Probabilistic Graphical Models

IKHAN CHOI

CONTENTS

1. MODELS

1.1. Statistical model.

Definition 1.1. A *statistical model* is an approximation scheme for unknown joint probability distribution.

The general purpose of many statistical models is to *estimate the joint probability distribution* of several random variables. The joint probability distrbution contains data about relations of the random variables. For example, suppose our goal is to obtain the most possible value of Y when given $X = x$, and we have already estimated the joint distribution function $f_{X,Y}$. Then, since the function $y \mapsto f_{X,Y}(x, y)$ describes the distribution of the random variable $Y|X = x$, what we want to find can be defined reasonably as

$$\widehat{y} = \arg \max_{y} f_{X,Y}(x, y).$$

Example 1.1. A random field, which we have not defined yet, is a way to represent several random variables together with their dependecies. Therefore, a paramterized random field gives a statistical model. In this case, training means an approximating process to find the best parameter, usually written as β.

Last Update: August 12, 2019.

1.2. **Random fields.**

Definition 1.2 (Random field)**.** A *random field* is a set of random variables parametrized by a topological space or a (directed or undirected) graph.

Definition 1.3. In this note, a term *random field* or *network* will be used to refer to random fields on a graph.

Example 1.2 (Markov chain)**.** Define a graph $G = (V, E)$ and $\mathcal{S}$ such that:
$$V = \mathbb{Z}_{t \geq 0},$$
$$E = \{(t, t+1)\}_{t \in V},$$
$$\mathcal{S} = \text{a finite set.}$$
An element $t \in V$ denotes the time t. Then, the set of $\mathcal{S}$-valued random variables $\{X_t\}_{t \in V}$ indexed by V defines a random field.

   The Markov property is given by
$$X_t \perp\!\!\!\perp X_s \mid X_{t-1}$$
for $s \leq t$. Since $\mathcal{S}$ is finite, alternatively we may rewrite it by
$$\Pr(X_t = x_t \mid X_{t-1} = x_{t-1}) = \Pr(X_t = x_t \mid X_{t-1} = x_{t-1}, \cdots, X_0 = x_0).$$

Example 1.3 (Maxwell-Boltzman distribution)**.** Let X be a $\mathcal{S}$-valued random field on a graph $G = (V, E)$ such that:
$$V = \{n\}_{n=1}^N,$$
$$E = \varnothing,$$
$$\mathcal{S} = \mathbb{Z}_x^3 \times \mathbb{Z}_p^3$$
for a large natural number N. The set V is the set of ideal gas particles, and the space of states $\mathcal{S}$ embodies the discretized phase space. At each particle $n \in V$ is attached an $\mathcal{S}$-valued random variable denoted by X_n. Our ultimate goal is to describe the joint probability distribution of X_n's; equivalently, the distribution of a $\mathcal{S}^V$-valued random variable $X = (X_1, \cdots, X_N)$. Elements of $\mathcal{S}^V$ are called *microstates*.

   Let $m > 0$ be a constant and fix a parameter $\beta = \frac{1}{k_B T} > 0$ which is called *coldness*. Define the *Boltzmann factor* as a function $\phi_n : \mathbb{R}_{\beta > 0} \times \mathcal{S} \to \mathbb{R}$ such that
$$\phi_n(\beta, i) := e^{-\beta E(i)},$$
where the *energy function* $E : \mathcal{S} \to \mathbb{R}$ is defined by
$$E(x, p) := \frac{\|p\|^2}{2m}.$$
The assumption for Boltzmann factors states that given β, the probability for a particle to be in the state i is proportional to the Boltzmann factor:
$$\Pr(X_n = i_n) \propto \phi_n(\beta, i_n) = e^{-\beta E_{i_n}}$$
for each state $i_n \in \mathcal{S}$ and a fixed particle $n \in V$. Thus, we can write
$$\Pr(X_n = i_n) = \frac{\phi_n(\beta, i_n)}{\sum_{j_n \in \mathcal{S}} \phi_n(\beta, j_n)} =: \frac{\phi_n(\beta, i_n)}{Z_n(\beta)}.$$

If we assume the independence of X_n's, then we get the disjoint probability distribution

$$\Pr(X = i) = \frac{\phi(\beta, i)}{\sum_{j \in \mathcal{S}^V} \phi(\beta, j)} =: \frac{\phi(\beta, i)}{Z(\beta)},$$

where

$$\phi(\beta, i) := \prod_{n=1}^{N} \phi_n(\beta, i_n), \qquad Z(\beta) = \prod_{n=1}^{N} Z_n(\beta).$$

The denominator $Z : \mathbb{R}_{\beta > 0} \to \mathbb{R}^+$ is called the *partition function*.

Example 1.4 (Ising model)**.** Let X be a $\mathcal{S}$-valued random field on a graph $G = (V, E)$ such that:

$$V = \mathbb{Z}^2,$$
$$E = \{\, (x, y) : \|x - y\| = 1 \,\},$$
$$\mathcal{S} = \{\pm 1\}.$$

1.3. **Bayesian networks.**

Definition 1.4 (Bayesian network)**.** Let G be a directed acyclic graph.

The graph acts as a parameter space. We want to investigate mutual effects among the paramtrized random variables.

Theorem 1.5 (Factorization of probablity)**.**

Example 1.6 (NBC, Naive Bayesian Classifier)**.**

Example 1.7 (HMM, Hidden Markov Model)**.**

1.4. **Markov networks.**

Definition 1.5 (Markov network)**.**

Markov networks are sometimes called MRF, Markov random field.

Example 1.8 (CRF, Conditional Random Field)**.** Consider a network with a graph G such that vertices are divided into two classes.

1.5. **Neural networks.** Probabilistic graphical models provide effective explanations of the neural networks, but neural networks are not confined only to graphical models.

Definition 1.6 (Neural network)**.** *Neural network* cannot be defined mathematically. It indicates statistical models that can solve problems with a collection of artificial neurons by adjusting connection strength among them.

Example 1.9 (MLP, Multi-layer Perceptron)**.**

Example 1.10 (RNN, Recurrent Neural Network)**.**

2. Inference

2.1. **Viterbi algorithm.**

3. Learning

3.1. **Gradient descent method.**

3.2. **Back propagation.**

Backpropagation refers to algorithms to train the weight matrices for minimizing the cost function J, which does not depend explicitly on any variables except the last layer vector $a^{(n)}$. However, since J is a function of the weight matrices implicitly, via $a^{(n)}$, we may find the representation of the gradiant of J as viewing it as a function on the space of weight matrices of each given layer. In other words, we want to find the coefficients of the differential form dJ on the basis $\{dW_{ij}^{(n-1)}\}_{i,j}$, $\{dW_{jk}^{(n-2)}\}_{j,k}$, or $\{dW_{kl}^{(n-3)}\}_{k,l}$, and so on.

Recall the definitions:

$$a_i^{(n)} = \sigma\left(\sum_j W_{ij}^{(n-1)} a_j^{(n-1)}\right).$$

Since the derivative of the sigmoid function is given by $\sigma' = \sigma - \sigma^2$, we can compute the following auxiliary relations

$$\frac{\partial a_i^{(n)}}{\partial a_j^{(n-1)}} = h(a_i^{(n)})W_{ij}^{(n-1)} \quad and \quad \frac{\partial a_i^{(n)}}{\partial W_{i'j}^{(n-1)}} = \delta_{ii'} h(a_i^{(n)})a_j^{(n-1)},$$

where $h(x) = x - x^2$.

Then, we can compute

$$dJ = \sum_i \frac{\partial J}{\partial a_i^{(n)}} \sum_j \frac{\partial a_i^{(n)}}{\partial W_{ij}^{(n-1)}} \, dW_{ij}^{(n-1)} = \sum_{i,j} \frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)})a_j^{(n-1)} \, dW_{ij}^{(n-1)},$$

which implies

$$\nabla J(W^{(n-1)}) = \left[\frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)})a_j^{(n-1)}\right] \frac{\partial}{\partial W_{ij}^{(n-1)}}.$$

Note that it is a function of a_i and a_j. The gradient descent method will take

$$W_{ij}^{(n-1)+} := W_{ij}^{(n-1)} - \alpha \cdot \frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)})a_j^{(n-1)}$$

with a proper parameter $\alpha > 0$.

By the same reason,

$$dJ = \sum_{i,j,k} \frac{\partial J}{\partial a_i^{(n)}} \frac{\partial a_i^{(n)}}{\partial a_j^{(n-1)}} \frac{\partial a_j^{(n-1)}}{\partial W_{jk}^{(n-2)}} \, dW_{jk}^{(n-2)}$$

$$= \sum_{i,j,k} \frac{\partial J}{\partial a_i^{(n)}} \cdot h(a_i^{(n)})W_{ij}^{(n-1)} \cdot h(a_j^{(n-1)})a_k^{(n-2)} \, dW_{jk}^{(n-2)},$$

which implies

$$\nabla J(W^{(n-2)}) = \left[\sum_i \frac{\partial J}{\partial a_i^{(n)}} \cdot h(a_i^{(n)})W_{ij}^{(n-1)} \cdot h(a_j^{(n-1)})a_k^{(n-2)}\right] \frac{\partial}{\partial W_{jk}^{(n-2)}}.$$

Therefore, the gradient descent method will take

$$W_{jk}^{(n-2)^+} := W_{jk}^{(n-2)} - \alpha \cdot \sum_i \frac{\partial J}{\partial a_i^{(n)}} h(a_i^{(n)}) W_{ij}^{(n-1)} h(a_j^{(n-1)}) a_k^{(n-2)}$$

$$= W_{jk}^{(n-2)} + (1 - a_j^{(n-1)}) a_k^{(n-2)} \sum_i (W_{ij}^{(n-1)^+} - W_{ij}^{(n-1)}) W_{ij}^{(n-1)}.$$

In similar way,

$$W_{kl}^{(n-3)^+} := W_{kl}^{(n-3)} + (1 - a_k^{(n-2)}) a_l^{(n-3)} \sum_i (W_{jk}^{(n-2)^+} - W_{jk}^{(n-2)}) W_{jk}^{(n-2)} (?)$$

3.3. **Maximum likelihood estimate.**

Definition 3.1. Let f be a distribution function on a measure space X. Let $\{f_\theta\}_\theta$ be a parametrized family of distrubution functions on X. The *likelihood* $L_n(\theta) : \Omega^n \to \mathbb{R}_{\geq 0}$ for a fixed parameter θ is a random variable defined by

$$L_n(\theta) := \prod_{i=1}^n f_\theta(x_i)$$

where $\{x_i\}_i$ is a family of i.i.d. X-valued random variables with a distriburion f.

The objective of the likelihood function is to find θ such that f_θ approximates the unknown distribution f. Write

$$\frac{1}{n} \log L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f_\theta(x_i).$$

By the law of large numbers, $\frac{1}{n} \log L_n(\theta)$ converges to a constant function

$$\mathbb{E}(\log f_\theta(x)) = \int_X f \log f_\theta$$

in measure as $n \to \infty$. This constant function is exactly what we call *cross entropy*.

The *Kullback-Leibler divergence* is a kind of asymmetric distance function defined from the difference with cross entropy

$$D_{KL}(f \| f_\theta) := \int_X f \log f - \int_X f \log f_\theta.$$

It is proved to be always nonnegative by the Jensen inequality:

$$\int_X f \log f_\theta - \int_X f \log f = \int_X f \log \frac{f_\theta}{f} \leq \log \left(\int_X f \frac{f_\theta}{f} \right) = 0.$$

Here, we exclude the region $f = 0$ from the integration region. Then, we can say, bigger $L_n(\theta)$ is, closer f_θ and f are.