Statistical Sequence Modelling

Jean-Philippe Vert

`Jean-Philippe.Vert@mines.org`

`http://www.dma.ens.fr/users/vert`

May 21, 2001

Outline

Outline

1. Adaptive Context Trees : Theory

Outline

1. Adaptive Context Trees : Theory

2. Adaptive Context Trees : Applications

Outline

1. Adaptive Context Trees : Theory

2. Adaptive Context Trees : Applications

3. Combining ACT with HMM?

Outline

1. Adaptive Context Trees : Theory
2. Adaptive Context Trees : Applications
3. Combining ACT with HMM?
4. From sequence modelling to classification...

Part 1

Adaptive Context Trees : Theory

Motivations

Motivations

- A sequence (DNA, protein...) is a complex object

Motivations

- A sequence (DNA, protein...) is a complex object
- A probabilistic model is a probability distribution over the set of sequences

Motivations

- A sequence (DNA, protein...) is a complex object
- A probabilistic model is a probability distribution over the set of sequences
- It can be used to compare two sequences or a sequence and a model

Classical models : Markov

Classical models : Markov

- A model is characterized by the conditional distribution $P(Y \mid X)$ where X is the past, Y is the next character.

Classical models : Markov

- A model is characterized by the conditional distribution $P(Y \mid X)$ where X is the past, Y is the next character.

- Markov models : $P(Y \mid X)$ only depends on the last D letters of X

Classical models : Markov

- A model is characterized by the conditional distribution $P(Y \mid X)$ where X is the past, Y is the next character.

- Markov models : $P(Y \mid X)$ only depends on the last D letters of X

- The number of parameters is exponential with D!

Context tree model

Context tree model

- Maps any past sequence X into its longest suffix $\mathcal{S}(X)$

Context tree model

- Maps any past sequence X into its longest suffix $\mathcal{S}(X)$
- A probability distribution θ_s is attached to each node $s \in \mathcal{S}$:

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

The estimation issue

- Let $P(X \mid Y)$ an unknown probability distribution

The estimation issue

- Let $P(X \mid Y)$ an unknown probability distribution
- We observe an i.i.d. sample

$$\mathcal{E} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$$

The estimation issue

- Let $P(X \mid Y)$ an unknown probability distribution
- We observe an i.i.d. sample

$$\mathcal{E} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$$

- How to guess a good tree S and a good set of parameters θ from $\mathcal{E}$?

Adaptive Context tree model : the recipe

1. Split $\mathcal{E}$ into $\mathcal{E}_1$ and $\mathcal{E}_2$

Adaptive Context tree model : the recipe

1. Split $\mathcal{E}$ into $\mathcal{E}_1$ and $\mathcal{E}_2$

2. Use $\mathcal{E}_1$ to estimate the parameters of every model S by:

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

Adaptive Context tree model : the recipe

1. Split $\mathcal{E}$ into $\mathcal{E}_1$ and $\mathcal{E}_2$

2. Use $\mathcal{E}_1$ to estimate the parameters of every model S by:

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

3. Chose a probability *a priori* on the set of trees $\pi(\mathcal{S})$

4. Use $\mathcal{E}_2$ to build a posterior distribution on the trees:

$$\rho(\mathcal{S}) = \frac{1}{Z}\pi(\mathcal{S}) \times \prod_{i \in \mathcal{E}_2} \hat{P}_{\mathcal{S}}(Y_i \mid X_i)^{\beta}$$

4. Use $\mathcal{E}_2$ to build a posterior distribution on the trees:

$$\rho(\mathcal{S}) = \frac{1}{Z}\pi(\mathcal{S}) \times \prod_{i \in \mathcal{E}_2} \hat{P}_{\mathcal{S}}(Y_i \mid X_i)^{\beta}$$

5. The ACT estimator is finally:

$$\hat{P}(Y \mid X) = \sum_{\mathcal{S}} \rho(\mathcal{S}) \hat{P}_{\mathcal{S}}(Y \mid X)$$

Is $\hat{P}(Y \mid X)$ a “good” estimation?

Is $\hat{P}(Y \mid X)$ a "good" estimation?

- The closeness of $P(Y|X)$ and $Q(Y|X)$ can be measured in conditional relative entropy

$$D(P||Q) = \sum_{(x,y)} P(x,y) \log \frac{Q(y|x)}{P(y|x)}$$

- For any unknown P the average loss of $\hat{P}$ satisfies

$$E[D(P||\hat{P})] \leq \inf_{\mathcal{S},\theta} \left\{ D(P||P_{\mathcal{S},\theta}) + C\frac{|\mathcal{S}|}{N} \right\}$$

- For any unknown P the average loss of $\hat{P}$ satisfies

$$E[D(P||\hat{P})] \leq \inf_{\mathcal{S},\theta}\left\{D(P||P_{\mathcal{S},\theta}) + C\frac{|\mathcal{S}|}{N}\right\}$$

- C is an optimal constant

Remarks

- No assumption is made on the unknown P

Remarks

- No assumption is made on the unknown P
- The implementation is efficient using a recursive algorithm (Context Tree Weighting method)

Remarks

- No assumption is made on the unknown P

- The implementation is efficient using a recursive algorithm (Context Tree Weighting method)

- The resulting distribution $\hat{P}(Y|X)$ is a mixture of all $P_{\mathcal{S},\theta}$ and not a particular one

Part 2

Adaptive Context Trees : Applications

Unsupervised Text clustering (1)

- Let:
  - ⋆ T_1 and T_2 two given texts (i.e. long strings)

Unsupervised Text clustering (1)

- Let:
  - ⋆ T_1 and T_2 two given texts (i.e. long strings)
  - ⋆ $\mathcal{E}_1$, $\mathcal{E}_1'$, $\mathcal{E}_2$, $\mathcal{E}_2'$ sampled i.i.d. from T_1 et T_2

Unsupervised Text clustering (1)

- Let:
  - ⋆ T_1 and T_2 two given texts (i.e. long strings)
  - ⋆ $\mathcal{E}_1$, $\mathcal{E}_1'$, $\mathcal{E}_2$, $\mathcal{E}_2'$ sampled i.i.d. from T_1 et T_2

- A pseudo-distance between T_1 et T_2 is:

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

Unsupervised Text clustering (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*)

Text database

Distance between text n.23 (Spinoza) and other texts

Text clustering (1.03 threshold)

Automatic text generation

`talk.politics.mideast`:

associattements in the greeks who be neven exclub no bribedom of spread marinary s trooperties 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

Biological sequences?

- The same method can be applied to cluster or classify proteins, DNA etc...

Biological sequences?

- The same method can be applied to cluster or classify proteins, DNA etc...

- Approach already tested with good results for protein family prediction by Bejerano/Rona (RECOMB 1999) and Eskin/Grundy/Singer (RECOMB 2000)

Part 3

ACT with HMM?

Definition of a CT-HMM (1)

- $\mathcal{H}$ a finite set of hidden states

Definition of a CT-HMM (1)

- $\mathcal{H}$ a finite set of hidden states

- $(\mathcal{S}, \theta) = \{(\mathcal{S}, \theta)_h\,,\, h \in \mathcal{H})\}$ a family of context tree models for each hidden state

Definition of a CT-HMM (1)

- $\mathcal{H}$ a finite set of hidden states
- $(\mathcal{S},\theta) = \{(\mathcal{S},\theta)_h\,,\, h \in \mathcal{H})\}$ a family of context tree models for each hidden state
- $\mu(h_2 \,|\, h_1)$ a transition probability

Definition of a CT-HMM (2)

- The CT-HMM distribution is:

$$\begin{aligned} &P_{\mathcal{S},\theta,\mu}(H_{n+1}, X_{n+1} \,|\, H_{-\infty}^{n}, X_{-\infty}^{n}) \\ &\qquad = \mu(H_{n+1} \,|\, H_n) \times P_{(\mathcal{S},\theta)_{H_i}}(X_i \,|\, X_{-\infty}^{n}) \end{aligned}$$

Definition of a CT-HMM (2)

- The CT-HMM distribution is:

$$P_{\mathcal{S},\theta,\mu}(H_{n+1}, X_{n+1} \,|\, H^n_{-\infty}, X^n_{-\infty}) \\ = \mu(H_{n+1} \,|\, H_n) \times P_{(\mathcal{S},\theta)_{H_i}}(X_i \,|\, X^n_{-\infty})$$

- It generalizes HMMs

How to guess the hidden state sequence?

- Let $x = (\ldots, x_0, \ldots, x_N)$ a observed sequence, generated by an unknown model supposed to be well approached by a CT-HMM.

How to guess the hidden state sequence?

- Let $x = (\ldots, x_0, \ldots, x_N)$ a observed sequence, generated by an unknown model supposed to be well approached by a CT-HMM.

- The classical approach (E-M algorithm) does not work

How to guess the hidden state sequence?

- Let $x = (\ldots, x_0, \ldots, x_N)$ a observed sequence, generated by an unknown model supposed to be well approached by a CT-HMM.
- The classical approach (E-M algorithm) does not work
- How to guess a good $h = (h_1, \ldots, h_N)$?

How to guess the hidden state sequence?

- Let $x = (\ldots, x_0, \ldots, x_N)$ a observed sequence, generated by an unknown model supposed to be well approached by a CT-HMM.
- The classical approach (E-M algorithm) does not work
- How to guess a good $h = (h_1, \ldots, h_N)$?
- A good sequence can be seen as a one which reduces the complexity of the observed sequence

A mixture approach

- Let $\pi(S, d\theta, d\mu)$ a probability *a priori* on the models and parameters

A mixture approach

- Let $\pi(S, d\theta, d\mu)$ a probability *a priori* on the models and parameters

- The mixture probability sums up the information contained in all models:

$$P_w(X) = \sum_{H,\mathcal{S}} \int_{\mu,\theta} P_{\mu,\mathcal{S},\theta}(X, H)\pi(d\mu, \mathcal{S}, d\theta)$$

Selection by Minimum Description Length

- $\log_2 P_w(h)$ bits to describe the hidden sequence

Selection by Minimum Description Length

- $\log_2 P_w(h)$ bits to describe the hidden sequence
- $\log_2 P_w(x \mid h)$ to describe the observation x given a hidden sequence h

Selection by Minimum Description Length

- $\log_2 P_w(h)$ bits to describe the hidden sequence
- $\log_2 P_w(x \mid h)$ to describe the observation x given a hidden sequence h
- MDL : Choose $h = \arg\max_h P_w(h) \times P_w(x|h)$

Application : E Coli genome segmentation

Part 3

From modelling to classification

Classification ?

- $X \in \mathcal{X}$ an objet

Classification ?

- $X \in \mathcal{X}$ an objet
- $Y \in \mathcal{Y}$ a class (typically $\{1, \ldots, k\}$)

Classification ?

- $X \in \mathcal{X}$ an objet
- $Y \in \mathcal{Y}$ a class (typically $\{1, \ldots, k\}$)
- Classifier $=$ mapping $f : \mathcal{X} \to \mathcal{Y}$

Classification and bioinformatics

- protein classification,

Classification and bioinformatics

- protein classification, structure prediction ...

Classification and bioinformatics

- protein classification, structure prediction ...
- (promoter-sequence based) gene classification...

Classification and bioinformatics

- protein classification, structure prediction ...
- (promoter-sequence based) gene classification...
- functional classification of enzymes, binding pairs...

Classical approach in bioinformatics

- build probabilistic models

Classical approach in bioinformatics

- build probabilistic models
- derive a score function

Classical approach in bioinformatics

- build probabilistic models
- derive a score function
- classify according to maximum score

Research proposal

- Efficient learning algorithms include:

Research proposal

- Efficient learning algorithms include:

⋆ Support Vector Machine ,

Research proposal

- Efficient learning algorithms include:
  - Support Vector Machine ,
  - Boosting ,

Research proposal

- Efficient learning algorithms include:
  - ⋆ Support Vector Machine ,
  - ⋆ Boosting ,
  - ⋆ Randomized classifiers ...

Research proposal

- Efficient learning algorithms include:
  - ⋆ Support Vector Machine ,
  - ⋆ Boosting ,
  - ⋆ Randomized classifiers ...
- First papers give impressive results

Research proposal

- Efficient learning algorithms include:
  - ⋆ Support Vector Machine ,
  - ⋆ Boosting ,
  - ⋆ Randomized classifiers ...
- First papers give impressive results
- What about a seminar?