Learning Bayes Networks

- Fully observed vs. hidden variables
- Known structure vs. unknown structure
- Maximum likelihood vs. Bayesian

Fully observed Case

- Classic example: Naive Bayes algorithm.
- Likelihood decomposes and we get separate estimate for parameters of every node.
- For discrete variables we get maximum likelihood for multinomial variables which is given by frequency counts.
- Bayesian solution: if we use a conjugate prior (Dirichlet for multinomial) then posterior in same family. Posterior and MAP given by adjusted frequency counts. Calculating posterior is simply inference in the Bayes network.

Unobserved Variables

- Classic example: Mixture of Gaussians.
  \[ \Pr \{ x, c \} = \Pr \{ c = c \} \mathcal{N}(x | \mu_c, \Sigma) \]
- Maximum likelihood can be calculated by EM algorithm.
- Note that we need to perform inference as part of the learning algorithm.
- In example we calculate
  \[ E[C = c | X, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k] \]

Structure Learning

- Hard in general case; tractable in some cases (tree; order known)
- Heuristic search: Define a score function and perform local search to optimize network and parameters.
- Maximum likelihood: \( \arg \max_c \arg \max \Pr \{ D | G, \theta \} \)
- Overfitting may be an issue.

Bayesian Structure Learning

- Define priors \( \Pr \{ G \} \) and \( \Pr \{ \theta | G \} \).
- Marginal likelihood (evidence function) for \( G \):
  \[ \Pr \{ D | G \} = \int \Pr \{ D | \theta, G \} \Pr \{ \theta | G \} d\theta \]
- Posterior: \( \Pr \{ G | D \} \sim \Pr \{ D | G \} \Pr \{ G \} \)
- With large sample size \( M \) score (log posterior) becomes:
  \[ \text{score} \{ G | D \} = \log \Pr \{ D | G, \hat{\theta} \} + \log \Pr \{ G \} - \frac{1}{2} \text{Dim}(G) \log M \]

- Can search for MAP using heuristic search. NB this is a model selection problem.
- Posterior may be hard to calculate. Sample from posterior for Bayesian averaging.
Parameter Learning in MRFs

- The likelihood does not decouple due to normalization factor $Z$.
- $i$ index of cliques; $j$ index over examples

\[ L = \Pr \{ D | \theta \} = \prod_i \frac{1}{Z} e^{-\sum_j \Psi_i(X_{Qj})} \]

\[ LL = - \sum_j \sum_i \Psi_i(X_{Qj}) \cdot \sum_j \ln Z = - \sum_j \sum_i \Psi_i(X_{Qj}) \cdot M \ln Z \]

\[ \frac{\partial LL}{\partial \Psi_{i,v}} = - \sum_j I(x_{Qj}^i = v) + M \Pr \{ X_{Qj} = v \} \]

Therefore at the solution we have:

\[ \Pr \{ X_{Qj} = v \} = \frac{N_{X_{Qj}^i}}{M} \]

Optimize with gradient descent or other methods.