3 Détection de la langue par comptage des lettres

Formalisation simplifiée du problème

Un message $\mathbf{s}$ est une suite de N caractères pris dans un alphabet $\mathcal{A}$ composé de α caractères : $s_1, \cdots, s_N$ avec $s_j \in \mathcal{A}$ (par exemple $\mathcal{A} = \{A, B, ..., Z\}$, $\alpha = 26$, les espaces entre mots ne sont pas pris en compte dans cette approche.). L'espace Ω des messages de longueur N contient α^N éventualités dont la probabilité dépend de la langue. On observe un message $\mathbf{m} = m_1, \cdots, m_N$; l'objectif est de déterminer, avec un taux d'erreur minimum, la langue de ce message.

Le message est écrit dans une langue l avec $l \in \{0, 1\}$. Les différentes hypothèses pour la langue sont notées H_l : sous l'hypothèse H_l, le message est écrit dans la langue l et ses caractères ont des probabilités *a priori* $\left\{p_c^{(l)}\right\}_{c\in\mathcal{A}}$ supposées connues.

$f_c(\mathbf{m})$, la fréquence empirique du caractère c dans le message $\mathbf{m}$, est l'estimation de cette loi obtenue par comptage des différents caractères dans le message observé $\mathbf{m}$. La fréquence empirique de chaque caractère $c \in \mathcal{A}$ est donnée par :

$$f_c(\mathbf{m}) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{m_j=c}\,,\ c \in \mathcal{A}$$

Remarques :

— Lorsque la longueur de message tend vers l'infini, pour un message $\mathbf{m}$ écrit dans la langue l, cette estimation $\{f_c(\mathbf{m})\}_{c\in\mathcal{A}}$ tend vers $\left\{p_c^{(l)}\right\}_{c\in\mathcal{A}}$ (loi des grands nombres).

— $f_c(\mathbf{m})$ est le nombre d'occurrences du caractère c dans le message $\mathbf{m}$ divisé par N.

La probabilité *a priori* (avant observation d'un message) pour que le message soit écrit dans la langue l est notée $\mathcal{P}_{H_l}$ (probabilité *a priori* de la langue l). Les $\mathcal{P}_{H_l}$ sont supposées connues (sans *a priori* on choisit $\mathcal{P}_{H_0} = \mathcal{P}_{H_1} = 1/2$).

On note $\widehat{H_\lambda}$ l'hypothèse décidée sur la langue et H_l l'hypothèse effective pour le message observé.

On note $\Pr\left[\widehat{H_\lambda}|H_l\right]$ la probabilité de décider la langue λ (décider l'hypothèse H_λ) sachant que la langue l est utilisée (hypothèse H_l effective). La probabilité d'erreur s'écrit :

$$P_e = \mathcal{P}_{H_0}\Pr\left[\widehat{H_1}|H_0\right] + \mathcal{P}_{H_1}\Pr\left[\widehat{H_0}|H_1\right] \quad \text{(Probabilité d'erreur.)} \tag{1}$$

Le but de ce TD est de répondre à deux questions :

— Dans un message donné, comment détecter la langue (avec un taux d'erreur minimum) à partir de la fréquence empirique d'apparition des différents caractères ?

— Quelle est la fiabilité (probabilité d'erreur P_e) de cette décision ?

L'idée consiste à partitionner l'espace Ω en 2 régions de décision D_0 et D_1 de telle sorte que :

— si le message $\mathbf{m}$ observé appartient à D_j on décide l'hypothèse H_j.

— les régions D_j sont choisies de sorte que la probabilité d'erreur P_e soit minimale. $\{D_j\}_{j\in\{0,1\}}$ est l'ensemble des messages pour lesquels on décide H_j.

Questions :

Pour simplifier la démarche, on modélise les caractères comme des variables aléatoires indépendantes de même loi.

En notant $\Pr[\mathbf{z}|H_l] = P_l(\mathbf{z})$ la probabilité du message $\mathbf{z}$ sous l'hypothèse H_l :

$$\Pr\left[\widehat{H_\lambda}|H_l\right] = \sum_{\mathbf{z}\in D_\lambda} \Pr[\mathbf{z}|H_l]\,.$$

1. Proposer une solution pratique pour la détection de la langue qui n'utilise que l'histogramme empirique des caractères dans le message et les histogrammes $a\ priori$ des deux langues en concurence.
2. Exprimer la probabilité d'erreur P_e en fonction de $\mathcal{P}_{H_0}$, $\mathcal{P}_{H_1} = 1 - \mathcal{P}_{H_0}$, $P_0(\mathbf{z})$ et $P_1(\mathbf{z})$ en sommant sur les ensembles D_0 et D_1 ; puis seulement sur l'ensemble D_1.
3. En déduire que la région D_1 à choisir pour que la probabilité d'erreur soit minimale est

$$D_1 = \{\mathbf{z} \in \Omega / \mathcal{P}_{H_0} P_0(\mathbf{z}) - \mathcal{P}_{H_1} P_1(\mathbf{z}) < 0\}\,.$$

4. Combien y-a-t-il d'éléménts dans Ω ? On considère un algorithme qui précalcule $\mathcal{D}_1$ pour tester ensuite l'appartenance d'un mot à $\mathcal{D}_1$. Quel est son coût ? Qu'en pensez-vous ?
5. Les caractères du message étant supposés indépendants, exprimer le test d'appartenance du message $\mathbf{m}$ observé à la région D_1 en fonction de $\mathcal{P}_{H_0}$, $\mathcal{P}_{H_1}$ et des $p_{m_k}^{(0)}$, $p_{m_k}^{(1)}$?
6. Réécrire le test établi à la question précédente en fonction des fréquences empiriques $f_c(\mathbf{m})$ des caractères dans le message observé $\mathbf{m}$.
7. Réécrire le test en fonction des deux divergences de Kullback $D\left(f_c(\mathbf{m})||p_c^{(1)}\right)$ et $D\left(f_c(\mathbf{m})||p_c^{(0)}\right)$.
   Commentez le cas sans $a\ priori$ $\mathcal{P}_{H_1} = \mathcal{P}_{H_0}$ et la manière dont l'$a\ priori$ impacte le test.
8. Donner un algorithme implémentant ce test et donner son nombre d'opération arithmétiques. Comparer au coût du test précédent (question 3).
9. Que devient le test lorsque deux langues ont des caractères distribués selon la même loi de probabilité $a\ priori$? Et si, de plus, $\mathcal{P}_{H_1} = \mathcal{P}_{H_0}$?
10. Si la performance obtenue s'avère insuffisante par rapport à l'objectif souhaité, quelle possibilité d'amélioration proposer ?
11. En supposant les caractères indépendants, montrer que la probabilité d'erreur s'écrit

$$P_e = \frac{1}{2} - \frac{1}{2} \sum_{\mathbf{z}\in\Omega} \left| \mathcal{P}_{H_1} \prod_{c\in\mathcal{A}} {p_c^{(1)}}^{N f_c(\mathbf{z})} - \mathcal{P}_{H_0} \prod_{c\in\mathcal{A}} {p_c^{(0)}}^{N f_c(\mathbf{z})} \right|$$

Que peut-on dire de l'utilisation pratique de ce résultat ?