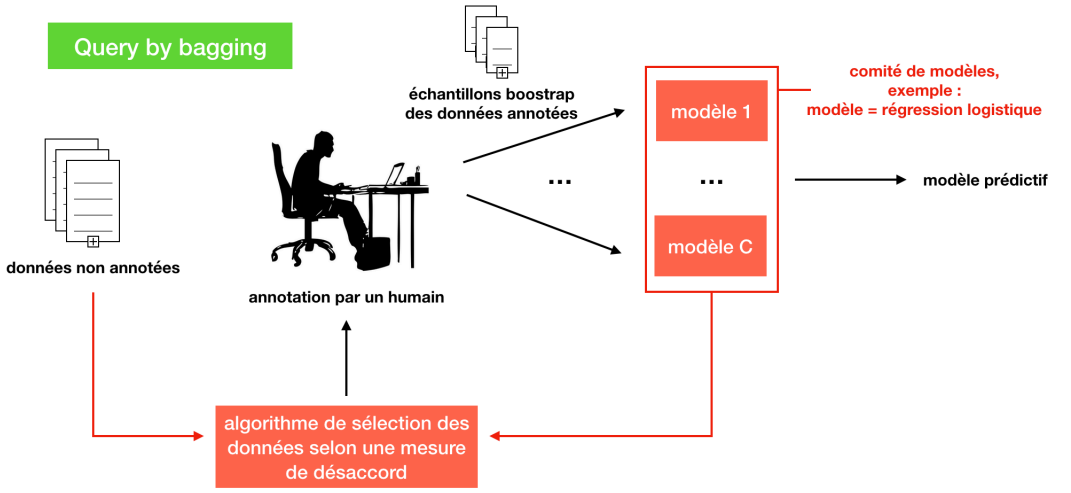
- L'approche QBC (Seung et al., 1992) consiste à **mettre en place, entraîner et maintenir** plusieurs modèles d'apprenant $\Theta = \{\theta^{(1)}, \cdots, \theta^{(N)}\}$.
- Pour $i \in \{1, \cdots, C\}$, chaque modèle $\theta^{(i)}$ peut voter sur le label de chaque candidat de la requête
- La requête la plus informative est l'**instance où les modèles sont le plus en désaccord**
- **Objectif en apprentissage :** donner le meilleur modèle dans l'espace de versions. **Objectif en apprentissage actif :** contraindre le plus possible la taille de l'espace avec le moins d'instances labellisées que possible.
- Afin d'implémenter un algorithme de séléction QBC nous devons préalablement :
  1. construire un comité de modèles qui représentent différentes régions de l'espace des versions (Mitchelle, 1982).
  2. avoir une mesure de désaccord parmi les membres du comité
- Cependant, il ne semble pas avoir de méthode idéale, ni un nombre optimal de membres à utiliser, car les résultats dépendent fortement de l'application.

1. Quel comité de modèles choisir ?

Out[12]:



Out[13]:



2. Quelle mesure de désaccord choisir ?

- Il existe différentes approches pour mesurer le désaccord. Nous allons seulement présenter les deux méthodes les plus connues :
  - Le **vote entropy** (Dagon et Engelson, 1995) vue comme une généralisation de x_H^* :

$$x_{VE}^* = \arg\max_{x \in U'} \left\{ -\sum_{i=1}^{|C|} \frac{V(y_i)}{N} \log \frac{V(y_i)}{N} \right\}$$

    où N est la taille de la comité et $V(y_i)$ le nombre de vote pour la classe y_i.

  - La **divergence moyenne de Kullback Leibler** (MacCallum et Nigam, 1998) :

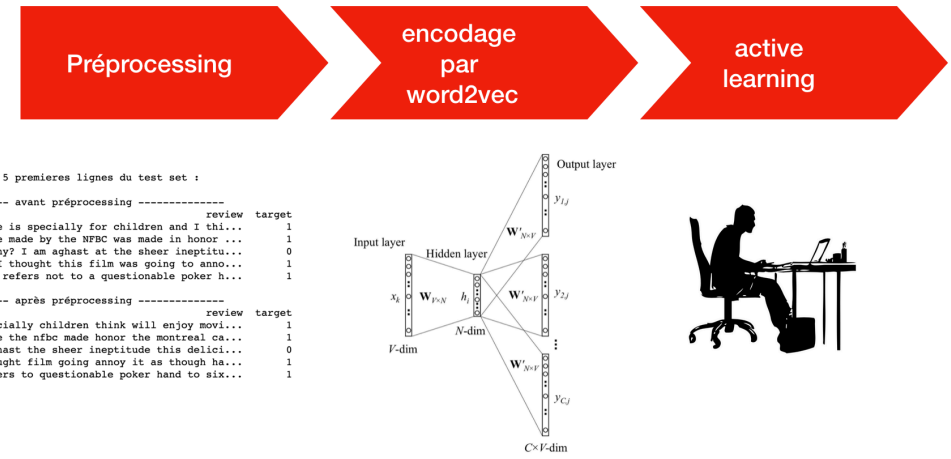
$$x_{KL}^* = \frac{1}{N} \sum_{i=1}^{N} D(P_{\theta^{(i)}} \,||\, P_N)$$

    où $D(P_{\theta^{(i)}} \,||\, P_N) = \sum_{j=1}^{|C|} P_{\theta^{(i)}}(y_j \mid x) \log \frac{P_{\theta^{(i)}}(y_j \mid x)}{P_N(y_j \mid x)}$

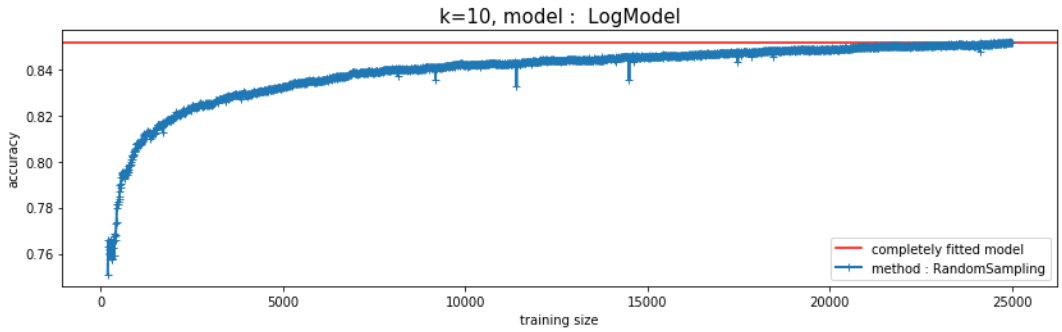
    avec $P_N(y_j \mid x) = \frac{1}{N} \sum_{j=1}^{|C|} P_{\theta^{(i)}}(y_j \mid x)$

4. Expérimentations

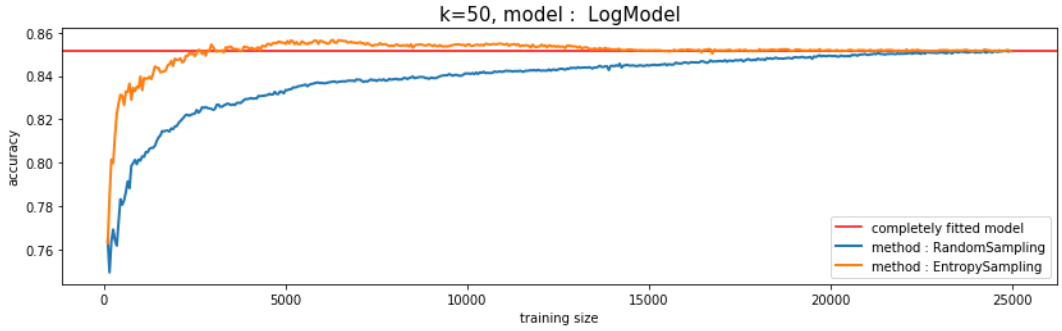
Out[15]:



Out[14]:



Out[17]:



résumé :

- **Active learning**
- deux méthodes classiques : **échantillonnage incertain, échantillonnage basé sur le désaccord**
- par rapport à un échantillonnage aléatoire, nous pouvons obtenu la même performance du modèle tout en **réduisant considérablement la taille des données d'apprentissage**

Conclusion

- **préprocessing** des données textuelles
- NLP dans un **contexte non-supervisé** : word embedding, topic modeling
- NLP dans un **contexte supervisé** : RNN, DRNN, BRNN, LSTM, GRU
- NLP dans un **contexte semi-supervisé** : active learning pour la classification de texte