Abstract

The need to operate on sequence data is prevalent across a range of real world applications including protein/DNA classification, speech recognition, intrusion detection, and text classification. Sequence data can be distinguished from the more-typical vector representation in that the length of sequences within a dataset can vary and that the order of symbols within a sequence carries meaning. Although it has become increasingly easy to collect large amounts of sequence data, our ability to infer useful information from these sequences has not kept pace. For instance, in the domain of biological sequences, experimentally determining the order of amino acids in a protein is far easier than determining the protein's physical structure or its role within a living organism. This asymmetry holds over a number of sequence data domains, and, as a result, researchers increasingly rely on computational techniques to infer properties of sequences that are either difficult or costly to collect through direct measurement.

This work explores a number of latent variable models over sequence data. These models were designed to produce alternate representations of sequences that distill relevant information, making them both easier to process with traditional machine-learning tools and potentially improving on benchmarks over standard inference tasks such as classification and motif finding.

In this presentation, I will discuss two latent variable models that incorporate structure from the Profile Hidden Markov Model (HMM), a model commonly used to represent biological sequences. These methods both simplify and enrich the mechanisms by which standard Profile HMMs operate.

The first model relaxes the discrete Profile HMM hidden state space to a continuous one. Placing a regularizer that encourages sparsity on this new continuous space produces a new model that shares many characteristics with a set of successful techniques known as Sparse Dictionary Learning. This relaxation is the basis of our Relevant Subsequence Sparse Dictionary Learning (RS-DL) model. Applied to continuous sequences, RS-DL is effective at extracting human-recognizable motifs. In addition, subsequences extracted using RS-DL can improve on classification performance over standard nearest neighbor and dynamic time warping techniques.

The next model I discuss involves incorporating Profile HMM structure into a family of purely discriminative models. These models, which we call Subsequence Networks, are similar to convolutional neural networks, which have garnered state-of-the-art results in a number of tasks in computer vision. Subsequence Networks compare favorably to state-of-the-art sequence Kernel approaches on protein sequence classification problems while using a significantly different mode of operation.