Méthodes Statistiques pour la Modélisation du Langage Naturel

Jean-Philippe Vert
(Directeur : Olivier Catoni)
Ecole Normale Supérieure
Département de Mathématiques et Applications

Jean-Philippe.Vert@mines.org
http ://www.dma.ens.fr/users/vert

30 mars 2001

Introduction

Contexte et Motivations

Modélisation statistique du langage naturel

Modélisation statistique du langage naturel

– But = décrire de manière probabiliste des contraintes du langage

Modélisation statistique du langage naturel

- But = décrire de manière probabiliste des contraintes du langage
- **Un modèle statistique est un processus stochastique P qui assigne une probabilité $P(T_1, \ldots, T_n)$ à toute suite de lettres (ou de mots)**

Modélisation statistique du langage naturel

- But = décrire de manière probabiliste des contraintes du langage
- **Un modèle statistique est un processus stochastique P qui assigne une probabilité $P(T_1, \ldots, T_n)$ à toute suite de lettres (ou de mots)**
- Il est censé "imiter" le processus d'écriture

Modélisation statistique du langage naturel

- But = décrire de manière probabiliste des contraintes du langage
- **Un modèle statistique est un processus stochastique P qui assigne une probabilité $P(T_1, \ldots, T_n)$ à toute suite de lettres (ou de mots)**
- Il est censé "imiter" le processus d'écriture
- Il peut être défini à plusieurs niveaux : pour une langue, un auteur, un thème, etc...

A quoi ça sert ? (1)

– Compression et transmission (cf Shannon)

A quoi ça sert ? (1)

– Compression et transmission (cf Shannon)
– Cadre Bayésien (reconnaissance de la parole, OCR, traduction automatique...) :

$$T^* = \arg\max_T P(T \mid I)$$
$$= \arg\max_T P(T)P(I \mid T)$$

(I =Input, T^* = texte résultat).

A quoi ça sert ? (2)

A quoi ça sert ? (2)

– Cadre Bayésien (bis) (classification de textes...)

$$k^* = \arg\max_{i=1,\ldots,k} P(i \mid W)$$

$$= \arg\max_{i=1,\ldots,k} P(i) P(W \mid i)$$

A quoi ça sert ? (2)

– Cadre Bayésien (bis) (classification de textes...)

$$k^* = \arg\max_{i=1,\ldots,k} P(i \mid W)$$

$$= \arg\max_{i=1,\ldots,k} P(i)P(W \mid i)$$

– Pour représenter un texte, il faut des **modèles locaux**

Modèles actuels

– Suprématie des **N-grammes**

Modèles actuels

– Suprématie des **N-grammes**
– Difficultés :

Modèles actuels

- Suprématie des **N-grammes**
- Difficultés :
  - Corrélations à longue distance

Modèles actuels

- Suprématie des **N-grammes**
- Difficultés :
  - Corrélations à longue distance
  - Dépend beaucoup du contexte

Modèles actuels

- Suprématie des **N-grammes**
- Difficultés :
  - Corrélations à longue distance
  - Dépend beaucoup du contexte
  - Très grandes dimensions

Modèles actuels

- Suprématie des **N-grammes**
- Difficultés :
  - Corrélations à longue distance
  - Dépend beaucoup du contexte
  - Très grandes dimensions
  - Taille des corpus d'entraînement limitée

But de cette thèse

But de cette thèse

– Proposer des méthodes d'estimation pour apprendre $P(T)$ dans un cadre :

But de cette thèse

- Proposer des méthodes d'estimation pour apprendre $P(T)$ dans un cadre :
  - non-paramétrique (peu d'hypothèses sur P) ;

But de cette thèse

- Proposer des méthodes d'estimation pour apprendre $P(T)$ dans un cadre :
  - non-paramétrique (peu d'hypothèses sur P) ;
  - non-asymptotique (corpus limité)

But de cette thèse

- Proposer des méthodes d'estimation pour apprendre $P(T)$ dans un cadre :
  - non-paramétrique (peu d'hypothèses sur P) ;
  - non-asymptotique (corpus limité)

- Proposer des applications en classification de textes

Plan

Plan

1. Formalisation

Plan

1. Formalisation
2. Arbres de contexte adaptatifs

Plan

1. Formalisation
2. Arbres de contexte adaptatifs
3. Double mélange et lien avec compression

Plan

1. Formalisation
2. Arbres de contexte adaptatifs
3. Double mélange et lien avec compression
4. Recodage du passé

Plan

1. Formalisation
2. Arbres de contexte adaptatifs
3. Double mélange et lien avec compression
4. Recodage du passé
5. Rééchantillonnage

Plan

1. Formalisation
2. Arbres de contexte adaptatifs
3. Double mélange et lien avec compression
4. Recodage du passé
5. Rééchantillonnage
6. Expériences

Plan

1. Formalisation
2. Arbres de contexte adaptatifs
3. Double mélange et lien avec compression
4. Recodage du passé
5. Rééchantillonnage
6. Expériences
7. Conclusion

Partie 1

Formalisation

Cadre statistique

– Soit X =le passé et Y =le prochain caractère

Cadre statistique

- Soit X =le passé et Y =le prochain caractère
- But = estimer $P(Y \mid X)$ à partir d'un ensemble d'entraînement $\mathcal{E}_N$ (typiquement N échantillons (X, Y) i.i.d. de loi P)

Risque d'un estimateur

Risque d'un estimateur

– Un estimateur $\hat{Q}$ utilise $\mathcal{E}_N$ pour construire $\hat{Q}(Y \mid X, \mathcal{E}_N)$ qui a une perte (entropie conditionnelle) :

$$D\left(P(Y \mid X) \,||\, \hat{Q}(Y \mid X, \mathcal{E}_N)\right) = E_{(X,Y)\sim P} \ln \frac{P(Y \mid X)}{\hat{Q}(Y \mid X, \mathcal{E}_N)}$$

Risque d'un estimateur

– Un estimateur $\hat{Q}$ utilise $\mathcal{E}_N$ pour construire $\hat{Q}(Y \mid X, \mathcal{E}_N)$ qui a une perte (entropie conditionnelle) :

$$D\left(P(Y \mid X) \,||\, \hat{Q}(Y \mid X, \mathcal{E}_N)\right) = E_{(X,Y)\sim P} \ln \frac{P(Y \mid X)}{\hat{Q}(Y \mid X, \mathcal{E}_N)}$$

– Le risque moyen de $\hat{Q}$ est :

$$\mathcal{R}(P, N, \hat{Q}) = E_{\mathcal{E}_N}\left[D\left(P(Y \mid X) \,||\, \hat{Q}(Y \mid X, \mathcal{E}_N)\right)\right]$$

Inégalité "oracle"

– Si $\hat{Q}$ est à valeur dans un ensemble $\mathcal{Q}$ on cherche une borne de la forme : $\forall P \in \mathcal{P}$,

$$\mathcal{R}(P, N, \hat{Q}) \leq \inf_{Q \in \mathcal{Q}} \mathcal{R}(P, N, Q) + b(N, \mathcal{P}, \mathcal{Q})$$

Inégalité "oracle"

– Si $\hat{Q}$ est à valeur dans un ensemble $\mathcal{Q}$ on cherche une borne de la forme : $\forall P \in \mathcal{P}$,

$$\mathcal{R}(P, N, \hat{Q}) \leq \inf_{Q \in \mathcal{Q}} \mathcal{R}(P, N, Q) + b(N, \mathcal{P}, \mathcal{Q})$$

– Si on a une famille de modèles $\{\mathcal{Q}_i, i \in I\}$ on cherche une inégalité "oracle" : $\forall P \in \mathcal{P}$,

$$\mathcal{R}(P, N, \hat{Q}) \leq \inf_{i \in I} \left\{ \inf_{Q \in \mathcal{Q}_i} \mathcal{R}(P, N, Q) + b(N, \mathcal{P}, \mathcal{Q}_i) \right\}$$

Partie 2

Arbres de contexte adaptatifs

Arbre de contexte

Arbre de contexte

– Définit une fonction de suffixe $\mathcal{S}(X)$

Arbre de contexte

- Définit une fonction de suffixe $\mathcal{S}(X)$
- Une distribution θ_s sur l'alphabet $\mathcal{A}$ est attachée à chaque nœud s :

$$Q_{\mathcal{S},\theta}(Y \mid X) = \theta_{\mathcal{S}(X)}(Y)$$

Estimation pour un arbre donné : Laplace

– $\mathcal{E}_N = (X_i, Y_i)_{i=1,\ldots,N}$

Estimation pour un arbre donné : Laplace

– $\mathcal{E}_N = (X_i, Y_i)_{i=1,\ldots,N}$
– L'estimateur de Laplace est :

$$\hat{Q}_S(Y \mid X, \mathcal{E}_N) = \frac{\#\{i : s(X_i) = s(X) \text{ et } Y_i = Y\} + 1}{\#\{i : s(X_i) = s(X)\} + |\mathcal{A}|}$$

Estimation pour un arbre donné : Laplace

– $\mathcal{E}_N = (X_i, Y_i)_{i=1,\dots,N}$
– L'estimateur de Laplace est :

$$\hat{Q}_{\mathcal{S}}(Y \mid X, \mathcal{E}_N) = \frac{\#\{i : s(X_i) = s(X) \text{ et } Y_i = Y\} + 1}{\#\{i : s(X_i) = s(X)\} + |\mathcal{A}|}$$

– Performance : $\forall P \in \mathcal{P}$,

$$\mathcal{R}(P, N, \hat{Q}_{\mathcal{S}}) \leq \min_{\theta} \mathcal{R}(P, N, Q_{\mathcal{S},\theta}) + \frac{|\mathcal{A}| - 1}{N} |\mathcal{S}|$$

Estimation pour un arbre donné : Laplace adaptatif (1)

– Soit $a(s) = \#\{y \in \mathcal{A} : \exists i, s(X_i) = s \text{ et } Y_i = y\}$

Estimation pour un arbre donné : Laplace adaptatif (1)

– Soit $a(s) = \#\{y \in \mathcal{A} : \exists i, s(X_i) = s \text{ et } Y_i = y\}$
– L'estimateur de Laplace adaptatif est :

$$\hat{Q}_S(Y \mid X, \mathcal{E}_N) = \frac{\#\{i : s(X_i) = s(X) \text{ et } Y_i = Y\} + \frac{a(s)}{|\mathcal{A}|}}{\#\{i : s(X_i) = s(X)\} + a(s)}$$

Estimation pour un arbre donné : Laplace adaptatif (2)

– Performance : $\forall P \in \mathcal{P}$,

$$\mathcal{R}(P, N, \hat{Q}_{\mathcal{S}}) \leq \min_{\theta} \mathcal{R}(P, N, Q_{\mathcal{S},\theta}) + \frac{\sum_{s \in \mathcal{S}} \left[a(s) - 1 + a(s) \left(1 - \frac{a(s)}{a} \right) \right]}{N}$$

– (au lieu de $\sum_{s \in \mathcal{S}} (a - 1)/N$)

Arbres de contexte adaptatifs (1)

– Pour chaque arbre $\mathcal{S}$ on calcule $\hat{Q}_{\mathcal{S}}(Y \mid X, \mathcal{E}_K)$ à partir d'un ensemble d'observations $\mathcal{E}_K$

Arbres de contexte adaptatifs (1)

- Pour chaque arbre $\mathcal{S}$ on calcule $\hat{Q}_{\mathcal{S}}(Y \mid X, \mathcal{E}_K)$ à partir d'un ensemble d'observations $\mathcal{E}_K$
- Une probabilité a priori π est choisie sur l'ensemble des arbres (de profondeur $\leq D$)
- Par exemple :

$$\pi(\mathcal{S}) = c^{|\mathcal{S}|}$$

Arbres de contexte adaptatifs (2)

– Un deuxième ensemble $\mathcal{E}_{N-K}$ sert à calculer une distribution de Gibbs a posteriori, d'exposant $\beta > 0$:

$$\rho_\beta(\mathcal{S}) = \frac{1}{Z}\pi(\mathcal{S}) \times \prod_{i=K+1}^{N} \hat{Q}_{\mathcal{S}}(Y_i \,|\, X_i, \mathcal{E}_K)^\beta$$

Arbres de contexte adaptatifs (2)

– Un deuxième ensemble $\mathcal{E}_{N-K}$ sert à calculer une distribution de Gibbs a posteriori, d'exposant $\beta > 0$:

$$\rho_\beta(\mathcal{S}) = \frac{1}{Z}\pi(\mathcal{S}) \times \prod_{i=K+1}^{N} \hat{Q}_{\mathcal{S}}(Y_i \,|\, X_i, \mathcal{E}_K)^\beta$$

– L'estimateur final est la moyenne selon ρ_β :

$$\hat{Q}(Y \,|\, X, \mathcal{E}_K, \mathcal{E}_{N-K}) = \sum_{\mathcal{S}} \rho_\beta(\mathcal{S}) \hat{Q}_{\mathcal{S}}(Y \,|\, X, \mathcal{E}_K)$$

Arbres de contexte adaptatifs (3)

Théorème 1. *Pour β "suffisamment" petit (de l'ordre de $\sqrt{2 \ln \ln N} / \ln N$) et un certain choix de K :*

$$\mathcal{R}(P, N, \hat{Q}) \leq \min_{\mathcal{S}} \left[\min_{\theta} \mathcal{R}(P, N, Q_{\mathcal{S},\theta}) + \frac{|\mathcal{S}| C_N}{N} \right]$$

avec

$$C_N = \left(\sqrt{\frac{1 + \log |\mathcal{A}|}{\beta}} + \sqrt{|\mathcal{A}| - 1} \right)^2 \left(1 + \frac{1}{N - 2} \right)$$

Arbres de contexte adaptatifs (4) : Implémentation

– On peut écrire l'arbre de contexte adaptatif sous la forme :

$$\hat{Q}(Y \mid X, \mathcal{E}_K, \mathcal{E}_{N-K}) = \frac{\sum_{\mathcal{S}} \prod_{s \in \mathcal{S}} w^{(1)}(s)}{\sum_{\mathcal{S}} \prod_{s \in \mathcal{S}} w^{(0)}(s)}$$

Arbres de contexte adaptatifs (4) : Implémentation

– On peut écrire l'arbre de contexte adaptatif sous la forme :

$$\hat{Q}(Y \mid X, \mathcal{E}_K, \mathcal{E}_{N-K}) = \frac{\sum_{\mathcal{S}} \prod_{s \in \mathcal{S}} w^{(1)}(s)}{\sum_{\mathcal{S}} \prod_{s \in \mathcal{S}} w^{(0)}(s)}$$

– Ces deux sommes se calculent récursivement selon la méthode du "Context Tree Weighting" de Willems, Shtarkov et Tjalkens, .

Arbres de contexte adaptatifs (4) : Améliorations

– Utilisation du mélange progressif au lieu du mélange de Gibbs

Arbres de contexte adaptatifs (4) : Améliorations

- Utilisation du mélange progressif au lieu du mélange de Gibbs
- Utilisation du Laplace adaptatif au lieu du Laplace

Arbres de contexte adaptatifs (4) : Améliorations

- Utilisation du mélange progressif au lieu du mélange de Gibbs
- Utilisation du Laplace adaptatif au lieu du Laplace
- π peut dépendre des données, ce qui permet de regrouper les arbres selon leur trace et d'améliorer l'inégalité oracle.

Partie 3

Double mélange et lien avec compression

Idée du Double Mélange

Au lieu de faire deux mélanges :

$$\begin{cases} \hat{Q}_{\mathcal{S}}(Y \mid X, \mathcal{E}_1) = \frac{1}{Z_1} \int Q_{\mathcal{S},\theta}(\mathcal{E}_1) Q_{\mathcal{S},\theta}(Y \mid X) \pi(d\theta) \\ \end{cases}$$

Idée du Double Mélange

Au lieu de faire deux mélanges :

$$\begin{cases} \hat{Q}_{\mathcal{S}}(Y \mid X, \mathcal{E}_1) = \frac{1}{Z_1} \int Q_{\mathcal{S},\theta}(\mathcal{E}_1) Q_{\mathcal{S},\theta}(Y \mid X) \pi(d\theta) \\ \hat{Q}(Y \mid X, \mathcal{E}_1, \mathcal{E}_2) = \frac{1}{Z_2} \sum_{\mathcal{S}} \pi(\mathcal{S}) \hat{Q}_{\mathcal{S}}(\mathcal{E}_2)^{\beta} \hat{Q}_{\mathcal{S}}(Y \mid X) \end{cases}$$

Idée du Double Mélange

Au lieu de faire deux mélanges :

$$\begin{cases} \hat{Q}_{\mathcal{S}}(Y \mid X, \mathcal{E}_1) = \frac{1}{Z_1} \int Q_{\mathcal{S},\theta}(\mathcal{E}_1) Q_{\mathcal{S},\theta}(Y \mid X) \pi(d\theta) \\ \hat{Q}(Y \mid X, \mathcal{E}_1, \mathcal{E}_2) = \frac{1}{Z_2} \sum_{\mathcal{S}} \pi(\mathcal{S}) \hat{Q}_{\mathcal{S}}(\mathcal{E}_2)^{\beta} \hat{Q}_{\mathcal{S}}(Y \mid X) \end{cases}$$

Peut-on faire un “double” mélange :

$$\hat{Q}^{(d)}(Y \mid X, \mathcal{E}) = \frac{1}{Z} \sum_{\mathcal{S}} \pi(\mathcal{S}) \int Q_{\mathcal{S},\theta}(\mathcal{E})^{\beta} Q_{\mathcal{S},\theta}(Y \mid X) \pi(d\theta)$$

Lien avec la compression universelle

$$\mathcal{R}^{(cum)}(P, N, \hat{Q}) = \sum_{i=1}^{N} \mathcal{R}(P, i, \hat{Q})$$

Lien avec la compression universelle

$$\mathcal{R}^{(cum)}(P, N, \hat{Q}) = \sum_{i=1}^{N} \mathcal{R}(P, i, \hat{Q})$$

– Inégalité "oracle" pour le double mélange (ex : CTW) :

$$\mathcal{R}^{(cum)}(P, N, \hat{Q}^{(d)}) \leq \min_{\mathcal{S}} \Big[\min_{\theta} \mathcal{R}^{(cum)}(P, N, Q_{\mathcal{S},\theta}) + \frac{(|\mathcal{A}| - 1)|\mathcal{S}|}{2} \ln N + \ln \frac{1}{\pi(\mathcal{S})} + C \Big]$$

Double mélange “à la Gibbs”

Performance de Gibbs (Catoni, 1998) : si β est “suffisamment petit” alors :

$$\mathcal{R}(P,N,\hat{Q}^{(d)}) \leq \inf_{\mathcal{S},\theta}\left[\mathcal{R}(P,N,Q_{\mathcal{S},\theta}) + \frac{\gamma(\mathcal{S},\theta)}{\beta(N+1)}\right]$$

avec

$$\gamma(\mathcal{S}_0,\theta_0) = E_{\mathcal{E}_{N+1}} \frac{E_{\pi(S,d\theta)} Q_{\mathcal{S},\theta}(\mathcal{E}_{N+1})^\beta \ln Q_{\mathcal{S},\theta}(\mathcal{E}_{N+1})^\beta}{E_{\pi(S,d\theta)} Q_{\mathcal{S},\theta}(\mathcal{E}_{N+1})^\beta \ln Q_{\mathcal{S}_0,\theta_0}(\mathcal{E}_{N+1})^\beta}$$

Calcul de la borne $\gamma(\mathcal{S}, \theta)$ (1)

Par le calcul (technique...) on montre que si $\pi(d\theta)$ est une distribution de Dirichlet de paramètre $1/2$:

$$\gamma(\mathcal{S}, \theta) \leq \frac{(|\mathcal{A}| - 1)|\mathcal{S}|}{2} + E_{\mathcal{E}_{N+1}} \ln \frac{\sum_{\mathcal{S}'} \tilde{\pi}(\mathcal{S}')}{\tilde{\pi}(\mathcal{S})}$$

Calcul de la borne $\gamma(\mathcal{S}, \theta)$ (1)

Par le calcul (technique...) on montre que si $\pi(d\theta)$ est une distribution de Dirichlet de paramètre $1/2$:

$$\gamma(\mathcal{S}, \theta) \leq \frac{(|\mathcal{A}| - 1)|\mathcal{S}|}{2} + E_{\mathcal{E}_{N+1}} \ln \frac{\sum_{\mathcal{S}'} \tilde{\pi}(\mathcal{S}')}{\tilde{\pi}(\mathcal{S})}$$

avec :

$$\tilde{\pi}(\mathcal{S}) \approx \pi(\mathcal{S}) \times \exp\left(\frac{(|\mathcal{A}| - 1)|\mathcal{S}|}{2}\right) \times \prod_{s \in \mathcal{S}} \left(cte \times n(s)^{-\frac{|\mathcal{A}|-1}{2}}\right)$$

Calcul de la borne $\gamma(\mathcal{S}, \theta)$ (1)

Si $\pi(\mathcal{S})$ est une distribution a priori il faut donc faire un double mélange "à la Gibbs" avec la distribution modifiée :

$$\bar{\pi}(\mathcal{S}) \sim \pi(\mathcal{S}) \times \prod_{s \in \mathcal{S}} \frac{\pi^{|\mathcal{A}|/2}}{\Gamma(|\mathcal{A}|/2)} \left(\frac{\beta n(s)}{2\pi e} \right)^{\frac{|\mathcal{A}|-1}{2}}$$

pour obtenir...

Performance du double mélange

Théorème 2. *Pour β "suffisamment" petit (de l'ordre de $\sqrt{2\ln\ln N}/(8\ln N)$), le double mélange selon $\bar{\pi}$ satisfait :*

$$\mathcal{R}(P,N,\hat{Q}) \leq \min_{\mathcal{S}}\left[\min_{\theta}\mathcal{R}(P,N,Q_{\mathcal{S},\theta}) + \frac{|\mathcal{S}|C_N}{\beta N}\right]$$

avec

$$C_N = 1 + \log|\mathcal{A}| + \frac{|\mathcal{A}|-1}{2} + O(N^{-1})$$

Bilan du double mélange

– Par rapport à la compression, il faut favoriser les probabilités a priori des modèles par un facteur en $N^{dim/2}$ pour prendre en compte les différences dans la vitesse d'estimation des paramètres

Bilan du double mélange

- Par rapport à la compression, il faut favoriser les probabilités a priori des modèles par un facteur en $N^{dim/2}$ pour prendre en compte les différences dans la vitesse d'estimation des paramètres
- Il est asymptotiquement équivalent de remplacer le mélange $\pi(d\theta|\mathcal{S})$ par le maximum de vraisemblance pénalisé par un facteur $\exp(-|\mathcal{S}|((|\mathcal{A}|-1)/2)$

Bilan du double mélange

- Par rapport à la compression, il faut favoriser les probabilités a priori des modèles par un facteur en $N^{dim/2}$ pour prendre en compte les différences dans la vitesse d'estimation des paramètres
- Il est asymptotiquement équivalent de remplacer le mélange $\pi(d\theta|\mathcal{S})$ par le maximum de vraisemblance pénalisé par un facteur $\exp(-|\mathcal{S}|((|\mathcal{A}|-1)/2)$
- $\pi(d\theta\,|\,\mathcal{S})$ est naturellement une Dirichlet de paramètre $1/2$ $(Jeffrey's\ prior)$

Partie 4

Recodage du passé

Idée du recodage (1)

Idée du recodage (2)

– L'entropie de l'anglais est de l'ordre de 1 bit / caractère

Idée du recodage (2)

- L'entropie de l'anglais est de l'ordre de 1 bit / caractère
- Au lieu de faire des arbres contextes sur les chaînes de caractères on peut les faire sur une représentation plus "efficace"

Idée du recodage (2)

– L'entropie de l'anglais est de l'ordre de 1 bit / caractère

– Au lieu de faire des arbres contextes sur les chaînes de caractères on peut les faire sur une représentation plus "efficace"

– Un code binaire de longueur D pour le passé est une application

$$\sigma : \mathcal{A}^{-\mathbb{N}} \to \{0,1\}^D$$

Choix du code σ

– Le but est d'avoir beaucoup d'information dans D bits, qui est similaire au problème du $variable\text{-}to\text{-}fixed\ length\ block\ coding$.

Choix du code σ

- Le but est d'avoir beaucoup d'information dans D bits, qui est similaire au problème du $variable\text{-}to\text{-}fixed\ length\ block\ coding$.
- Si la distribution de $T_{-\infty}^{-1} = (\ldots, T_{-2}, T_{-1})$ est connue alors il est possible de créer un code qui contienne en D bits l'information contenue en moyenne dans $D/H(T_{-\infty}^{-1})$ caractères

Code de Tunstall (distribution connue)

Recodage itératif pour processus stationnaire (1)

– On part d'un code $\sigma(T_{-\infty}^{-1})$ du passé

Recodage itératif pour processus stationnaire (1)

- On part d'un code $\sigma(T_{-\infty}^{-1})$ du passé
- On utilise des observations i.i.d. de $(\ldots, T_{-2}, T_{-1}, T_0)$ pour estimer $P(T_0 \,|\, T_{-i}^{-1})$ $(i \geq 0)$ par l'arbre de contexte adaptatif qui utilise le code

Recodage itératif pour processus stationnaire (1)

- On part d'un code $\sigma(T_{-\infty}^{-1})$ du passé
- On utilise des observations i.i.d. de $(\ldots, T_{-2}, T_{-1}, T_0)$ pour estimer $P(T_0 \,|\, T_{-i}^{-1})$ $(i \geq 0)$ par l'arbre de contexte adaptatif qui utilise le code
- On en déduit une estimation de la densité du futur par :

$$\hat{Q}(T_1^K) = \hat{Q}(T_1) \times \hat{Q}(T_2 \,|\, T_1) \times \ldots \times \hat{Q}(T_K \,|\, T_1^{K-1})$$

Recodage itératif pour processus stationnaire (1)

- On part d'un code $\sigma(T_{-\infty}^{-1})$ du passé
- On utilise des observations i.i.d. de $(\ldots, T_{-2}, T_{-1}, T_0)$ pour estimer $P(T_0 \,|\, T_{-i}^{-1})$ $(i \geq 0)$ par l'arbre de contexte adaptatif qui utilise le code
- On en déduit une estimation de la densité du futur par :

$$\hat{Q}(T_1^K) = \hat{Q}(T_1) \times \hat{Q}(T_2 \,|\, T_1) \times \ldots \times \hat{Q}(T_K \,|\, T_1^{K-1})$$

– on en déduit un code (de Tunstall) du futur : $\sigma(T_1^\infty)$

Recodage itératif pour processus stationnaire (2)

– On part d'un code $\sigma(T_1^\infty)$ du futur

Recodage itératif pour processus stationnaire (2)

- On part d'un code $\sigma(T_1^\infty)$ du futur
- On utilise des observations i.i.d. de $(T_0, T_1, T_2, \ldots)$ pour estimer $P(T_0 \,|\, T_1^i)$ $(i \geq 0)$ par l'arbre de contexte adaptatif qui utilise le code

Recodage itératif pour processus stationnaire (2)

- On part d'un code $\sigma(T_1^\infty)$ du futur
- On utilise des observations i.i.d. de $(T_0, T_1, T_2, \ldots)$ pour estimer $P(T_0 \,|\, T_1^i)$ $(i \geq 0)$ par l'arbre de contexte adaptatif qui utilise le code
- On en déduit une estimation de la densité du passé par :

$$\hat{Q}(T_{-K}^{-1}) = \hat{Q}(T_{-1}) \times \hat{Q}(T_{-2} \,|\, T_{-1}) \times \ldots \times \hat{Q}(T_{-K} \,|\, T_{-K+1}^{-1})$$

Recodage itératif pour processus stationnaire (2)

- On part d'un code $\sigma(T_1^\infty)$ du futur
- On utilise des observations i.i.d. de $(T_0, T_1, T_2, \ldots)$ pour estimer $P(T_0 \,|\, T_1^i)$ $(i \geq 0)$ par l'arbre de contexte adaptatif qui utilise le code
- On en déduit une estimation de la densité du passé par :

$$\hat{Q}(T_{-K}^{-1}) = \hat{Q}(T_{-1}) \times \hat{Q}(T_{-2} \,|\, T_{-1}) \times \ldots \times \hat{Q}(T_{-K} \,|\, T_{-K+1}^{-1})$$

– on en déduit un code (de Tunstall) du passé : $\sigma(T_{-\infty}^{-1})$

Performance du recodage (1)

– L'efficacité d'une itération se mesure en redondance moyenne du code (pour $k > 0$)

$$R_k(P, N, \hat{Q}) = E_{P(dT_1^k)} \ln \frac{P(T_1^k)}{\hat{Q}(T_1^k)}$$

Performance du recodage (1)

– L'efficacité d'une itération se mesure en redondance moyenne du code (pour $k > 0$)

$$R_k(P, N, \hat{Q}) = E_{P(dT_1^k)} \ln \frac{P(T_1^k)}{\hat{Q}(T_1^k)}$$

– Elle se décompose en $R_k = \sum_{i=1}^{k} r_i$ avec :

$$r_i(P, N, \hat{Q}) = E_{P(dT_1^i)} \ln \frac{P(T_i \,|\, T_1^{i-1})}{\hat{Q}(T_i \,|\, T_1^{i-1})}$$

Performance du recodage (2)

Théorème 3. *Pour β "suffisamment" petit (de l'ordre de $\sqrt{2 \ln \ln N / \ln N}$) :*

$$ER_n(P, \hat{Q}) \leq \sum_{i=1}^{n} \left\{ \min_{\mathcal{S}} \left[\min_{\theta} r_i(P, \hat{Q}_{\mathcal{S} \circ \sigma_i}) + \frac{C_N |\mathcal{S}|}{N} \right] \right\}$$

avec

$$C_N = \left(\sqrt{2\beta^{-1}} + \sqrt{|\mathcal{A}| - 1} \right)^2 \left(1 + \frac{1}{N-2} \right)$$

Partie 5

Rééchantillonnage

Idée du rééchantillonnage

– Dans tout ce qui précède N observations i.i.d. du processus $T_{-\infty}^{1}$ sont supposées disponibles

Idée du rééchantillonnage

- Dans tout ce qui précède N observations i.i.d. du processus $T_{-\infty}^{1}$ sont supposées disponibles
- Dans un cadre plus réaliste on dispose d'une seule (longue) réalisation du processus à partir duquel on effectue un rééchantillonnage

Idée du rééchantillonnage

- Dans tout ce qui précède N observations i.i.d. du processus $T_{-\infty}^{1}$ sont supposées disponibles
- Dans un cadre plus réaliste on dispose d'une seule (longue) réalisation du processus à partir duquel on effectue un rééchantillonnage
- Est-il possible de prouver les inégalités "oracle" en prenant en compte le rééchantillonnage ?

Rééchantillonnage selon la loi empirique(1)

– Supposons que le vrai processus suive la loi d'un certain arbre de contexte $P = P_{\mathcal{S}_0,\theta_0}$ ergodique, où $\mathcal{S}_0$ a une profondeur $\leq D$

Rééchantillonnage selon la loi empirique(1)

– Supposons que le vrai processus suive la loi d'un certain arbre de contexte $P = P_{\mathcal{S}_0,\theta_0}$ ergodique, où $\mathcal{S}_0$ a une profondeur $\leq D$
– Soit T_{1-D}^{L} une (longue) observation du processus

Rééchantillonnage selon la loi empirique(1)

– Supposons que le vrai processus suive la loi d'un certain arbre de contexte $P = P_{\mathcal{S}_0,\theta_0}$ ergodique, où $\mathcal{S}_0$ a une profondeur $\leq D$
– Soit T_{1-D}^{L} une (longue) observation du processus
– La distribution empirique d'une chaine z de longueur $D+1$ est :

$$\hat{P}_L(z) = \frac{1}{L}\#\{1 \leq i \leq L : T_{i-D}Nn = z\}$$

Rééchantillonnage selon la loi empirique(1)

– Soit $\hat{\mathcal{E}}_N$ un ensemble de N variables (X, Y) tirées selon la loi empirique $\hat{P}_L$

Rééchantillonnage selon la loi empirique(1)

- Soit $\hat{\mathcal{E}}_N$ un ensemble de N variables (X, Y) tirées selon la loi empirique $\hat{P}_L$
- Soit $\hat{Q}(Y \mid X, \hat{\mathcal{E}}_N)$ estimé avec un arbre de contexte adaptatif

Performance du rééchantillonnage

Théorème 4. *Le risque de l'estimateur* $\hat{Q}(Y \mid X, \hat{\mathcal{E}}_N)$ *est borné par :*

$$E_{P(dT_{1-D}^L)}\mathcal{R}(P, N, \hat{Q}) \leq \left[1 + O\left(\sqrt{\frac{\ln L}{L}}\right)\right] \times$$

$$\min_{\mathcal{S}} \left[\min_{\theta} \mathcal{R}(P, N, Q_{\mathcal{S},\theta}) + \frac{|\mathcal{S}|C_N}{N}\right] + O\left(\sqrt{\frac{\ln L}{L}}\right)$$

Idée de la preuve

– Utilise une inégalité de déviation de la forme :

$$P\left(\sup_{s\in\mathcal{S}_o}\left|\frac{\hat{P}_L(s)}{P(s)}-1\right|>\epsilon\right)\leq 2D|\mathcal{S}|e^{-CN\epsilon^2}$$

pour une constante C.

Idée de la preuve

– Utilise une inégalité de déviation de la forme :

$$P\left(\sup_{s\in\mathcal{S}_o}\left|\frac{\hat{P}_L(s)}{P(s)}-1\right|>\epsilon\right)\leq 2D|\mathcal{S}|e^{-CN\epsilon^2}$$

pour une constante C.

– On obtient une borne pour le risque moyen en fonction de ϵ qui peut ensuite être optimisé

Partie 6

Expériences

Comparaison avec des modèles de Markov d'ordre fixe

- Utilise le texte $Far\ from\ the\ madding\ crowd$ de T. Hardy (du $Calgary\ corpus$)
- Tire un ensemble d'entraînement pour calculer les estimateurs
- Calcule la log-vraisemblance sur un ensemble test i.i.d. (de taille 5000)

Classification de textes non supervisée (1)

Soient :

– T_1 et T_2 deux textes

– $\mathcal{E}_1$, $\mathcal{E}'_1$, $\mathcal{E}_2$, $\mathcal{E}'_2$ tirés i.i.d. de T_1 et T_2

Classification de textes non supervisée (1)

Soient :

– T_1 et T_2 deux textes
– $\mathcal{E}_1$, $\mathcal{E}'_1$, $\mathcal{E}_2$, $\mathcal{E}'_2$ tirés i.i.d. de T_1 et T_2

Une pseudo-distance entre T_1 et T_2 est :

$$d(T_1, T_2) = \ln \frac{\hat{Q}(\mathcal{E}'_1 \,|\, \mathcal{E}_1)}{\hat{Q}(\mathcal{E}'_2 \,|\, \mathcal{E}_1)} + \ln \frac{\hat{Q}(\mathcal{E}'_2 \,|\, \mathcal{E}_2)}{\hat{Q}(\mathcal{E}'_1 \,|\, \mathcal{E}_2)}$$

Classification de textes non supervisée (2)

Text Number	Extracted from
1-5	Wintson Churchill (*The Crossing*)
6-10	Joseph Conrad (*The Arrow of gold*)
11-15	Arthur Conan Doyle (*The hound of the Baskervilles*)
16-20	Karl Marx (*Manifesto of the communist party*)
21-25	Baruch Spinoza (*Political treatise*)
26-30	Jonathan Swift (*Gulliver's travel*)
31-35	Francois Marie Arouet Voltaire (*Candide*)
36-40	Virginia Woolf (*Night and day*)

Base de données de textes

Distance entre le texte 23 (Spinoza) et les autres

Similarités entre textes (distance seuillée à 1.03)

Classification de textes supervisée (1)

– But $=$ classer un nouveau texte T dans une catégorie parmi k

Classification de textes supervisée (1)

- But $=$ classer un nouveau texte T dans une catégorie parmi k
- Soit $\mathcal{C}$ une catégorie et $\hat{Q}_{\mathcal{C}}$ un arbre de contexte adaptatif entraîné dessus

Classification de textes supervisée (1)

- But $=$ classer un nouveau texte T dans une catégorie parmi k
- Soit $\mathcal{C}$ une catégorie et $\hat{Q}_{\mathcal{C}}$ un arbre de contexte adaptatif entraîné dessus
- Le score de $\mathcal{C}$ sur T est :

$$s_T(\mathcal{C}) = \frac{1}{|T|} \log \hat{Q}_{\mathcal{C}}(T)$$
$$= -h(\hat{P}_T) - \mathcal{D}(\hat{P}_T \,||\, \hat{Q}_{\mathcal{C}})$$

Classification de textes supervisée (2)

– Si $\mathcal{C}_1$ et $\mathcal{C}_2$ sont deux catégories concurrentes :

$$s_T(\mathcal{C}_1) - s_T(\mathcal{C}_2) = \mathcal{D}(\hat{P}_T \,||\, \hat{Q}_{\mathcal{C}_2}) - \mathcal{D}(\hat{P}_T \,||\, \hat{Q}_{\mathcal{C}_1})$$

Classification de textes supervisée (2)

– Si $\mathcal{C}_1$ et $\mathcal{C}_2$ sont deux catégories concurrentes :

$$s_T(\mathcal{C}_1) - s_T(\mathcal{C}_2) = \mathcal{D}(\hat{P}_T \,||\, \hat{Q}_{\mathcal{C}_2}) - \mathcal{D}(\hat{P}_T \,||\, \hat{Q}_{\mathcal{C}_1})$$

– On choisit la catégorie avec le score le plus élevé

Expérience : Reuters-21578 database

Categorie	B-E point
earn	93
acq	91
money-fx	71
grain	74
crude	79
trade	56
interest	63
ship	75
wheat	58
corn	41

Expérience : 20 Newsgroup Database

– 20 Newsgroups x 1000 messages

Expérience : 20 Newsgroup Database

– 20 Newsgroups x 1000 messages
– Il faut choisir une catégorie parmi 20

Expérience : 20 Newsgroup Database

- 20 Newsgroups x 1000 messages
- Il faut choisir une catégorie parmi 20
- Textes peu formatés et ambigus

Expérience : 20 Newsgroup Database

- 20 Newsgroups x 1000 messages
- Il faut choisir une catégorie parmi 20
- Textes peu formatés et ambigus
- **Précision = 90,0 % de bonne classification**

Expérience : Génération automatique de texte

`talk.politics.mideast` : associattements in the greeks who be neven exclub no bribedom of spread marinary s troo-perties savi tack acter i ruthh jake bony
`soc.religion.christian` : that must as a friend one jerome unimovingt ail serving are national atan cwru evid which done joseph in response of the wholeleaseriend

Partie 7

Conclusion

Ce qui est fait...

– Présentation d'estimateurs "de mélange" (idées de O. Catoni) aboutissant à des résultats théoriques (inégalités oracle) et pratiques ("presque" implémentables)

Ce qui est fait...

- Présentation d'estimateurs "de mélange" (idées de O. Catoni) aboutissant à des résultats théoriques (inégalités oracle) et pratiques ("presque" implémentables)
- Méthode de recodage et de rééchantillonnage

Ce qui est fait...

- Présentation d'estimateurs "de mélange" (idées de O. Catoni) aboutissant à des résultats théoriques (inégalités oracle) et pratiques ("presque" implémentables)
- Méthode de recodage et de rééchantillonnage
- Résultats expérimentaux prometteurs

...et ce qui est à faire

– Lien avec sélection de modèles

...et ce qui est à faire

- Lien avec sélection de modèles
- Inégalités en déviation et pas seulement en moyenne

...et ce qui est à faire

– Lien avec sélection de modèles
– Inégalités en déviation et pas seulement en moyenne
– Autres critères (risque L_1, régression, classification...)

...et ce qui est à faire

- Lien avec sélection de modèles
- Inégalités en déviation et pas seulement en moyenne
- Autres critères (risque L_1, régression, classification...)
- Optimiser la classification ($Kernel\ methods$...), la segmentation automatique...

...et ce qui est à faire

- Lien avec sélection de modèles
- Inégalités en déviation et pas seulement en moyenne
- Autres critères (risque L_1, régression, classification...)
- Optimiser la classification ($Kernel\ methods$...), la segmentation automatique...
- Domaines d'applications : langage, économétrie/finance, biologie ($post\text{-}génomique,\ protéomique$...)

...et ce qui est à faire

- Lien avec sélection de modèles
- Inégalités en déviation et pas seulement en moyenne
- Autres critères (risque L_1, régression, classification...)
- Optimiser la classification ($Kernel\ methods$...), la segmentation automatique...
- Domaines d'applications : langage, économétrie/finance, biologie ($post\text{-}génomique,\ protéomique$...)

MERCI !