SAND2022-12109C

Sandia National Laboratories

Exceptional service in the national interest

Compression Analytics

Presenter: Eduardo Cardenas-Torres

Sandia Mentor: Alexander Foss

MARTIANS Symposium

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

What is Compression Analytics?

Data Compression: The process of encoding information using fewer bit than the original representation

- **What's the Intuition behind this?**
  - Reduce the number of unique symbols in your data (smallest possible "alphabet")
  - Encode more frequent symbols with fewer bits (fewest bits for most common "letters")

Challenges in Data Compression

- Best choice of compression algorithm depends on data type
- Transforming data first before compressing (depending on method)
- Many different compression algorithms to choose from

Lossy and **Lossless Compression:**

- Lossless Compression is a type of class of data compression that allows data to be reconstructed from compressed data without any loss of information.
- Lossy Compression does not allow data to be perfectly reconstructed

Using Compression Algorithms for Machine Learning

How can compression be utilized for machine learning?

- Compression algorithms involve modeling the data and then encoding it
- Modeling often includes an algorithm for predicting symbol probabilities
- The encoding stage uses these probabilities to substitute high-probability symbols with smaller codes
- Compression Analytics uses predicted symbol probabilities for machine learning
- **Today's focus:** Discriminant Analysis for classification of data by using those trained probabilities from compression algorithms

Background: Discriminant Analysis

- The idea is to model the distribution of X in each of the classes separately , P(X=x | G = k), and then use Bayes theorem to flip things around and obtain P(G=k | X=x).

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}.$$

How does compression come into play?

- Typical compression algorithms follow the follow algorithm:
- **P(σ|s), σ represents the predicting symbol and the s represents the string given**
- Can see where compression algorithms may tie in with Discriminant Analysis and help utilize statistical properties

Prediction by Partial Matching (PPM)

Original Paper: Data Compression Using Adaptive Coding and Partial String Matching by John G. Cleary and Ian H. Witten (1984)

- PPM is a type of lossless compression
- Different Versions of PPM for handling unobserved characters in a string

Basic Idea: PPM uses the last few characters (“context”) in the input stream to predict the upcoming one

- Example: Suppose we have a text document about cars (includes the word car frequently)
- Predict the 3rd letter, given preceding observed characters
- **P(σ|”c”)**
- **P(σ|”ca”)**
- **P(σ|”car”)**
- PPM algorithms have specialized systems for combining different context lengths

How does PPM work?

How does the Algorithm work? (Generally for all PPM)

$$\hat{P}(\sigma|s_{n-D+1}^{n}) = \begin{cases} \hat{P}_D(\sigma|s_{n-D+1}^{n}), & \text{if } s_{n-D+1}^{n}\sigma \text{ apeared in the training sequence;} \\ \hat{P}_D(escape|s_{n-D+1}^{n}) \cdot \hat{P}(\sigma|s_{n-D+2}^{n}), & \text{otherwise.} \end{cases}$$

- s represents the sequence prior to **σ**, the predicted symbol
- s_a^b denotes the sequence from index a to b
- **n** indexes position in the sequence
- P(**σ** |"ab") = P(**σ** $|s_1^2$)
- D represents the **order** of a model
- What does the order exactly mean? The number of symbols read prior to the predicted symbol
- e.g. P(σ|"ab") is a second order model because we are given two characters "ab"

PPM* Examples

Order $k = 2$					Order $k = 1$					Order $k = 0$				Order $k = -1$			
Predictions			c	p	Predictions			c	p	Predictions			c	p	Predictions		c	p
ab	→	r	2	$\frac{2}{3}$	a	→	b	2	$\frac{2}{7}$	→	a	5	$\frac{5}{16}$	→	A	1	$\frac{1}{\lvert A\rvert}$
	→	*Esc*	1	$\frac{1}{3}$		→	c	1	$\frac{1}{7}$	→	b	2	$\frac{2}{16}$				
						→	d	1	$\frac{1}{7}$	→	c	1	$\frac{1}{16}$				
ac	→	a	1	$\frac{1}{2}$		→	*Esc*	3	$\frac{3}{7}$	→	d	1	$\frac{1}{16}$				
	→	*Esc*	1	$\frac{1}{2}$						→	r	2	$\frac{2}{16}$				
					b	→	r	2	$\frac{2}{3}$	→	*Esc*	5	$\frac{5}{16}$				
ad	→	a	1	$\frac{1}{2}$		→	*Esc*	1	$\frac{1}{3}$								
	→	*Esc*	1	$\frac{1}{2}$													
					c	→	a	1	$\frac{1}{2}$								
br	→	a	2	$\frac{2}{3}$		→	*Esc*	1	$\frac{1}{2}$								
	→	*Esc*	1	$\frac{1}{3}$													
					d	→	a	1	$\frac{1}{2}$								
ca	→	d	1	$\frac{1}{2}$		→	*Esc*	1	$\frac{1}{2}$								
	→	*Esc*	1	$\frac{1}{2}$													
					r	→	a	2	$\frac{1}{3}$								
da	→	b	1	$\frac{1}{2}$		→	*Esc*	1	$\frac{1}{3}$								
	→	*Esc*	1	$\frac{1}{2}$													
ra	→	c	1	$\frac{1}{2}$													
	→	*Esc*	1	$\frac{1}{2}$													

Table 1: PPM model after processing the string *abracadabra* (maximum order 2)

PPM* and Discriminant Analysis

Discriminant Analysis classifies an observation x into class c such that

$$c = argmax_k P(Y = k)P(X = x|Y = k)$$

- **P(Y=k):** prior probability of drawing from class k
- **P(X=x|Y=k):** Conditional probability of drawing an observation x from given class k

- How can PPM* be utilized in this?

$$\hat{P} = (\sigma|s_{n-D+1}^{n})$$

- Where the probabilities of the symbols can replace the conditional distribution **P(X=x|Y=k)**

Combining PPM* and Discriminant Analysis

General Discriminant Analysis

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{\sum_{l=1}^{K} P(X = x|Y = l)P(Y = l)}$$

PPM* Algorithm

$$\hat{P}(\sigma|s_{n-D+1}^{n}) = \begin{cases} \hat{P}_D(\sigma|s_{n-D+1}^{n}), & \text{if } s_{n-D+1}^{n}\sigma \text{ apeared in the training sequence;} \\ \hat{P}_D(escape|s_{n-D+1}^{n}) \cdot \hat{P}(\sigma|s_{n-D+2}^{n}), & \text{otherwise.} \end{cases}$$

PPM* Discriminant Analysis

$$\hat{P}(Y = k|X = x) = \frac{\hat{P}(X = x|s_{n-D+1}^{n}, Y = k)\hat{P}(Y = k)}{\sum_{l=1}^{K} \hat{P}(X = x|s_{n-D+1}^{n}, Y = l)\hat{P}(Y = l)}$$

$$\hat{P}(Y = k) = \frac{\#\{items\ in\ class\ k\}}{\#\{Total\ items\}}$$

Variable Order Markov Model for PPM* Discriminant Analysis

Variable Order Markov Model is a type of Markov Model that gives predicted probabilities for the current symbol based on a given context

- VOMMs have multiple conditional distributions for a range of context lengths
- Standard Markov Models require a single choice of context length
- VOMMs can flexibly use different context lengths for each prediction

$$P(X = x | S_{n-D+1}^{n})$$

Variable Order Markov Model for PPM* Discriminant Analysis

Variable Order Markov Model can help finding the conditional distribution P(X=x|Y=k) and similar to PPM* compression

- By using PPM* strength of escaping as a transition probability from one model to a lower one

Example

$$P(X = x|s_{n-D+1}^{n}) \sim Multinomial(\vec{p}_D)$$

$$\vec{p}_D \sim Dirichlet$$

$$P(X = x|s_{n-D+1}^{n})\ Multinomial(\vec{p}_D)$$

$$\vec{p}_D\ Dirichlet(\alpha_{D-1})$$

References

- Proceedings DCC ‘95 Data Compression Conference. (1995) Unbounded Length Contexts for PPM. Institute of Electrical and Electronics Engineers
- Journal of Artificial Intelligence Research 22. (2004) On Prediction Using Variable Order Markov Models. Journal of Artificial Intelligence Research. R. Begleiter, R. El-Yaniv, G. Yona
- Proceedings of Machine Learning Research. (2010) Bayesian Variable Order Markov Models. Christos Dimitrakakis

Thank you for Listening!