An Entropic Estimator for Structure Discovery in Hidden Markov Models

Presented by Derek Anderson
• Discovery and Segmentation of Activities in Video

• Other Brand references (where the real details are!)
  - An entropic estimator for structure discovery
    - TR-98-19 August 1998
  - Pattern Discovery via Entropy Minimization
    - Artificial Intelligence and Statistics, 1999
  - Structure Discovery in Conditional Probability Models via an Entropic Prior and Parameter Extinction
    - Neural Computation, Vol. 11, no. 5, pp. 1,155-1.182, 1999
SOME of the Problems with “Classical HMMs”

• Models tend to not be *interpretable*
• Model parameters
  - How many *mixtures, number of states*, and *transition probs*
• What constitutes a *good* model?
  - Individually the most likely w.r.t. observation data
  - Discriminative
  - Concise and compact representation
  - Interpretable model
  - All of the above?
• Number of training samples (not unique to HMM!)
  - *Too few* observations to estimate many of the parameters
  - HMMs can consist of A LOT of parameters and relatively few samples to estimate their values from
Some Approaches to Learning and Structure Discovery

• Heuristic generate-and-test search
• Evolutionary computing
• Discriminative training methods
• Clustering approaches
• I have seen some references about, but have not looked into, the following
  ◦ Minimum message length (MML) and minimum description length (MDL)
The Big Picture!

- Desire the simultaneous learning of model structure and parameters
- Brand advocates replacing the M-step in the Baum-Welch with estimators that minimize entropy
  - Maximize the information content of the model parameters
  - A desire for the smallest, least ambiguous, most specific model compatible with the data
Problems Surrounding the EM

- Expectation Maximization (EM) used for the ML solution (talking about Baum–Welch)
  - Brand says not valid for small data sets
  - Most of the parameters are only supported by small subsets of the data
  - Approach is riddled with local optima
  - Use entropy to discover the regularities and hidden structure while simultaneously addressing accidental properties (noise and sampling artifacts)
Entropic Estimator for Structure Discovery in HMMs

• Brand includes an entropic prior and provides a solution for the maximum a posteriori (MAP) estimator

• What is a MAP estimator?
  • Similar to ML estimator, but MAP employs an augmented optimization objective which incorporates a prior distribution over the quantity one wants to estimate (numerator in Bayes formulae)

\[
\theta^* = \arg \max_{\theta} \left[ P(\theta|X) \propto P(X|\theta) e^{-H(\theta)} \right]
\]

EQ.1

\[
\exp \left[ \sum_{i}^{N} \theta_i \log \theta_i \right] = e^{-H(\theta)}
\]
Relation of EQ.1 to EQ.2

- MAP estimate minimizes the entropy of
  - (a) Data’s expected sufficient statistics
    - Sufficiency is the property possessed by a statistic, with respect to a parameter, "when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter" (really the “minimum” definition)
    - In the EM, the E-step reduces to computing expected sufficient statistics for the parameters (i.e. expected transition and emission counts for an HMM)
    - In the M-step, parameters are updated using normalized sufficient statistics
  - (b) Relative entropy between model and data
  - (c) Measure of the model itself

$$\theta^* = \arg \min_{\theta} \left[ H(\omega) + D(\omega || \theta) + H(\theta) \right]$$  \hspace{1cm} \text{EQ.2}

Typically \( w \) represents the "true" distribution of data, observations, or a precise calculated theoretical distribution. The measure \( \theta \) typically represents a theory, a model, a description or an approximation of \( w \).
The Calculation

\[ \theta^* = \arg \min_{\theta} \left[ H(\omega) + D(\omega \| \theta) + H(\theta) \right] \]

(a) \[ -\sum_i \omega_i \log \omega_i + \sum_i \omega_i \log \frac{\omega_i}{\theta_i} - \sum_i \theta_i \log \theta_i \]
MAP Estimator for Gaussian

- State with a single Gaussian (no state mixture model!)
  - Entropy favors minimum volume covariance's
  - He assumes a zero mean
  - Estimator is
    \[
    \hat{K} = \frac{\sum_{i=1}^{N} x_i x_i^T}{N + Z}
    \]
  - \( Z \) is a negative temperature term
  - \( Z \) varies the strength of the prior under the control of a temperature variable \( T = 1 - Z \)

- \[ \Sigma_{\ell}^{\text{new}} = \frac{\sum_{i=1}^{N} P(\ell|x_i, \Theta^g)(x_i - \mu_{\ell}^{\text{new}})(x_i - \mu_{\ell}^{\text{new}})^T}{\sum_{i=1}^{N} P(\ell|x_i, \Theta^g)} \]
  - for MM and HMM

- \[ \Sigma_{i\ell} = \frac{\sum_{t=1}^{T} \gamma_{i\ell}(t)(o_t - \mu_{i\ell})(o_t - \mu_{i\ell})^T}{\sum_{t=1}^{T} \gamma_{i\ell}(t)} \]

- Drive to \( Z = 1 \) over training
- Gives deterministic annealing within EM

\[
P(\theta|X, T, T_0) \overset{\text{def}}{=} P(X|\theta)P(\theta)^{T_0-T}\delta(T)
\]

\[ \hat{\theta} = \arg\max_{\theta} [\log P(X|\theta) - ZH(\theta)] \]
MAP Estimator for the Transition Probabilities

• Entropic prior favors near-deterministic odds
  ♦ Around chance (0.5), no information
  ♦ At the extremes {0, 1}, very informative

• Estimator is (derivation in the Pattern Discovery Via Entropy Minimization paper)

\[
\hat{\lambda} = \frac{1}{N} \sum_{i}^{N} \frac{\omega_i}{\theta_i} + Z \log \theta_i + Z,
\]

\[
\hat{\theta}_i = \frac{-\omega_i/Z}{W(-\omega_i e^{1-\lambda/Z}/Z)},
\]

\[W \text{ is the Lambert inverse function satisfying } W(x)e^{W(x)} = x\]
Symbol Recognition

a. conventional  
b. entropic

initialization  
final model  
conventionally trained
Office Activity

- Ellipse fitting the single largest connected set of active pixels in the image
- Uses two consecutive frames
- The observation vector consisted of
  - Mean x
  - Mean y
  - Change in x
  - Change in y
  - Mass
  - Change in mass
  - Elongation
  - Eccentricity
Learned Model
Conclusions

• Interesting approach to learning structure and parameters simultaneously
• Still has the same limitations as most HMMs
  ◆ Most likely classifier
    ▪ Pick 1 of K!
    ▪ What about “I do not know”?
• Brand also extend this work to
  ◆ Multi-observation-mixture+counter
  ◆ Good for situations in which one wants to simultaneously monitor multiple processes within a single hidden variable structure