Diversity-Promoting Bayesian Learning of Latent Variable Models

Pengtao Xie¹, Jun Zhu¹,², Eric Xing¹
¹Machine Learning Department, Carnegie Mellon University
²Department of Computer Science and Technology, Tsinghua University

Motivation

Latent Variable Models (LVMs)

Patterns and Components

How to capture joint distribution
Facilitate efficient PA

Experiments on mixture of experts model
Basic idea

How to prevent overfitting
Naturally extends to Bayesian
Develop a mutual angular prior
Mutual angular Bayesian network prior

How to reduce model size
Promote angle
Encourage components to be diverse
Maximize angle

Decompose a component vector \( a_i \) into its magnitude \( g = \| a_i \|_2 \) and direction \( \hat{a}_i = a_i / g \)

Bayesian network

Local probability based on von Mises-Fisher distribution
\[ p(\hat{a}_i|p(a_j)) = C_F(x) \exp(-\sum_{i=1}^2 \hat{a}_i \hat{a}_j) \]

Prior over magnitude \( G = \{ g \} \)

Generative process of component vectors
- Sample \( \hat{a}_i \sim \text{MFD}(\alpha_i, \omega) \)
- For \( i = 2, \ldots, K \), sample \( \hat{a}_i \sim \text{MFD}(\sum_{j \neq i} \hat{a}_j, \omega) \)
- For \( i = 1, \ldots, K \), sample \( g_i \sim \text{Gamma}(\alpha_1, \alpha_2) \)
- For \( i = 1, \ldots, K \), \( \alpha_i = \omega \)

Joint distribution
\[ p(A) = C_F(\omega) \prod_{i=1}^K C_F(g_i) \exp(-\sum_{i=1}^2 \hat{a}_i \hat{a}_j) \]

Variational Inference
- Maximize a variational lower bound
  \[ \sup_{q(A)} \mathbb{E}_q[p(D|A)p(A)] - \mathbb{E}_q[\log q(A)] \]
- \( 1 / \sum_{i=1}^2 \| a_i \|_1 \) in \( p(\hat{a}_i|p(a_j)) \) is not amenable for variational inference
Reparametrize \( p(\hat{a}_i|p(a_j)) \)

- Derive upper bound for the log of partition function

Study Case: Mixture of Experts (ME)

Methods

Bayesian Learning of LVMs
- Component vectors are random variables
- Infer a posterior distribution over the components
- Advantages over point estimation
  -- Alleviates overfitting via model averaging
  -- Quantify uncertainty

Diversification LVMs in Bayesian Learning
- Basic idea: define diversity-biased priors and “propagate” the diversity to posterior via Bayes rule

Mutual Angular Prior
- Angle-based notion of diversity: component vectors are deemed to be more diverse if their pairwise angles are larger
- Facilitate efficient posterior inference
- Naturally extends to Bayesian nonparametrics

Diversification Improves Accuracy

Diversification Reduces Model Size without Sacrificing Modeling Power

Results

Diversification Effectively Captures Infrequent Patterns

State of the Art Performance

Conclusions

- Promote diversity in Bayesian latent variable models
  -- Capture infrequent patterns
  -- Reduce model size without sacrificing modeling power
  -- Prevent overfitting

- Mutual angular Bayesian network prior
  -- Facilitate variational inference
  -- Naturally extend to Bayesian nonparametrics
  -- Constructed based on Bayesian network and von Mises-Fisher distribution

- Variational inference
  -- Reparametrize local probability distribution
  -- Derive upper bound for the log of partition function

- Experiments on mixture of experts model demonstrate the efficacy of our method