Why

Which is the optimal tree to use for tree distribution approximation?

Definition

We want to choose a tree whose corresponding approximator for the given distribution achieves minimum relative entropy with the given distribution among all tree distribution approximators. We call such a distribution an *optimal* tree approximator of the given distribution. We call a tree according to which an optimal tree approximator factors and *optimal* approximator tree.

Result

Proposition **1.** *Let $A_1, \ldots, A_n$ be finite non-empty sets. Define $A = \prod_{i=1}^{n} A_i$. Let $q : A \to [0,1]$ a distribution. A tree T on $\{1, \ldots, n\}$ is an optimal approximator tree if and only if it is a maximal spanning tree of the mutual information graph of q.*

Proof. First, denote the optimal tree distribution approximator of q for tree T by p_T^*. Express

$$p_T^* = q_1 \prod_{i \neq 1} q_{i|\mathbf{pa}_i}$$

Second, express $d(q,p) = H(q,p) - H(q)$. Since $H(q)$ does

not depend on T, p_T^* is a minimizer (w.r.t. T) of $d(q, p_T^*)$ if and only if it is a minimizer of $H(q, p_T^*)$.

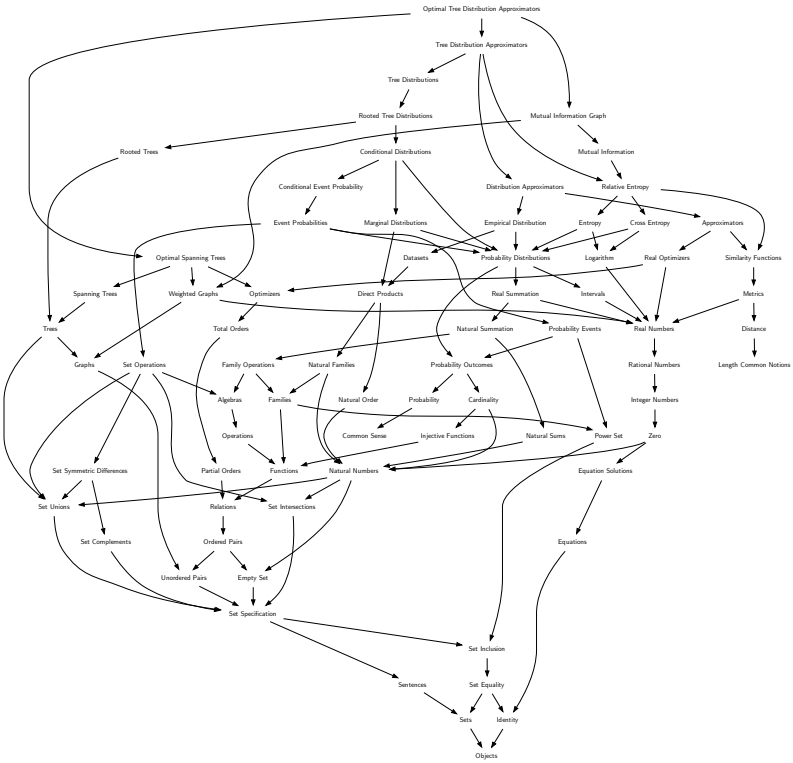
Third, express the cross entropy of p_T^* relative to q as

$$\begin{aligned}
H(q, p_T^*) &= H(q_1) - \sum_{i \neq 1} \sum_{a \in A} q(a) \log q_{i|pai}(a_i, a_{\mathbf{pa}_i}) \\
&= H(q_1) - \sum_{i \neq 1} \sum_{a \in A} q(a) \big(\log q_{i,\mathbf{pa}_i}(a_i, a_{\mathbf{pa}_i}) - \log q_{\mathbf{pa}_i}(a_{\mathbf{pa}_i}) \big) \\
&= H(q_1) - \sum_{i \neq 1} \sum_{a \in A} q(a) \big(\log q_{i,\mathbf{pa}_i}(a_i, a_{\mathbf{pa}_i}) - \log q_{\mathbf{pa}_i}(a_{\mathbf{pa}_i}) - \log q_i \\
&= \sum_{i=1}^{n} H(q_i) - \sum_{i \neq 1} I(q_i, q_{\mathbf{pa}_i}) \\
&= \sum_{i=1}^{n} H(q_i) - \sum_{\{i,j\} \in T} I(q_i, q_j)
\end{aligned}$$

where $\mathbf{pa}_i$ denotes the parent of vertex i in T rooted at vertex 1 ($i = 2, \ldots, n$). For $i = 1, \ldots, n$, $H(q_i)$ does not depend on the choice of tree. Therefore selecting a tree to minimize the second term in the final expression above is equivalent to choosing a maximal spanning tree from the weighted graph with mutual information edge weights; namely, the mutual information graph of q.

□

Proposition 1 says that to we should first select a maximum spanning tree of the mutual information graph of the distribution we are approximating. Then, we should pick the best approximator to q which factors according to that three.


Optimal Tree Distribution Approximators
Tree Distribution Approximators
Tree Distributions
Rooted Tree Distributions
Mutual Information Graph
Rooted Trees
Conditional Distributions
Mutual Information
Conditional Event Probability
Distribution Approximators
Relative Entropy
Event Probabilities
Marginal Distributions
Empirical Distribution
Entropy
Cross Entropy
Approximators
Datasets
Probability Distributions
Logarithm
Real Optimizers
Similarity Functions
Optimal Spanning Trees
Spanning Trees
Weighted Graphs
Optimizers
Direct Products
Real Summation
Intervals
Metrics
Trees
Total Orders
Natural Summation
Probability Events
Real Numbers
Distance
Graphs
Set Operations
Family Operations
Natural Families
Probability Outcomes
Rational Numbers
Length Common Notions
Algebras
Families
Natural Order
Probability
Cardinality
Integer Numbers
Operations
Common Sense
Injective Functions
Natural Sums
Power Set
Zero
Set Symmetric Differences
Partial Orders
Functions
Natural Numbers
Equation Solutions
Set Unions
Relations
Set Intersections
Set Complements
Ordered Pairs
Equations
Unordered Pairs
Empty Set
Set Specification
Set Inclusion
Sentences
Set Equality
Sets
Identity
Objects