# Inference and Representation, Fall 2016

## Problem Set 3: Exact inference & Structure learning

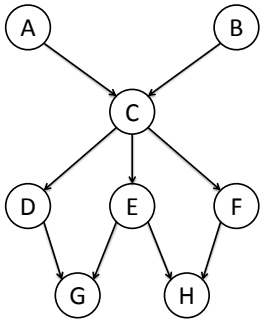
**Due: Monday, October 3, 2016 at 3pm (as a PDF file uploaded to Gradescope)**

**Important:** *See problem set policy on the course web site.*

---

*For question 3, you are allowed to use basic graph packages (e.g., for representing and working with undirected graphs, or for finding the maximum spanning tree), but are* **not** *permitted to use any machine learning, graphical models, or probabilistic inference packages.*

1. Consider the Bayesian network shown below, and answer the following questions. You may assume that the random variables are binary-valued, i.e. take states in $\{0, 1\}$.




   (a) Moralize the Bayesian network (submit a drawing of the new graph). What edges are added?

   (b) Give a perfect elimination ordering, i.e. one that yields no fill edges.

   (c) Give an elimination ordering that results in the induced graph having $\geq 5$ nodes in one (or more) cliques.

   (d) Suppose we want to compute the query $\Pr(B = 0 \mid G = 1)$. Prove that H and F are irrelevant variables with respect to this query. That is, show that it is possible to prune the Bayesian network, removing H and F, while not changing the value of $\Pr(B = 0 \mid G = 1)$.

   (e) Walk through the execution of the variable elimination algorithm to compute $\Pr(B = 0 \mid G = 1)$, using as few computations as necessary (i.e., using the elimination ordering given in (b), and using the simplication given by your answer to (d)).

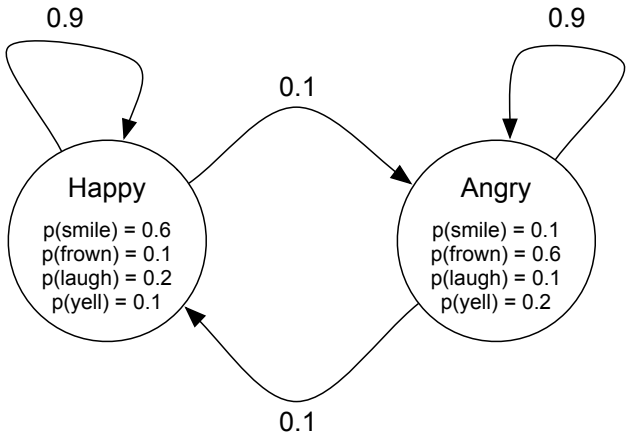
2. **Hidden Markov models**. Harry lives a simple life. Some days he is Angry and some days he is Happy. But he hides his emotional state, and so all we can observe is whether he smiles, frowns, laughs, or yells. Harry's best friend is utterly confused about whether Harry is actually happy or angry and decides to model his emotional state using a hidden Markov model.

   Let $X_d \in \{\text{Happy, Angry}\}$ denote Harry's emotional state on day $d$, and let $Y_d \in \{\text{smile, frown, laugh, yell}\}$ denote the observation made about Harry on day $d$. **Assume that on**

**day 1 Harry is in the Happy state**, i.e. $X_1 = \text{Happy}$. Furthermore, assume that Harry transitions between states exactly once per day (staying in the same state is an option) according to the following distribution: $p(X_{d+1} = \text{Happy} \mid X_d = \text{Angry}) = 0.1$, $p(X_{d+1} = \text{Angry} \mid X_d = \text{Happy}) = 0.1$, $p(X_{d+1} = \text{Angry} \mid X_d = \text{Angry}) = 0.9$, and $p(X_{d+1} = \text{Happy} \mid X_d = \text{Happy}) = 0.9$.

The observation distribution for Harry's Happy state is given by $p(Y_d = \text{smile} \mid X_d = \text{Happy}) = 0.6, p(Y_d = \text{frown} \mid X_d = \text{Happy}) = 0.1, p(Y_d = \text{laugh} \mid X_d = \text{Happy}) = 0.2$, and $p(Y_d = \text{yell} \mid X_d = \text{Happy}) = 0.1$. The observation distribution for Harry's Angry state is $p(Y_d = \text{smile} \mid X_d = \text{Angry}) = 0.1, p(Y_d = \text{frown} \mid X_d = \text{Angry}) = 0.6, p(Y_d = \text{laugh} \mid X_d = \text{Angry}) = 0.1$, and $p(Y_d = \text{yell} \mid X_d = \text{Angry}) = 0.2$.

**All of this is summarized in the following figure**:




**Be sure to show all of your work for the below questions.** Note, the goal of this question is to get you to start thinking deeply about probabilistic inference. Thus, although you could look at Chapter 17 for an overview of HMMs, try to solve this question based on first principles (also: no programming needed!).

(a) What is $p(X_2 = \text{Happy})$?

(b) What is $p(Y_2 = \text{frown})$?

(c) What is $p(X_2 = \text{Happy} \mid Y_2 = \text{frown})$?

(d) What is $p(Y_{80} = \text{yell})$?

(e) Assume that $Y_1 = Y_2 = Y_3 = Y_4 = Y_5 = \text{frown}$. What is the most likely sequence of the states? That is, compute the MAP assignment $\arg\max_{x_1,\ldots,x_5} p(X_1 = x_1, \ldots, X_5 = x_5 \mid Y_1 = Y_2 = Y_3 = Y_4 = Y_5 = \text{frown})$.

3. **Chow-Liu algorithm.** When trying to do object detection from computer images, *context* can be very helpful. For example, if "car" and "road" are present in an image, then it is likely that "building" and "sky" are present as well (see Figure 1). In recent work, a tree-structured Markov random field (see Figure 2) was shown to be particularly useful for modeling the prior distribution of what objects are present in images and using this to improve object detection [1].

   You will replicate some of the results from [1] (it is not necessary to read this paper to complete this assignment). Specifically, you will implement the Chow-Liu algorithm

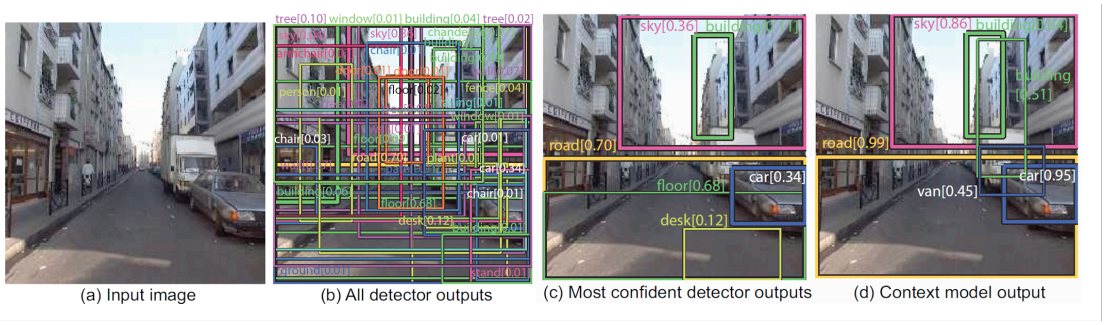



Figure 1: Using context within object detection for computer vision. [1]

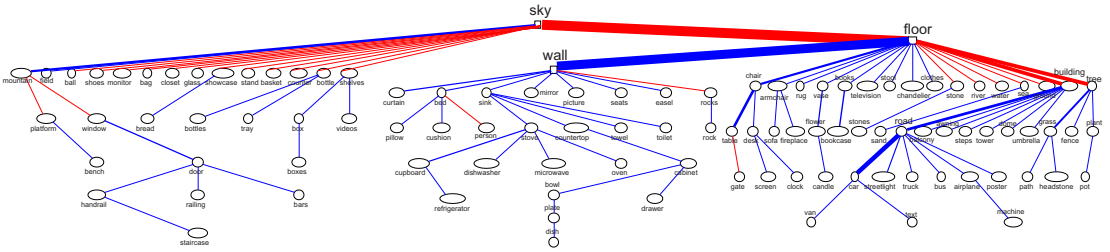



Figure 2: Pairwise MRF of object class presences in images [1]. Red edges denote negative correlations between classes. The thickness of each edge represents the strength of the link. You will be learning this MRF in question 3.

(1968) for maximum likelihood learning of tree-structured Markov random fields [2]. See also Murphy's book Section 26.3 for a brief overview (the Murphy book is available online for free for NYU students; see course website).

The goal of learning is to find the tree-structured distribution $p_T(\mathbf{x})$ that maximizes the log-likelihood of the training data $\mathcal{D} = \{\mathbf{x}\}$:

$$\max_T \max_{\theta_T} \sum_{\mathbf{x} \in \mathcal{D}} \log p_T(\mathbf{x}; \theta_T).$$

We will show in Lecture 9 that for a fixed structure $T$, the maximum likelihood parameters for a MRF have a property called *moment matching*, meaning that the learned distribution will have marginals $p_T(x_i, x_j)$ equal to the empirical marginals $\hat{p}(x_i, x_j)$ computed from the data $\mathcal{D}$, i.e. $\hat{p}(x_i, x_j) = count(x_i, x_j)/|\mathcal{D}|$ where $count(x_i, x_j)$ is the number of data points in $\mathcal{D}$ with $X_i = x_i$ and $X_j = x_j$. Thus, using the factorization from Eq. (2) of question 4 of PS2, the learning task is reduced to solving

$$\max_T \sum_{\mathbf{x} \in \mathcal{D}} \log \left[ \prod_{(i,j) \in T} \frac{\hat{p}(x_i, x_j)}{\hat{p}(x_i)\hat{p}(x_j)} \prod_{j \in V} \hat{p}(x_j) \right].$$

We can simplify the quantity being maximized over $T$ as follows (let $N = |\mathcal{D}|$):

$$
\begin{aligned}
&= \sum_{\mathbf{x}\in\mathcal{D}} \Big( \sum_{(i,j)\in T} \log\left[\frac{\hat{p}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}\right] + \sum_{j\in V} \log\left[\hat{p}(x_j)\right] \Big) \\
&= \sum_{(i,j)\in T} \sum_{\mathbf{x}\in\mathcal{D}} \log\left[\frac{\hat{p}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}\right] + \sum_{j\in V} \sum_{\mathbf{x}\in\mathcal{D}} \log\left[\hat{p}(x_j)\right] \\
&= \sum_{(i,j)\in T} \sum_{x_i,x_j} N\hat{p}(x_i,x_j) \log\left[\frac{\hat{p}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}\right] + \sum_{j\in V} \sum_{x_i} N\hat{p}(x_i) \log\left[\hat{p}(x_j)\right] \\
&= N\Big( \sum_{(i,j)\in T} I_{\hat{p}}(X_i,X_j) - \sum_{j\in V} H_{\hat{p}}(X_j) \Big),
\end{aligned}
$$

where $I_{\hat{p}}(X_i,X_j) = \sum_{x_i,x_j} \hat{p}(x_i,x_j) \log \frac{\hat{p}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}$ is the empirical *mutual information* of variables $X_i$ and $X_j$, and $H_{\hat{p}}(X_i)$ is the empirical *entropy* of variable $X_i$. Since the entropy terms are not a function of $T$, these can be ignored for the purpose of finding the maximum likelihood tree structure. **We conclude that the maximum likelihood tree can be obtained by finding the maximum-weight spanning tree in a complete graph with edge weights $I_{\hat{p}}(X_i,X_j)$ for each edge $(i,j)$.**

The Chow-Liu algorithm then consists of the following two steps:

(a) Compute each edge weight based on the empirical mutual information.

(b) Find a maximum spanning tree (MST) via Kruskal or Prim's Algorithm.

(c) Output a pairwise MRF with edge potentials $\phi_{ij}(x_i,x_j) = \frac{\hat{p}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}$ for each $(i,j)\in T$ and node potentials $\phi_i(x_i) = \hat{p}(x_i)$.

We have one random variable $X_i \in \{0,1\}$ for each object type (e.g., "car" or "road") specifying whether this object is present in a given image. For this problem, you are provided with a matrix of dimension $N \times M$ where $N = 4367$ is the number of images in the training set and $M = 111$ is the number of object types. This data is in the file "chowliu-input.txt", and the file "names.txt" specifies the object names corresponding to each column.

Implement the Chow-Liu algorithm described above to learn the maximum likelihood tree-structured MRF from the data provided. Your code should output the MRF in the standard UAI format described here:

`http://www.hlt.utdallas.edu/~vgogate/uai14-competition/modelformat.html`

# References

[1] Myung Jin Choi, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. Exploiting hierarchical context on a large database of object categories. In *IEEE Conference on Computer VIsion and Pattern Recognition (CVPR)*, 2010.

[2] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.