Knowledge as a Constraint on Uncertainty for Unsupervised Classification: A Study in Part-of-Speech Tagging

Thomas J. Murray IV, Panayiotis G. Georgiou, and Shrikanth S. Narayanan
Speech Analysis and Interpretation Laboratory
University of Southern California, Los Angeles, CA USA
Focus of the study

• Setting: No labeled data, but knowledge source that may limit or bias classifier choices

• As knowledge increases, how well does reduction in entropy of label distribution predict performance? How precise does the knowledge need to be?

• Evaluate other effects of constrained parameter space: stability, convergence, label assignment
• For a given input \( x \), prior knowledge constrains choice of label \( y \); if we view as distribution \( p(y \mid x) \), \( H(Y \mid X) \) is one measure of task difficulty

• No labeled data required; but note we compare different types of knowledge, fixed model type

• Fano’s equality (Fano, 1961): Conditional entropy part of lower bound on \( p(\text{Error}) \)
Summary

- Label entropy is a reasonable indicator of performance with different knowledge sets; and no labeled data is required.

- Even simple constraints can have big benefits for accuracy, training stability, and convergence.

- Uncertainty is of course not the only factor in accuracy. For future work: more complete predictive measures.