Text Categorization Using Adaptive Context Trees

Jean-Philippe Vert

Ecole Normale Supérieure
Department of Mathematics and Applications

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

February 22, 2001

Plan

- Bag-of-words representation
- Statistical Language Models
- Adaptive context trees
- Experimental results

Introduction

- Text categorization

Introduction

- Text categorization
- Many solutions (Bayesian classifiers, k-nearest neighbors, boosting, SVM...)

Introduction

- Text categorization
- Many solutions (Bayesian classifiers, k-nearest neighbors, boosting, SVM...)
- All based on the *bag-of-words* representation for texts

Introduction

- Text categorization
- Many solutions (Bayesian classifiers, k-nearest neighbors, boosting, SVM...)
- All based on the *bag-of-words* representation for texts
- We propose an alternative representation based on *statistical language modelling*

Bag-of-words representation

- A text T is mapped to a vector $\overrightarrow{v}_T$ of $\mathbb{R}^{|\mathcal{A}|}$

Bag-of-words representation

- A text T is mapped to a vector $\overrightarrow{v}_T$ of $\mathbb{R}^{|\mathcal{A}|}$
- Trade-off when chosing $\mathcal{A}$:

Size of $\mathcal{A}$	Semantic content	Stability
Large (e.g. 20,000 words)	High	Low
Small (e.g. 26 letters)	Low	High

Bag-of-words representation

- A text T is mapped to a vector $\overrightarrow{v}_T$ of $\mathbb{R}^{|\mathcal{A}|}$
- Trade-off when chosing $\mathcal{A}$:

Size of $\mathcal{A}$	Semantic content	Stability
Large (e.g. 20,000 words)	High	Low
Small (e.g. 26 letters)	Low	High

- Control on $|\mathcal{A}|$: word stemming, thesaurus, stop words removal, feature selection...

Statistical Language Modelling (SLM)

- Model language $=$ stochastic process $P(X_1 \ldots, X_n)$ on $\mathcal{A}$ which “mimics” the language generating process

Statistical Language Modelling (SLM)

- Model language = stochastic process $P(X_1 \ldots, X_n)$ on $\mathcal{A}$ which "mimics" the language generating process
- Bayes decision framework (speech recognition, OCR, machine translation...):

$$W^* = \arg\max_W P(W \mid I)$$
$$= \arg\max_W P(W) P(I \mid W)$$

- Bayes decision framework (bis) for document classification or information retrieval:

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

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

- Bayes decision framework (bis) for document classification or information retrieval:

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

- Text modelling: we need **local models**

Building local models

- $P_T(X_1 \,|\, X_{-\infty}^0)$ is very rich even if $|\mathcal{A}|$ is small

Building local models

- $P_T(X_1 \,|\, X_{-\infty}^0)$ is very rich even if $|\mathcal{A}|$ is small
- Constraints:
  - ⋆ No assumption on the "true" P (non-parametric)

Building local models

- $P_T(X_1 \,|\, X_{-\infty}^0)$ is very rich even if $|\mathcal{A}|$ is small
- Constraints:
  - ⋆ No assumption on the "true" P (non-parametric)
  - ⋆ Small number of observations (non-asymptotic)

Building local models

- $P_T(X_1 \,|\, X_{-\infty}^0)$ is very rich even if $|\mathcal{A}|$ is small
- Constraints:
  - ⋆ No assumption on the "true" P (non-parametric)
  - ⋆ Small number of observations (non-asymptotic)
- Trade-off between *complexity* of a model and *precision* of the estimation

Mathematical Formulation

- If P is a process distribution the conditional relative entropy of $Q(X_1 \| X_{-\infty}^0)$ is:

$$\mathcal{D}(P \,||\, Q) =$$

$$\sum_{x_{-\infty}^0 \in \mathcal{A}^\infty} P(x_{-\infty}^0) \sum_{x_1 \in \mathcal{A}} P(x_1 \,|x_{-\infty}^0) \log \frac{P(x_1 \,|x_{-\infty}^0)}{Q(x_1 \,|x_{-\infty}^0)}$$

Mathematical Formulation

- If P is a process distribution the conditional relative entropy of $Q(X_1 \| X_{-\infty}^0)$ is:

$$\mathcal{D}(P \,||\, Q) = \sum_{x_{-\infty}^0 \in \mathcal{A}^\infty} P(x_{-\infty}^0) \sum_{x_1 \in \mathcal{A}} P(x_1 \,|x_{-\infty}^0) \log \frac{P(x_1 \,|x_{-\infty}^0)}{Q(x_1 \,|x_{-\infty}^0)}$$

- An observation is $Z = (X_{-\infty}^0, X_1)$

- An *estimator* $\hat{P}$ maps a series of observations $Z_1^N = (Z_1, \ldots, Z_N)$ into a conditional distribution:

$$\hat{P}_{Z_1^N}(X_1 \mid X_{-\infty}^0)$$

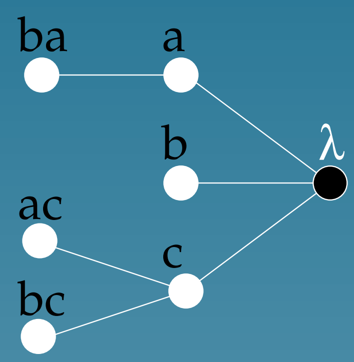
- An *estimator* $\hat{P}$ maps a series of observations $Z_1^N = (Z_1, \dots, Z_N)$ into a conditional distribution:

$$\hat{P}_{Z_1^N}(X_1 \,|\, X_{-\infty}^0)$$

- The average risk of $\hat{P}$ to estimate P is:

$$R(\hat{P}) = E_{Z_1^N \sim P} \mathcal{D}(P \,||\, \hat{P}_{Z_1^N})$$

Context tree model



- Variable-length Markov models
- A distribution θ_s on $\mathcal{A}$ is attached to each node s:

$$P_{\mathcal{S},\theta}(X_1 \mid X_{-\infty}^0) = \theta_{s(X_{-\infty}^0)}(X_1)$$

Adaptive context tree model

- A test set Z_1^K is used to estimate the continuous parameters of all models $\mathcal{S}$

Adaptive context tree model

- A test set Z_1^K is used to estimate the continuous parameters of all models $\mathcal{S}$
- A validation set Z_{K+1}^N is used to build a posterior Gibbs distribution $\rho(d\mathcal{S})$ on the set of models

Adaptive context tree model

- A test set Z_1^K is used to estimate the continuous parameters of all models $\mathcal{S}$
- A validation set Z_{K+1}^N is used to build a posterior Gibbs distribution $\rho(d\mathcal{S})$ on the set of models
- The resulting estimator $\hat{P}$ is:

$$\hat{P}(X_1 \,|\, X_{-\infty}^0) = \sum_{\mathcal{S}} \rho(S) \hat{P}_{\mathcal{S}}(X_1 \,|\, X_{-\infty}^0)$$

Performance

Theorem 1. *The adaptive context tree estimator* $\hat{P}$ *satisfies:*

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

with

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

Application: Text representation

- T a text to represent

Application: Text representation

- T a text to represent
- Sample an i.i.d. set Z_1^N from T by repreatedly choosing a random position in the text

Application: Text representation

- T a text to represent
- Sample an i.i.d. set Z_1^N from T by repreatedly choosing a random position in the text
- Use Z_1^N to estimate $\hat{P}_T$

Application: Scoring a category

- Let $\mathcal{C}$ a category and $\hat{P}_{\mathcal{C}}$ its representation

Application: Scoring a category

- Let $\mathcal{C}$ a category and $\hat{P}_{\mathcal{C}}$ its representation
- The score of $\mathcal{C}$ w.r.t. a text T is:

$$\begin{aligned} s_T(\mathcal{C}) &= \log P_{\mathcal{C}}(T) \\ &= -h(P_T) - \mathcal{D}(P_T \,||\, \hat{P}_{\mathcal{C}}) \end{aligned}$$

Application: Text categorization

- For two categories $\mathcal{C}_1$ and $\mathcal{C}_2$:

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

Application: Text categorization

- For two categories $\mathcal{C}_1$ and $\mathcal{C}_2$:

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

- Chose the category with highest score (naive)

Experiments: `Reuters-21578` database

Category	B-E point
earn	93
acq	91
money-fx	71
grain	74
crude	79
trade	56
interest	63
ship	75

Experiment: 20 Newsgroup Database

- Maps any new text into one out of 20 categories
- Accuracy $=$ 90,0 %

Experiment: 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

Conclusion

- Representation of a complex object with statistical methods

Conclusion

- Representation of a complex object with statistical methods
- Encouraging results in spite of obvious drawbacks (what about language?)

Conclusion

- Representation of a complex object with statistical methods
- Encouraging results in spite of obvious drawbacks (what about language?)
- General method which can be applied to any natural language (e.g. Japanese) or other complex strings (e.g. biological sequences)

Conclusion

- Representation of a complex object with statistical methods
- Encouraging results in spite of obvious drawbacks (what about language?)
- General method which can be applied to any natural language (e.g. Japanese) or other complex strings (e.g. biological sequences)