Graphical Models: The Why
10708, Fall 2020
Pradeep Ravikumar

1 Statistical Models

Artificial Intelligence is concerned with "intelligently" reasoning (and potentially acting based on this reasoning) in complex domains. Most modern approaches to do so are based on so-called statistical models. Here, we proceed in two stages. The first is a model building stage: we extract and summarize information in "data" via a statistical model. The second is a probabilistic reasoning stage: we use the statistical model to then draw intelligent "inferences" about the complex domain.

What is a statistical model? Here, we assume that key aspects of the complex system can be characterized via random variables, and the overall complex system can then be characterized by a joint distribution over these random variables: a statistical model. A key advantage of such statistical models is their ability to handle **uncertainty**. When we observe data from any complex domain, we are faced with three forms of uncertainty. The first is due to the limitations of the measurement and observation processes themselves. For instance in a medical diagnosis setting, there is likely inherent noise when we measure symptoms of patients. The second is due to inability to observe or model all aspects of the complex domain. We likely will not be able to get a complete picture of all the symptoms of the patient, so that the relationships between the symptoms we do observe and diseases cannot be typically expressed via deterministic relationships. The third is that there could be inherent non-determinism in the complex domain. While this is certainly the case at the level of quantum mechanics, this is also typically true even at the macro-scopic level. For instance, for any disease, the manifested symptoms could vary, in the very least with respect to severity. Due to all these reasons, it is critical that models of complex domains be able to represent uncertainty.

Given such a statistical model, one can then use *probabilistic reasoning*, to draw inferences about the underlying complex system. For instance, if we observe that the patient has a headache, and a rash on their back, but no fever, what are the probabilities of various diseases, or even simply, what is the most probable disease?

So far we have talked about two of three main legs of the use of AI in a complex domain: **representation**, and **reasoning**. The last leg is **learning**: we should be able to learn the statistical model in an automated way from data, ideally also incorporating any existing domain knowledge seamlessly in the process.

To summarize the overall approach, it is useful to think about the observation process underlying the complex domain which provides us with the given data, as a "forward mapping"

from some unknown domain state to the given observations. We use a statistical model to represent this forward mapping, where the unknown domain state is represented by the statistical model parameters. The model building stage is then the "inverse" problem of inductively inferring the unknown domain state – represented by the statistical model params – given observations.And then finally, we use probabilistic reasoning to draw inferences from the statistical model, this is a largely deductive process, but which gets non-trivial for large-scale models, where it then involves function approximation and optimization theory.

1.1 Do we really need statistical models

In the statistical model driven approach to ML discussed above, "learning" involves inferring the statistical model given the data, which can then be used for probabilistic reasoning, and making intelligent decisions. But in many ML contexts, one can directly "learn" the decision function without necessarily going through the route of inferring a statistical model over the data. This is particularly the case in supervised learning tasks of classification and regression, where we aim to learn a decision function that predicts an output given an input. And also for the unsupervised learning tasks such as clustering where we aim to group a set of data points. This uses the framework of statistical decision theory: we specify a class of target decision functions, and then a loss function that can evaluate the goodness of fit of any candidate decision function to the given data, and then finally, we optimize this goodness of fit measure to estimate the "optimal" decision function.

But in many cases such a decision-theoretic approach is not ideal, and we might want to take the scenic statistical model route instead:

(a) We might not have any specific target decision function in mind, or perhaps we have multiple decision functions in mind.

(b) We are not clear how to evaluate a good decision function (i.e. we are not sure about the loss function to use).

(c) We aim to summarize the overall data generation process via a statistical model. Use cases include general purpose "probabilistic reasoning," as well as generating additional "simulated" data from the statistical model.

Even if these use-cases are not applicable, there are *theoretical reasons* why we might want to go through the route of first inferring a statistical model:

(a) In many cases, it might not be possible to directly learn the optimal decision function from the given data, when the loss function explicitly involves the unknown statistical model parameters.

(b) If one were to take a Bayesian inferential philosophical viewpoint, the "rational" approach is to first update our uncertainty of the underlying statistical model given data, and then use this to compute the optimal downstream decision function.

For all these reasons, it is useful to learn a statistical model given data.

2 Why Probabilistic Graphical Models

OK, so we want to use statistical models. But which class of models to use? The field of statistical machine learning has byzantine multitudes of statistical models to choose from. From your intro ML classes, it might seem that ML is essentially a curated zoo of models and tools, with not much commonality among these. However not many of these models are suitable for large scale systems with a large number of variables. The key challenge with the use of statistical models in such large scale contexts is computational. In order to represent complex systems, we need joint distributions over a large number of variables. Under general settings, the representational complexity of statistical models however scales exponentially with the number of variables, which is obviously infeasible for large domains. And even if we are somehow able to represent such large scale distributions, probabilistic reasoning is in turn computationally hard, in the worst case scaling exponentially with the number of variables.

The main attraction of Probabilistic Graphical Models (PGMs) is that they allow for tractable representation and reasoning; drawing from tools in graph theory, probability theory, functional analysis, and statistical physics, among others. And over the last decade we now also have a suite of rigorous learning algorithms with provable guarantees, to learn PGMs from data. Technically PGMs are not really just a single model class per se, but more of a meta-model class. It provides a natural set of model class restrictions that allows us to represent large scale models compactly i.e. using few parameters.

Another key attraction of PGMs is that they are very interpretable: one can visualize the distribution using a graph (hence, *graphical* model). Relatedly, one could focus merely on extracting this graph structure (rather than a full statistical model), which has a much better notion of "direct dependence" among variables: a crucial object of interest in many scientific applications. In this lecture, we will be looking at these two facets of graphical models in greater detail. But before doing so, let's look at the varied applications of graphical models:

- Computer Vision
- Natural Language Processing
- Computational Biology

- Medical Diagnosis
- Computer Graphics
- Document Analysis
- Finance and Economics
- ...

We will be covering the first three application domains in various case studies throughout the course.

3 Compact Representation of Probability Distributions

Consider a discrete random vector $X = (X_1, \ldots, X_p)$ where each variable $X_i \in [r] := \{1, \ldots, r\}$ takes r possible values. The set of possible values $x \in [r]^p$,for the entire random vector is then r^p. Thus, to store the entire distribution $P(X)$ via a "probability table", with one real number for each configuration, would then we require we store $r^n - 1$ values (one less since they have to sum to one). This is obviously untenably large for large n, even for binary variables with $r = 2$.

3.1 Directed Graphical Models (DGMs)

Consider a directed acyclic graph (DAG) $G = (V, E)$, where each node $s \in V := [p]$ is associated with a random variable X_s, and the edge set $E \subseteq V \times V$ allows no directed cycles. For any node $i \in V$, the following graph-theoretic terms will be handy:

- parents: $\text{PA}_i \subseteq V$
- ancestors: $\text{ANCES}_i \subseteq V$
- children: $\text{CHILD}_i \subseteq V$
- descendants: $\text{DESC}_i \subseteq V$
- non-descendants: $\text{NON-DESC}_i \subseteq V$

Any such DAG G then corresponds to a *set of distributions* that "respect" the graph. Specifically, the joint distribution P over $X = (X_1, \ldots, X_p)$ is said to **factor according to the**

graph G if it can be written as:

$$P_{\text{fact}}(x) = \prod_{i=1}^{p} P(X_i | X_{\text{PA}_i}). \tag{1}$$

Now, the RHS is a product of a bunch of conditional probabilities. In general, it need not correspond to a valid joint distribution. But we have the following convenient proposition.

Proposition 1 *The function $P_{fact}(\cdot)$ specified in Eqn.* (1) *is a valid joint distribution over X when graph G is a DAG.*

We can show this by induction: consider leaf nodes, with no children. Let X_s be such a node. Then, by the factorization above, $P_{\text{fact}}(X) = P(X_s | X_{\text{PA}_s}) P_{\text{fact}}(X_{-s})$. By induction, if $P_{\text{fact}}(X_{-s})$ is a valid joint distribution over X_{-s}, then the above is a valid joint distribution over $X = (X_s, X_{-s})$.

Why is it interesting to consider such factorized distributions? The main reason to be excited is that this is a potentially very compact representation: the global joint distribution over all of the p variables is a product of "local" factors, each of which only depends on a node and its parents.

In the sequel, unless otherwise specified, we will focus on discrete random variables, where each $X_s \in \mathcal{X}_s$, for some discrete set $\mathcal{X}_s$ of size r_s. For simplicitly, suppose $r_s = r$, and $\mathcal{X}_s = [r]$. To represent a generic joint distribution $P(X)$ via a probability table, as noted earlier, requires $r^p - 1$ values.

But suppose we know that the distribution factors according to the graph G. And suppose the graph G has:

$$\max_{s \in V} |\text{PA}_s| \leq d.$$

We can then see that each local factor $P(X_s | X_{\text{PA}_s})$ can be represented using just $r^{d+1} - 1$ values, for a total of $p(r^{d+1} - 1) = O(pr^d)$ values. Thus for bounded d, we go from exponential to linear dependence on the number of variables p, thus making it now amenable to specify joint distributions over a very large number of variables.

3.2 Undirected Graphical Models (UGMs)

Consider now an undirected graph (UG) $G = (V, E)$, where each node $s \in V := [p]$ is associated with a random variable X_s, and the edge set $E \subseteq V \times V$ consists of unoriented edges, but where we now allow for undirected cycles. For any node $i \in V$, the following graph-theoretic terms will be handy:

- neighbors: $\text{NEIGHB}_i \subseteq V$

- non-neighbors: $\text{NON-NEIGHB}_i \subseteq V$

With DGMs, we had a very natural notion of locality: a node and its parents. UGs have no notion of parents, only neighbors. But the notion of neighbor is only pairwise: involving two variables. We can broaden this scope by considering the notion of **cliques**: which are fully connected subgraphs. Let $\mathcal{C}$ denote the set of cliques in the graph G. Another convenient notion is that of a maximal clique: any clique which does not belong to a larger clique. We can then see that cliques, and maximal cliques, provide a natural notion of locality in UGs.

As with DGMs, any UG G then corresponds to a *set of distributions* that "respect" the graph. Specifically, the joint distribution P over $X = (X_1, \ldots, X_p)$ is said to **factor according to the graph** G if it can be written as:

$$P_{\text{fact}}(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Phi_C(x_C), \tag{2}$$

where $\mathcal{C}$ is the set of cliques in G, and where $Z = \sum_x \prod_{C \in \mathcal{C}} \Phi(x_C)$, is the normalization constant. The factors $\Phi_C : \prod_{s \in C} \mathcal{X}_s \mapsto \mathbb{R}_+$ are non-negative functions over the local set of variables belonging to the clique $C \subseteq V$. Such models have a long history in statistical physics, and thus much of the terminology in UGMs is borrowed at least in part from statistical physics. The factors $\{\Phi(x_C)\}$ are also called clique "potentials", and the normalization constant Z is called the partition function.

Unlike with DGMs, here it is very clear that the UGM factorization yields a valid joint distribution.

Proposition 2 *The function $P_{fact}(\cdot)$ specified in Eqn.* (1) *is a valid joint distribution over X whenever the clique potential functions are non-negative, and non-trivial (i.e. not identically zero).*

Here, too this results in a huge reduction in representation cost. Suppose:

$$\max_{C \in \mathcal{C}} |C| \leq d.$$

Then just as with DGMs, each local factor $\Psi_C(x_C)$ can be represented using atmost r^d values, for a total of $p\, r^d$ values. Thus for bounded d, we go from exponential to linear dependence on the number of variables p, thus making it now amenable to specify joint distributions over a very large number of variables.

3.3 But what do potentials mean?

It is tempting to think of the potentials as being related to popular probabilistic quantities, such as marginals or conditionals. Suppose $P(X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \Phi_C(X_C)$. Then, would the

potentials $\Phi_C(X_C)$ be proportional to the marginal $P(X_C)$? In general, no. This is because the marginal gets contributions from multiple factors C' s.t. $C' \cap C \neq \varnothing$. Would the potentials $\Phi_C(X_C)$ be proportional to the conditional $P(X_C \mid X_{-C})$? Here again, in general no. This is because $P(X_C \mid X_{-C}) = P(X)/P(X_{-C}) = \Phi_C(X_C) \prod_{C' \neq C} \Phi_{C'}(X_{C'})/P(X_{-C})$. But as with the earlier argument, $P(X_{-C})$ will be have contribution from $\Phi_C(X_C)$ as well so long as $C \cap C' \neq \varnothing$ for some other $C' \in \mathcal{C}$.

So how can we interpret the clique potentials? We could assign them a "pre-probabilistic" notion of specifying "local" information regarding the overall probability. Thus, if we wish to encourage X_s and X_t to take similar values, we could do so via potentials $\Phi(X_s, X_t)$ that do so, without necessarily worrying about whether Φ has a specific probabilistic interpretation in terms of the overall distribution P. Another interpretation, when the clique potentials are all positive, is via the notion of "energy of the distribution". For such strictly positive clique potentials, we can define the clique energy function:

$$\phi_C(X_C) = -\ln \Phi_C(X_C),$$

where $\phi : \prod_{s \in C} \mathcal{X}_s \mapsto \mathbb{R}$ can be an arbitrary real-valued function, and need no longer be strictly positive. The sum of the clique energies:

$$E_G(X) = \sum_{C \in \mathcal{C}} \phi(x_C),$$

is also called the energy of the distribution. We can then write the overall distribution as:

$$P_{\text{fact}}(x) = \exp(-E_G(X) - \log Z).$$

This terminology is because the probability of a physical system depends inversely on the "energy" of a physical system. The overall distribution then is called a Gibbs distribution, which is thus simply a strictly positive UGM. Thus, with this setup, each clique energy function specifies the local contribution — from the clique C — to the overall energy.

4 The catch: constraints satisfied by PGMs

We thus saw that PGMs have a compact representation for the joint distribution in terms of local factors. But this compact representation has to come with a catch: the resulting distribution has to satisfy some constraints. But what exactly are these constraints? The factorization of the PGMs followed from their corresponding graph structure. As we will see, the resulting constraints can also be read off *just from the underlying graph structure*. The specific constraints will be called *Markov properties* of the corresponding PGM, and which we will be studying in some detail for the two classes, DGMs, and UGMs, separately. But at a high level, the absence of an edge indicates that the resulting variables are conditionally independent given an appropriate subset of the rest of the graph. In other words, the presence

of an edge connotes some notion of "direct dependence" that cannot be "explained away" by the rest of the graph. We will make these notions clearer when we study the Markov properties of PGMs in the sequel. But before doing so, it is worthwhile to consider alternative approaches by which we might characterize such direct dependences among variables.

5 Measures of Association/Dependence

Given a random vector $X = (X_1, \ldots, X_p)$, how do we characterize the "direct dependence" between any pair of variables X_s and X_t?

5.1 Marginal Dependence

The simplest approach is to consider "marginal" dependence i.e. look at the marginal distribution of (X_s, X_t) to assess the dependence.

The most popular measure in this context is Pearson's Correlation:

$$\rho(X_s, X_t) = \frac{\text{Cov}(X_s, X_t)}{\sqrt{\text{Var}(X_s)}\sqrt{\text{Var}(X_t)}}.$$

Some key properties:

- Captures linear dependency: linear regression coefficient of X_t on X_s is $\frac{\text{Cov}(X_s, X_t)}{\sqrt{\text{Var}(X_s)}}$
- $X_s \perp\!\!\!\perp X_t \implies \rho(X_s, X_t) = 0$
- But the converse is not true: $\rho(X_s, X_t) = 0$ does not imply $X_s \perp\!\!\!\perp X_t$ (since correlation focuses on linear dependency)
- Converse holds when (X_s, X_t) are bivariate Gaussian.

Recall the definition of marginal independence: we say that X_s is marginally independent of $X_t$iff:

$$P_{X_s X_t} = P_{X_s} P_{X_t}.$$

We could thus measure the "distance" between LHS and RHS to measure the degree of independence. A classical measure is mutual information:

$$I(X_s, X_t) = KL(P_{X_s X_t}, P_{X_s} P_{X_t}).$$

It is known that:

$$I(X_s, X_t) = 0 \text{ iff } X_s \perp\!\!\!\perp X_t.$$

One caveat is that it is difficult to estimate just given samples. It is thus useful to keep in mind that we could consider any of the large class of "divergence dependence" measures. Let D be any divergence between probability distributions such that $D(P, Q) = 0$ iff $P = Q$ a.e. We could then use this to define a dependence measure D (overloading notation here) via:

$$D(X_s, X_t) = D(P_{X_s X_t}, P_{X_s} P_{X_t}),$$

so that it follows that:

$$D(X_s, X_t) = 0 \text{ iff } X_s \perp\!\!\!\perp X_t.$$

Examples of useful divergences include f-divergences, which are generalizations of KL divergence above, and the class of "integral probability metrics". A key advantage of these is that they can be estimated efficiently from samples of X_s and X_t.

5.1.1 IPMs

Suppose B is a symmetric, convex set of measurable functions. Then, we can define the following dual norm:

$$\|\alpha\|_B = \max_{f \in B} \int_{\mathcal{X}} f(x) d\alpha(x),$$

which can then be used to define a metric over measures as

$$D(\alpha, \beta) = \|\alpha - \beta\|_B.$$

The total variation distance is an instance of this class, with

$$B = \{f \in \mathcal{C}(\mathcal{X}) \,:\, \|f\|_\infty \leq 1\}.$$

Another popular class is the so-called MMD distance. Given an RKHS $\mathcal{H}$, a natural class of functions to specify the dual norm is simply:

$$B = \{f \in \mathcal{H} \,:\, \|f\|_{\mathcal{H}} \leq 1\}.$$

Given the kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ corresponding to the RKHS $\mathcal{H}$, this can be simplified to:

$$\|\alpha\|_{\mathcal{H}_k} = \int_{\mathcal{X} \times \mathcal{X}} k(x, y) d\alpha(x) d\alpha(y),$$

which have been called "Maximum Mean Discrepancy" or "kernel norms". The above expression can also be written more compactly as:

$$\|\alpha\|_{\mathcal{H}_k} = \mathbb{E}_{X, X' \sim \alpha}[k(X, X')].$$

Given n samples $\{x_i\}_{i=1}^n$ drawn iid from α, and m samples $\{y_j\}_{j=1}^m$ drawn iid from β, let $\widehat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and $\widehat{\beta}_n = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$. Then, a natural estimator of $D(\alpha, \beta)$ is simply $D(\widehat{\alpha}_n, \widehat{\beta}_m)$.

However, for TV distance, and indeed most f-divergences, $D(\widehat{\alpha}_n, \widehat{\beta}_m)$ does not converge to $D(\alpha, \beta)$. For instance, $\|\widehat{\alpha}_n - \widehat{\beta}_n\|_{TV} = 2$ with probability 1 since the supports of the two discrete measures will likely not overlap. One needs to devise careful smoothing of these empirical measures, which forms a large body of work on non-parametric and parametric estimation of distributions. Even if they converge, the rate might scale exponentially with the dimension for instance as $O(n^{-1/d})$. A key attraction of MMD distances is the interesting result that the rate does not depend on the ambient dimension (Sriperumbudur et al., 2012):

$$\mathbb{E}[\|\widehat{\alpha}_n - \widehat{\beta}_n\|_{\mathcal{H}_k}] - \|\alpha - \beta\|_{\mathcal{H}_k} = O(n^{-1/2}).$$

5.2 Conditional Dependence

The key disadvantage of marginal notions of association is that it is not "direct" and in particular you will typically see more such associations than you might expect. As an example, suppose X_s = height of a person, and X_t = how well they do on say the SAT. The marginal dependence will be very high, but that is likely because toddlers and infants tend not to do well on the SAT. The dependence will likely vanish if we condition on X_u = the age of the student, so that $X_s \perp\!\!\!\perp X_t \,|\, X_u$.

To address this we thus need to consider conditional notions of dependence. The most popular notion is that of linear conditional dependence, namely partial correlation. The partial correlation of X_s and X_t given X_u measures the correlation of the residuals of X_s and X_t after eliminating the linear effect of X_u:

$$\rho(X_s, X_t \,|\, X_u) = \rho(e_{su}, e_{tu}),$$

where

$$e_{su} = X_s - (\beta_{su}^T X_u + b_{su})$$
$$e_{tu} = X_t - (\beta_{tu}^T X_u + b_{tu}).$$

Similar to marginal Pearson correlation, we have that: Some key properties:

- Captures linear conditional dependency
- $X_s \perp\!\!\!\perp X_t \,|\, X_u \implies \rho(X_s, X_t \,|\, X_u) = 0$
- But the converse is not true: $\rho(X_s, X_t \,|\, X_u) = 0$ does not imply $X_s \perp\!\!\!\perp X_t \,|\, X_u$ (since partial correlation focuses on linear conditional dependency)
- Converse holds if (X_s, X_t, X_u) are multivariate Gaussian.

Partial Correlation Graph We can construct a graph $G = (V, E)$ associated with $X = (X_1, \ldots, X_p)$ by associated lack of edges with zero partial correlation:

$$(u, v) \in E \implies \rho(X_u, X_v \mid X_{-uv}).$$

As we will see that this exactly corresponds to a UGM, also called a Gaussian graphical model, when X is multivariate Gaussian. But when X is not multivariate Gaussian, this need not capture true conditional dependences.

General Conditional Independence. To capture general conditional dependence, we could again work off the definition of conditional independence.

We say that X_s is conditionally independent of X_t given X_u iff:

$$P_{X_s X_t \mid X_u} = P_{X_s \mid X_u} P_{X_t \mid X_u}.$$

We could thus measure the "distance" between LHS and RHS to measure the degree of independence. Given any divergence measure D, we could thus define:

$$D(X_s, X_t \mid X_u) = D(P_{X_s X_t \mid X_u}, P_{X_s \mid X_u} P_{X_t \mid X_u}).$$

PGMs use graphs where the lack of an edge connotes some such conditional independence. The exact connotation of an edge depends on whether we are considering undirected, directed or chain (which involve both directed and undirected edges) graphical models.