Data Compression with Probabilistic Inference

Pranav Maneriker, Dhruv Singal , Ankesh Pandey
IIT Kanpur

Objectives

The primary objectives of our project are to:

- Gain an overview of various probabilistic machine learning approaches in lossless data compression
- Understand Prediction by Partial Matching and its derivatives
- Read about data compression using nonparametric Bayesian inference in Sequence Memoizer
- Explore and experiment with the PAQ series of archivers

Lossless compression

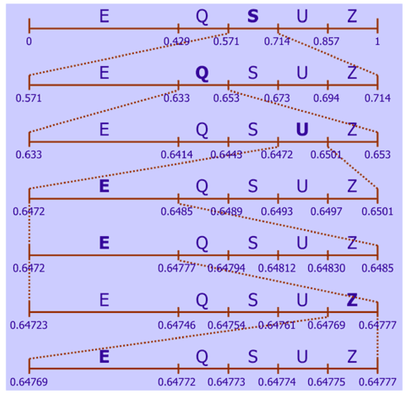

Figure 1: Arithmetic coding [1]

1. Arithmetic coding uses probability distribution over symbols to encode efficiently Decoding is the reverse operation.
2. Adaptive arithmetic coding involves updating probability distribution after observing each symbol in the sequence
3. The model updates must be deterministic so that the cumulative probabilities during both compression and decompression are consistent

Prediction by Partial Matching (PPM)

This algorithm, originally proposed by Cleary and Witten[2], sequentially compresses one symbol at a time while learning context dependent probability for improving compression.

1. A Markov model with context depth dependent on computational restrictions (higher order requires more storage)
2. For order o, the current prediction is a function of the previous o predictions
3. The model uses the longest matching context cleverly (with a fixed set of hyper parameters) to generate probabilities

PPM-A

Let $C = \sum_{\phi \in \Sigma} c(\phi)$ be the times a context i has been seen.
Probability of a symbol ϕ occurring in the same context:

$$e(\phi) = p(\phi) = \frac{c(\phi)}{1+C}, c(\phi) > 0$$

If q characters have occurred and $|\Sigma| = a$, for each of the $a - q$ characters not seen yet,

$$p(\phi) = \frac{1}{1+C} \times \frac{1}{a-q}, c(\phi) = 0$$

Advancements in PPM

Primarily, different PPM versions use modified update rules and models to define the probability of the next symbol and how the escape mechanism is handled.
Of the developments in PPM, the PPM-DP[4] algorithm is one of the most recent ones. The major ideas in this algorithm are:

1. Usage of a generalised blending mechanism with distinct hyper parameters based on depth and fanout to combine contexts
2. Dynamic updates to hyperparameters using gradient information

Sequence Memoizer

The Sequence Memoizer by Wood et al.[6] is a deep (unbounded) smoothing Markov model. It takes the non-parametric Bayesian approach to compress sequences of data. The main features are:

1. It encodes the fact that natural language sequence data exhibits power-law properties using the Pitman-Yor process as a prior
2. By extending the hierarchical PYP model it includes the set of distributions in all contexts of finite length
3. Due to the coagulation property of the PYP inference in full sequence model with infinite number of parameters is equivalent to inference in compact context tree with linear number of parameters

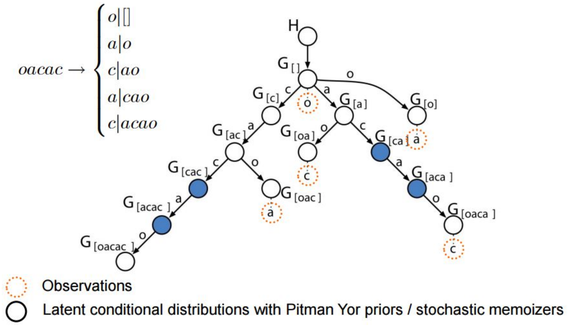

Figure 2: Graphical model Trie[5]

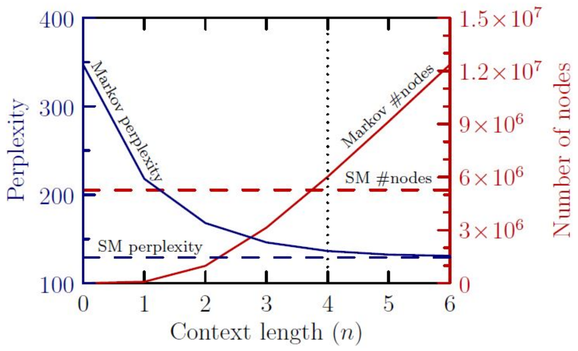

Figure 3: Performance of SM vs. n^{th} order Markov models[6]

PAQ

State of the art in data compression

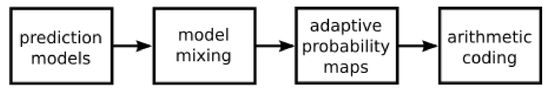

Figure 4: PAQ8 architecture[3]

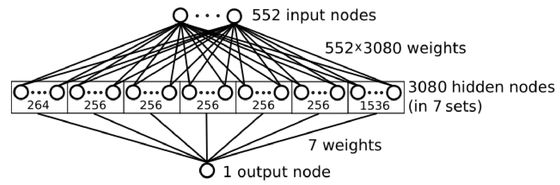

Figure 5: PAQ8 model mixer[3]

References

[1] Context adaptive binary arithmetic coding.

[2] J. G. Cleary and I. H. Witten.
Data compression using adaptive coding and partial string matching.
Communications, IEEE Transactions on, 32(4):396–402, 1984.

[3] B. Knoll.
A Machine Learning Perspective on Predictive Coding with PAQ8 and New Applications.
PhD thesis, The University of British Columbia (Vancouver, 2011.

[4] C. Steinruecken, Z. Ghahramani, and D. MacKay.
Improving ppm with dynamic parameter updates.
In *Proc. Data Compression Conference*, 2015.

[5] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W. Teh.
A stochastic memoizer for sequence data, 2009.

[6] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W. Teh.
The sequence memoizer.
Communications of the ACM, 54(2):91–98, 2011.

Acknowledgements

We are thankful to Professor Piyush Rai for providing us with a great learning environment in the course Probabilistic Machine Learning.