Likelihood Scores

Lecture XX

Reminder from Information Theory

- Mutual Information:

$$I(X,Y) = \sum\nolimits_{x\epsilon X} \sum\nolimits_{y\epsilon Y} P(x,y) log \frac{P(x,y)}{P(x)P(y)}$$

- Conditional Mutual Information:

$$I(X,Y,Z) = \sum\nolimits_{x\epsilon X} \sum\nolimits_{y\epsilon Y} \sum_{z\epsilon Z} P(x,y,z) log \frac{P(x,y|z)}{P(x|z)P(y|z)}$$

- Entropy: Conditional Mutual Information:

$$H(x) = \sum_{x\epsilon X} -P(x) log P(x)$$

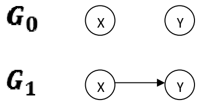
Scoring Maximum Likelihood Function

- When scoring function is the Maximum Likelihood, the model would make the data as probable as possible by choosing the graph structure that would produce the highest score for the MLE estimate of the parameter, we define:

$$Score(G; D) = log P(D|G, \theta_{ML})$$

Two Graph Structures

- Consider two simple graph structures:



$$Score(G_0; D) = \sum_{m \epsilon D} log\hat{\theta}_{x[m]} + log\hat{\theta}_{y[m]}$$

$$Score(G_1; D) = \sum_{m \epsilon D} log\hat{\theta}_{x[m]} + log\hat{\theta}_{y[m]|x[m]}$$

- The difference is:

$$Score(G_1; D) - Score(G_0; D) = \sum_{m \epsilon D} log\hat{\theta}_{y[m]|x[m]} - log\hat{\theta}_{y[m]}$$

Ex Continued

- **By counting how many times each conditional probability parameter appears in this term:**

$$= \sum_{x,y} M[x,y] log\hat{\theta}_{y|x} - M[y] log\hat{\theta}_{y}$$

- **When $\hat{P}$ is empirical distribution observed in the data**

$$= M \sum_{x,y} \hat{P}(x,y) log \frac{\hat{P}(y|x)}{\hat{P}(y)} = M \times I_{\hat{p}}(X,Y)$$

where $I_{\hat{p}}(X,Y)$ is the mutual information between X and Y in the distribution $\hat{P}$

The goal is to maximize the mutual information

Likelihood Score: General Networks

Entropy: to penalize addition of too many arcs between the nodes to prevent over fitting.

- Proposition:

$$\boldsymbol{Score}(G;D)$$

$$= M \sum_{i=1}^{n} I_{\hat{p}}(x_i, Pa_{x_i}^{G}) - M \sum_{i=1}^{n} H_{\hat{p}}(x_i)$$

Looping over all variables

Looping over all rows of CPT

- Proof:

$$= \sum_{i=1}^{n} \left[\sum_{\vec{u}_i \epsilon Val(Pa_{x_i}^{G})} \sum_{x_i \epsilon X_i} M[x_i, \vec{u}_i] log\hat{\theta}_{x_i|\vec{u}_i} \right]$$

All settings of this variable

$$= \frac{1}{M} \sum_{\vec{u}_i} \sum_{x_i \epsilon X_i} M[x_i, \vec{u}_i] log\hat{\theta}_{x_i|\vec{u}_i}$$

Proof Cont.

$$= \sum_{\vec{u}_i} \sum_{x_i \epsilon X_i} \hat{P}(x_i, \vec{u}_i) log \hat{P}(x_i|\vec{u}_i)$$

$$= \sum_{\vec{u}_i} \sum_{x_i \epsilon X_i} \hat{P}(x_i, \vec{u}_i) log \left(\frac{\hat{P}(x_i|\vec{u}_i)}{\hat{P}(\vec{u}_i)} \frac{\hat{P}(x_i)}{\hat{P}(x_i)} \right)$$

$$= \sum_{\vec{u}_i} \sum_{x_i \epsilon X_i} \hat{P}(x_i, \vec{u}_i) log \frac{\hat{P}(x_i, \vec{u}_i)}{\hat{P}(\vec{u}_i)\hat{P}(x_i)} - \sum_{x_i \epsilon X_i} \left(\sum_{\vec{u}_i} \hat{P}(x_i, \vec{u}_i) \right) log \hat{P}(x_i)$$

$$= I_{\hat{p}}(X_i, \vec{U}_i) - \sum_{x_i \epsilon X_i} \hat{P}(x_i) log \frac{1}{\hat{P}(x_i)}$$

$$= I_{\hat{p}}(X_i, \vec{U}_i) - H_{\hat{p}}(X_i)$$

Tree-Augmented Naïve Bayes (TAN) Model

- Bayesian network in which one node is distinguished as the Class node
- Arc from Class to every other node (feature node), as in naïve Bayes
- Remaining arcs form a directed tree among the feature nodes

TAN Learning Algorithm (guarantees maximum likelihood TAN model)

- Compute (based on data set) conditional mutual information between each pair of features, conditional on Class
- Compute the maximum weight spanning tree of the complete graph over features, with each edge weighted by conditional mutual information computed above
- Choose any feature as root and direct arcs from root, to get directed tree over features
- Add Class variable with arcs to all feature nodes, to get final network structure
- Learn parameters from data as for any other Bayes net