Generalization of the Class Conditional Independence Assumption

Robert M. Haralick

Computer Science, Graduate Center
City University of New York

The Class Conditional Independence Assumption

- Measurement tuple $d = (d_1, \ldots, d_N)$
- Class c
- Then the class conditional independence assumption is

$$P(d \mid c) = P(d_1, \ldots, d_N \mid c) = \prod_{n=1}^{N} P(d_n \mid c)$$

- It is this assumption that is used in the Naive Bayes Classifier.
- There are many other kinds of conditional independence assumptions.

Markov Class Conditional Independence Assumption

$$P(x_n \mid x_{n+1} \dots x_N) = P(x_n \mid x_{n+1}), \;\; n = 1, \dots N-1$$

Conditioned by class

$$\begin{aligned} P(x_1 \dots x_N \mid c) &= \left[\prod_{n=1}^{N-1} P(x_n \mid x_{n+1} \dots x_N \mid c) \right] P(x_N \mid c) \\ &= \left[\prod_{n=1}^{N-1} P(x_n \mid x_{n+1} \mid c) \right] \; P(x_N \mid c) \end{aligned}$$

Assign $(x_1, \dots x_N)$ to class c^* when

$$\begin{aligned} P(x_1 \dots x_N \mid c^*) &> P(x_1 \dots x_N \mid c), c \neq c^* \\ \left[\prod_{n=1}^{N-1} P(x_n \mid x_{n+1}, c^*) \right] P(x_N \mid c^*) &> \left[\prod_{n=1}^{N-1} P(x_n \mid x_{n+1}, c) \right] P(x_N \mid c) \end{aligned}$$

for all other c

Markov Dependence Tree

$$P(x_1, \ldots, x_7 \mid c) = P(x_1 \mid x_2, c)P(x_2 \mid x_3, c)P(x_3 \mid x_4, c)P(x_4 \mid x_5, c)P(x_5 \mid x_6, c)P(x_6 \mid x_7, c)P(x_7 \mid c)$$

Precedence Function

i	j(i)
7	6
6	5
5	4
4	3
3	2
2	1

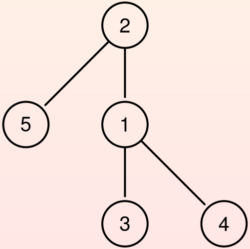
Dependence Trees

$$P(x_1, x_2, x_3, x_4, x_5 \mid c) = P(x_1 \mid x_2, c)P(x_5 \mid x_2, c)P(x_3 \mid x_1, c)P(x_4 \mid x_1, c)P(x_2 \mid c)$$

$$\sum_{x_2}\sum_{x_1}\sum_{x_4}\sum_{x_3}\sum_{x_5} P(x_1 \mid x_2, c)P(x_5 \mid x_2, c)P(x_3 \mid x_1, c)P(x_4 \mid x_1, c)P(x_2 \mid c) = 1$$

i	**j(i)**
1	2
5	2
3	1
4	1

Precedence Function



Product Probability Expansion

- If $Q(x|c)$ is a probability function and
- If $\{J_1, \ldots, J_K\}$ is a partition of $[1, N]$ then

$$Q(x \mid c) = \prod_{k=1}^{K} Q_k \left(\pi_{J_k}(x) \mid c \right) \tag{1}$$

- Is an example of a more general conditional independence assumption
- It is one of many kinds of conditional probability assumptions
- We can say that $Q(x \mid c)$ has a product probability expansion
- If it is not known that $Q(x \mid c)$ has product probability expansion (1), we may invoke the product probability expansion as an approximation

Product Probability Approximations

Definition

A Product Probability Approximation to a unknown joint distribution Q of N variables has the form

$$Q(x \mid c) = \prod_{k=1}^{K} Q_k \left(\pi_{J_k}(x) \mid c \right)$$

where

- $\{J_1, \ldots, J_K\}$ is a cover of $[1, N]$
- Q_k are arbitrary functions $Q_k > 0$
- $\sum_x Q(x \mid c) = 1$

Inverse Projection

Definition

- Let $I = [1, N]$ be the index set for the full space S
- Let the respective range sets for the N dimensions be $L_i,\ i \in I$
- Then $S = \times_{i \in I} L_i$, or in the indexed notation (I, S), is the full space
- Let $J \subset I$
- Let y be a tuple in the subspace indexed by J so that $y \in \times_{j \in J} L_j$; (J, y) is an indexed tuple

Then the Inverse Projection of (J, y) from the subspace indexed by $J \subset I$ is defined by

$$\pi_J^{-1}(J, y) = \{(I, x) \in (I, S) \mid \pi_J(I, x) = (J, y)\}$$

Extension

Definition

- Let there be N variables whose index set $I = \{1, \ldots, N\}$
- Let the range set for a variable whose index is j be L_j
- Let P be a probability distribution defined on the space $\times_{i \in I} L_i$
- Let $J_k \subset I$
- Define to P_{J_k} be the marginal distribution of P defined on the subspace indexed by J_k
  - $P_{J_k} : \times_{j \in J_k} L_j \to [0, 1]$
  - $P_{J_k}(J_k, y) = \sum_{(I,x) \in \pi_{J_k}^{-1}(J_k, y)} P(I, x)$

Then a probability distribution P defined on the subspace indexed by I is said to be an Extension of the given functions $Q_{J_k} : \times_{j \in J_k} L_j \to [0, 1],\ k = 1, \ldots, K$ if and only if

$$P_{J_k} = Q_{J_k},\ k = 1, \ldots, K$$

Maximum Entropy

- Lewis proved that of all distributions that are extensions of the given marginals
- The Product Approximation
  - Is the *closest* by the Kullback Liebler Divergence
  - And has the maximum entropy

P.M. Lewis, Approximating Probability Distributions to Reduce Storage Requirements, Information and Control Vol 2, 1959, pp. 214-225.

Kullback Liebler Divergence

Definition

The Kullback-Liebler Divergence (relative entropy) of a distribution P' to a reference distribution P is given by

$$D_{KL}(P||P') = \sum_{x} P(x) log \frac{P(x)}{P'(x)}$$

Although it is not a distance, it is said to be a measure of closeness of P' to the reference distribution P

Mutual Information

The largest entropy Product Probability Approximation problem was solved by Chow and Liu in 1968 for second order marginal probabilities using the optimal Kruskal's spanning tree algorithm. He used mutual information.

Definition

The Mutual Information between a random variable x and a random variable y is given by

$$I(x, y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

When $P(x, y) = P(x)P(y)$ the mutual information will be zero.

C.K. Chow and C.N. Liu, Approximating Discrete Probability Distributions with Dependence Trees, IEEE Transactions on Information Theory, Vol IT-14, No. 3, 1968 pp. 462-467.

Chow and Liu 1968

- Construct a weighted graph
  - If there are N variables, make an N-node graph
  - Label the nodes with the index of its variable
  - On the edge connecting node i and node j put the mutual information weight $I(x_i, x_j)$
- Use Kruskal's maximum spanning tree algorithm to find the spanning tree having maximum sum of weights
- The result will be a dependence tree
- With the precedence function of the tree, the joint probability will be the product of the conditional probabilities $P_{i|j}$ where j precedes i on the tree times the probability of the variable of the root

Another Generalization

It is also possible to generalize the class conditional independence assumption in a principled way and allow for overlapping index sets. We will illustrate with a small concrete example.

$$\frac{P_{134}(x_1, x_3, x_4) P_{352}(x_3, x_5, x_2)}{P_3(x_3)}$$

Notice that this form does define a probability distribution. Since each of the terms are positive, the fraction is non-negative. And the sum over all values for x_1, x_2, x_3, x_4, x_5 equals 1. To see how this works, sum on x_1, x_4 and discover that the total is 1.

$$\sum_{x_3,x_5,x_2} \sum_{x_1,x_4} \frac{P_{134}(x_1, x_3, x_4)P_{352}(x_3, x_5, x_2)}{P_3(x_3)}$$

$$= \sum_{x_3,x_5,x_2} \frac{P_3(x_3)P_{352}(x_3, x_5, x_2)}{P_3(x_3)}$$

$$= \sum_{x_3,x_5,x_2} P_{352}(x_3, x_5, x_2) = 1$$

Successive Marginals Overlap with One Variable

Let $J_1, \ldots, J_M$ be the index sets defining the subspaces.

$$J_a \cap J_b = \emptyset \text{ if } b > a + 1 \tag{2}$$

$$|J_a \cap J_{a+1}| \leq 1, \; a \in [1, M-1] \tag{3}$$

If $J_a \cap J_{a+1} \neq \emptyset$,

$$J_a \cap J_{a+1} \neq J_b \cap J_{b+1}, \quad a, b \in [1, M-1], \; a \neq b \tag{4}$$

- Constraint (2) requires that non-successive index sets in the ordering $\langle 1, 2, \ldots, M \rangle$ have no elements in common
- Constraint (3) requires that successive index sets have only one element in common
- Constraint (4) implies that the at most one element in common of successive index sets is unique

Successive Marginals Overlap with One Variable

If constraints (2), (3) and (4) are satisfied and $\{j_m\} = J_m - J_{m+1}, m = 1, \ldots, M-1$ then

$$\begin{aligned} P(I,x) &= \frac{\prod_{m=1}^{M} P_{J_m}\left(\pi_{J_m}(I,x)\right)}{\prod_{m=1}^{M-1} P_{j_m}\left(\pi_{j_m}(I,x)\right)} \\ &= \left(\prod_{m=1}^{M-1} \frac{P_{J_m}(\pi_{J_m}(I,x))}{P_{j_m}(\pi_{j_m}(I,x))}\right) P_{\pi_{J_M}}(\pi_{J_M}(I,x)) \\ &= \left(\prod_{m=1}^{M-1} P_{J_m}\left(\pi_{J_m - \{j_m\}}(I,x) \mid \pi_{j_m}(I,x)\right)\right) P_{J_M}\left(\pi_{J_M}(I,x)\right) \end{aligned}$$

is the largest entropy extension of the marginals $P_{J_1}, \ldots, P_{J_M}$

Dependence Tree Fourth Order

- $I = \{1, \ldots, N\}$
- N is dividable by 2
- There are $Q = N(N-1)/2$ size 2 subsets of I
- Call the subsets $W_1, \ldots, W_Q$
- Form a graph of Q nodes
- Connect node W_a with node W_b if and only if $W_a \cap W_b = \emptyset$
- On the edge between node W_a and W_b put mutual information weight $I(W_a, W_b)$ defined by

$$I(W_a, W_b) = \sum_x P_{W_a \cup W_b}\left(\pi_{W_a \cup W_b}(x)\right) log \frac{P_{W_a \cup W_b}\left(\pi_{W_a \cup W_b}(x)\right)}{P_{W_a}\left(\pi_{W_a}(x)\right) P_{W_b}\left(\pi_{W_b}(x)\right)}$$

Greedy Algorithm

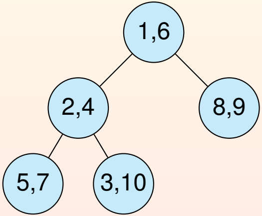
We construct a dependence tree with $N/2$ nodes and $N/2 - 1$ edges

- Choose the pair of nodes whose edge has the highest mutual information
- Successively connect a node to the tree being constructed having no overlap with those already selected and not forming a loop and having highest mutual information
- Continue until the tree has $N/2$ nodes whose associated index subsets form a partition of I

Dependence Tree Example

$$P(x_1, \ldots, x_{10} \mid c) = P(x_2, x_4 \mid x_1, x_6, c) P(x_5, x_7 \mid x_2, x_4, c) \times$$
$$P(x_3, x_{10} \mid x_2, x_4, c) P(x_8, x_9 \mid x_1, x_6, c)$$

$$P((I, x) \mid c) = P_{\{24|16\}} \left(\pi_{\{24\}}(I, x) \mid \pi_{\{16\}}(I, x), c\right) P_{\{57|24\}} \left(\pi_{\{57\}}(I, x) \mid \pi_{\{24\}}(I, x), c\right) \times$$
$$P_{\{3,10|24\}} \left(\pi_{\{3,10\}}(I, x) \mid \pi_{\{24\}}(I, x), c\right) P_{\{89|16\}} \left(\pi_{\{89\}}(I, x) \mid \pi_{\{16\}}(I, x), c\right)$$



Shows a dependence tree example for a measurement tuple with 10 components. Since each node has 2 indexes, the tree has five nodes. Each edge is associated with the pair of indexes in the upper node combined with the pair of indexes in the lower node thus forming a size 4 index set, indicating an explicit dependence among the index sets.

Graphical Models

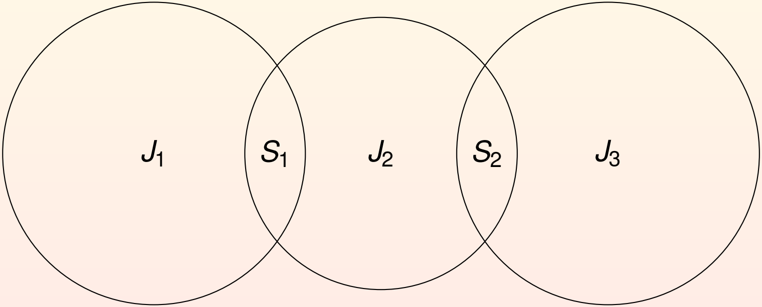
- All the examples we have shown are specializations
- Let I be the index set for the random variables; $I = \{1, \ldots, N\}$
- Let $\langle J_1, \ldots J_M \rangle$ be ordered index sets
  - Require $J_m \cap J_{m+1} = S_m \neq \emptyset, \;\; m = 1, \ldots, M-1$
  - Require $J_m \cap J_n = \emptyset, \;\; n > m+1$
  - Construct a graph. Make $J_1, \ldots, J_M$ be complete subgraphs
  - Verify that $J_1, \ldots, J_M$ are cliques of the graph
  - The graph will be chordal
- $S_m, \;\; m = 1, \ldots, M-1$ are the separators

Then

$$P_I(I, x) = \frac{\prod_{m=1}^{M} P_{J_m}\left(\pi_{J_m}(I, x)\right)}{\prod_{m=1}^{M-1} P_{S_m}\left(\pi_{S_m}(I, x)\right)}$$

is the largest entropy distribution that is an extension of $P_{J_m}, m = 1, \ldots, M$

$$P(I,x) = \frac{P_{J_1}\left(\pi_{J_1}(I,x)\right) P_{J_2}\left(\pi_{J_2}(I,x)\right) P_{J_3}\left(\pi_{J_3}(I,x)\right)}{P_{S_1}\left(\pi_{S_1}(I,x)\right) P_{S_2}\left(\pi_{S_2}(I,x)\right)}$$



S_1 is called a separator of the nodes in J_1 and the nodes in J_2 because if the nodes in S_1 are deleted, what remains of J_1 and J_2 are separated. In fact, if the nodes in S_1 are deleted, $J_1 - S_1$ and $(J_2 - S_2) \cup J_3$ are separated.

Impact on the N-tuple Subspace Classifier

The Bledsoe and Browning N-tuple subspace classifier breaks the full space into mutually exclusive subspaces and for each subspace, estimates the class conditional probabilities.

Suppose $\{H_1, H_2, \ldots H_Y\}$ and $\{J_1, J_2, \ldots J_Z\}$ are each covers of I, the index set for the full space, satisfying the conditions of the previous slides. Then for any class c, we can obtain two class conditional probabilities P_H and P_J. How can we utilize these two class Conditional Probability Functions? There are two natural possibilities:

- $S((I,x)|c) = P_H((I,x)|c)P_J((I,x)|c)$
- $S((I,x)|c) = P_H((I,x)|c) + P_J((I,x)|c)$

Then assign (I, x) to class c satisfying

$$S((I,x)|c) > S((I,x)|c'),\ c' \neq c$$