STAT 535 Lecture 1
September 28, 2011
Basics of multivariate inference
©Marina Meilă
mmp@stat.washington.edu

For the first part of 535, we will be concerned with

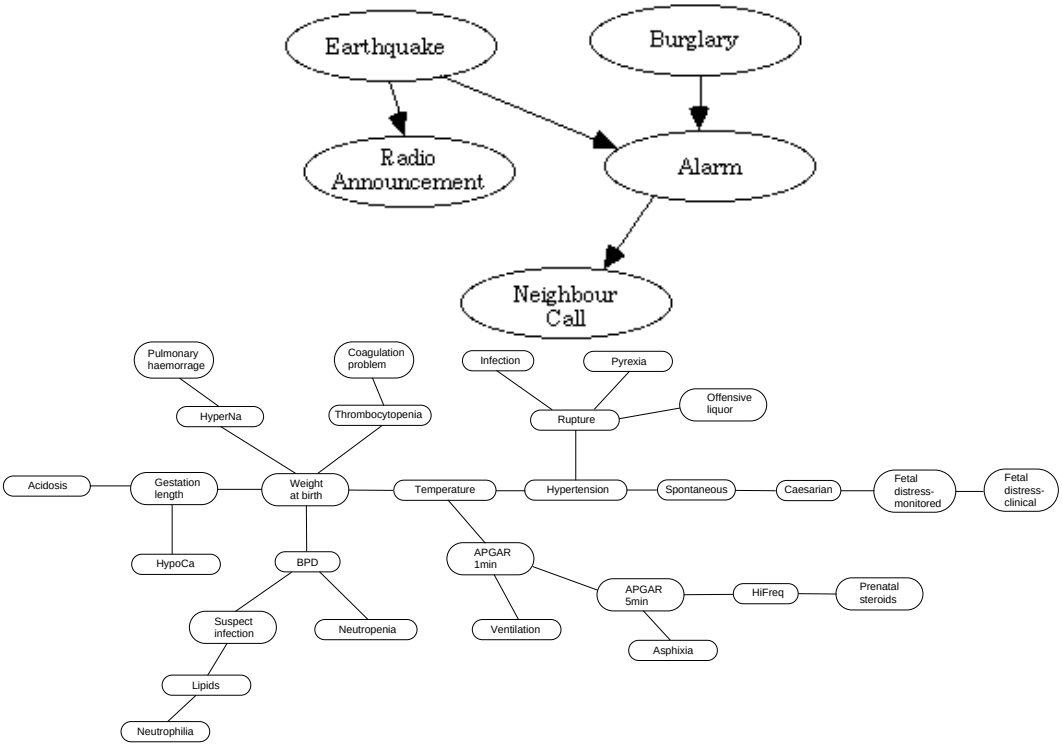
- Conditional independence (why?)
- Graphical probablity models/Belief networks (a language for expressing conditional independencies)
- Algorithms, especially related to graphs, and to statistical inference (how to make inference efficient? how fast can it be?)

- What the above three have in common is **efficient multivariate statistical inference**

1 Belief networks

Belief networks are probability models over multivariate domains. In this course, to simplify matters, we will be concerned with domains where all the variables are discrete, but everything we learn applies at least conceptually to continuous domains.

Here is are two example of multivariate domains described by a belief network:



Why should we learn about them?

- Powerful models (can represent a very rich class of distributions)
- Can represent distributions compactly
- Are intuitive
- Efficient and general algorithms for statistical inference.
  Statistical inference = computing $P(X = x|Y = y)$
- Efficient and general algorithms for estimating the model parameters (by e.g Maximum Likelihood)

2 Algorithms

- graph algorithms: e.g Minimum Spanning Tree, Matrix Tree Theorem
- belief network algorithms: e.g the Junction Tree algorithm, the Forward-Backward algorithm for HMM's revisited, message-passing algorithms
- other general purpose algorithms: e.g algorithms for disjoint sets

A common theme will be that big, complex sample spaces can often be partitioned into subsets that can be tackled independently, after a set of "conditioning" variables have been fixed. For example, given two consecutive exact position measurements for a GPS satellite at times t and $t+1$ (two so as to derive velocity information), prior history should not influence future behavior, and the sets of random variables $\leq t+1$ and $\geq t$ can be analyzed independently. In computer science, this is commonly known as *divide-and-conquer*, a close cousin of *dynamic programming*.

3 Multivariate distributions and statistical inference

Notations:

$V = \{X_1, X_2, \ldots X_n\}$	the domain
$n = \lvert V \rvert$	the number of variables, or the dimension of the domain
$\Omega(X_i)$	the domain of variable X_i (sometimes denoted Ω_{X_i})
$r_i = \lvert \Omega(X_i) \rvert$	(we assume variables are discrete)
$P_{X_1,X_2,\ldots X_n}$	the joint distribution (sometimes denoted $P(X_1, X_2, \ldots X_n)$
$A, B \subseteq V$	disjoint subsets of variables, $C = V \setminus (AB)$

Example: The "Chest clinic" example - a domain with several discrete variables.

$$\text{Smoker} \in \{Y, N\} = \Omega(S)$$

$$\text{Dyspnoea} \in \{Y, N\} = \Omega(D)$$

$$\text{Lung cancer} \in \{\text{no}, \text{incipient}, \text{advanced}\} = \Omega(L)$$

$$\text{Bronchitis} \in \{Y, N\} = \Omega(B)$$

$$V = \{S, D, L, B\}$$

The domain has 4 variables, $2 \times 2 \times 3 \times 2 = 24$ possible configurations. The **joint probability distribution** $P_{SDLB}(s, d, l, b)$ is real valued function on $\Omega(\{S, D, L, B\}) = \Omega(S) \times \Omega(D) \times \Omega(L) \times \Omega(B)$. We sometimes call it a **multidimensional probability table**.

The **marginal** distribution of S, L is

$$P_{SL}(s, l) = \sum_{d \in \Omega(D)} \sum_{b \in \Omega(B)} P_{SDLB}(s, d, l, b)$$

The **conditional** distribution of Lung cancer given Smoking is

$$P_{L|S}(l|s) = \frac{P_{SL}(s, l)}{P_S(s)}$$

Computing the probabilities of some variables of interest (L) when we observe others (S) and we don't know anything about the rest (B, D) is a fundamental operation in probabilistic reasoning called "statistical inference in the model P_{SDLB}".

We define by **statistical inference** the (partial) computation of the conditional distribution $P_{A|B=b}$, where A, B are as defined above, and b is in $\Omega(B)$. We call A the **variable of interest** or the **query variable**, B are the **observed variables** or the **conditioning variables**, and $B = b$ is the **evidence**.

Sometimes, inference will denote computing only a single value $P_{A|B}(a|b)$, and sometimes it will denote computing a statistic under $P_{A|B=b}$, like for instance the *mode* $\max_a P_{A|B=b}$ and $\operatorname{argmax}_a P_{A|B=b}$.

Examples of inference operations

- Bayesian estimation. Model $P(X|\theta)$, prior $P(\theta)$, evidence=data=sample $X_1, \ldots X_n$; inference is calculating the posterior $P(\theta|X_1, \ldots X_n)$

- Maximum Likelihood (ML) estimation (if the uniform prior exists on $\Omega(\theta)$. Assume the prior $P(\theta)$ is uniform. Then ML estimation is equivalent to the $\underset{\theta}{\operatorname{argmax}} P(\theta|X_1, \ldots X_n)$
- Regression and classification
- Diagnosis
- Real life (or literary fiction)
- Science and engineering: next section

4 Multivariate distributions in applications

We will generally be concerned with sample spaces expressed as a product over a finite set of random variables, often discrete:

- **Spatial localization** Variables: $X_{it} \in \mathbb{R}^3, Y_{it} \in \mathbb{R}, L_t \in \mathbb{R}^2$, $t = 1:T$, $i = 1:4$

  Sample space is path of ground observer $L_{1:T}$ and constellation of 4 GPS satellites $X_{1:4,1:T}$ over a sequence of time steps, and signal delay measurement used by observer $Y_{1:4,1:t}$ to estimate position (using model of satellites). Note that in this case the observer can not directly measure satellite positions. (In common usage, problem also incorporates inertial guidance measurements.) A typical problem: what is most probable ground location at the moment, given recent measurements?

- **Medical diagnosis** Variables: $D_i, F_j \in \{0, 1\}$, $i = 1:d$, $j = 1:f$

  D_i represent the presence or absence of an underlying disease; F_j is the result of a diagnostic *finding* (a medical test, a condition of the patient). A typical problem: what is posterior marginal probability of each disease given a small set of known findings? Another problem: what is the further information value of performing each test?

- **Theoretical physics** Variables: $Q_{ij} \in \{-1, +1\}$, $i, j = 1:k$

Sample space is idealized 2D lattice of electrons with spins. A typical problem: what is covariance of two spatially separated electrons? Another: what is average size of connected components with identical spins? Another: how do these answers depend on the temperature?

- **Language modeling** Variables: [parses, meanings], $X_i \in$ [`A-Z`], $i = 1 \ldots N$

  Sample space is sentences of length N. A typical problem: which of two sentences are more probable? Find the suject/object of this sentence. Translate it into Chinese. Is this sentence about cars? Is this sentence making a positive/negative comment about topic X?

- **Image processing** Variables: [contents descriptors], $B_{ij} \in [0, 1]$, $i, j = 1 \ldots 1000$

  Sample space is binary images of width and height 1000. A typical problem: find most probably reconstruction of image that is missing pixels. Does this image contain a car/a cheetah/a person? Which direction is the person facing? How many red blood cells are in this image?

On each of these sample spaces we will assume a joint probability distribution (or pdf in the case of continuous variables) over the primitive events - particular *configurations* or *assignments* to the variables. This probability distribution $P_{X_1 \ldots X_n}(x_1 \ldots x_n)$ could be:

- estimated from data (for example, from patient case histories in the disease/test case, in which case records are likely to be missing values for many of the variables);
- elicited from expert opinion (for example, by asking doctors to describe their internal models or using gambling-based games to extract them);
- derived from models (in the case of GPS, from Newton's laws, special and general relativity, stochastic models of observer motion and atmospheric effects on signal propagation);
- assumed axiomatically (as in the case of an idealized electron lattice).

Notice that in these examples one would expect strong dependencies between at least some of the variables (perhaps those that are close neighbors in time or space), and that the sizes of the sample space are quite large, likely far too large to

- store the joint distribution as a table associating a probability with each event;
- directly estimate the entries in such a table from a dataset;
- use a computer to process each configuration individually, for example to compute $E_S[f] = \sum_x P(x)f(x)$ or $\text{argmax}_x P(x)$.

In some cases it's easy to imagine incremental causal generative models where some variables are the "cause" of others, as as in the GPS and language modeling examples; in others like the electron spin example there doesn't seem a natural causal order among the variables.

5 How complex are operations with multivariate distributions?

Number of configurations $|\Omega(V)| = \prod_{i=1}^{n} r_i \geq 2^n$. Required storage depends **exponentially** on n!

Sampling: can be done in logarithmic time in the size of $\Omega(V)$, thus is $\mathcal{O}(n)$.

Returning the probability of a configuration is also $\mathcal{O}(n)$.

Computing the marginal of $X_1, \ldots X_k$ takes $\left(\prod_{i=1}^{k} r_i\right)\left(\prod_{i=k+1}^{n} r_i\right) = |\Omega(V)|$ additions. Also exponential.

Computing conditional distributions: they are ratios of two marginals $\Rightarrow$ also exponential.

$$P_{A|B} = \frac{P_{AB}}{P_B} \tag{1}$$

$$P_{AB}(a,b) = \sum_{c\in\Omega(C)} P_V(a,b,c) \tag{2}$$

$$P_B(b) = \sum_{a\in\Omega(A)} \sum_{c\in\Omega(C)} P_V(a,b,c) \tag{3}$$

$$= \sum_{a\in\Omega(A)} P_{AB}(a,b) \tag{4}$$

Hence, $P_B(b)$ is the normalization constant that turns $P_{AB}(\cdot, b)$ into $P_{A|B}(\cdot|b)$.

In conclusion, a multivariate probability table becomes intractable when the number of variables is large (practically over $10 - 20$). A solution to alleviate this problem (but ONLY in special cases) is offered by **graphical probability models**. They have the potential for compact representation and for efficient computations.

6 The many views of statistical inference

	Probability	**AI**	**Computation**
P_V	joint distribution	(state of) knowledge	multidimensional array
P_A, $A \subseteq V$	marginal	(state of) belief (about A)	sum over subarrays
$B = b$	evidence	observation	setting an index
$P_{V\setminus B\|B=b}$	conditional distribution	revised knowledge	extract sub-array, normalize (sum, divide by const) (corresponding to $B = b$)
$P_{A\|B=b}$	inference	revised belief in A	sum in sub-array (corresponding to $B = b$), normalize (sum, divide by const)