Probabilistic Graphical Models

IKHAN CHOI

Contents

1. Introduction

1.1. Statistical model.

Definition 1.1. A *statistical model* is a set of statistical assumptions that provides with an approximation scheme for unknown probability distribution.

A typical way to define statistical models is to give a parametrized family of probability distributions of special form. A model is said to be *parametric* if the parameter space has finite dimensional. Although there are nonparametric models, we only deal with parametrized models.

In the family, we want to find a particular distribution, i.e. particular parameter, that is sufficiently approximated to the unknown ideal distribution. This parameter finding procedure from samples is called *(statistical) estimation*.

Example 1.1. Let X be a real-valued random variable with probability density f. Suppose we have a set of trials $\{X_i\}_{i\in\mathbb{N}}$ of a same experiment designed for finding f. We can define trials X_i as i.i.d. real-valued random vaiables with distribution f. Assuming a fate $\omega \in \Omega$ has been determined, the notation x_i then now denotes the sample data $X_i(\omega) = x_i$ obtained from the trial X_i. We want to find the approximated distribution for f from the sampled dataset $\{x_i\}_i$ after ω is realized.

Consider a statiscal model defined with normal distributions

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} = \Theta.$$

First Written : October 6, 2019.
Last Updated : October 6, 2019.

The dimension of the model is the dimension of the parameter space $\Theta = \mathbb{R} \times \mathbb{R}_{>0}$, hence 2. In other words, we set a statistical assumption that the ideal distribution would follow the normal distribution.

We can find the parameter θ by a classical method called maximum likelihood estimate(MLE). The *likelihood* $L_n : \Theta \times \Omega \to \mathbb{R}_{\geq 0}$ is defined by a parametrized random variable

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

Intuitionally, data x_i are distributed on regions where the original density function f is large. If we apply the law of large numbers on the random variables $\{\log f_\theta(X_i)\}_{i\in\mathbb{N}}$ for each fixed θ, then since $\frac{1}{n}\log L_n$ is the average of $\log f_\theta(X_i)$ for $1 \leq i \leq n$, we get a convergence to a constant random variable, also called *cross entropy*:

$$\frac{1}{n}\log L_n(\theta) \overset{\text{in measure}}{\to} \mathbb{E}(\log f_\theta(X)) = \int_{\mathbb{R}} f \log f_\theta \quad \text{as} \quad n \to \infty.$$

Define *Kullback-Leibler divergence*, a kind of asymmetric distance function, by the difference of cross entropies

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

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

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

Then, we can say, bigger $L_n(\theta)$ is, closer f_θ and f are. Therefore, if ω is given, we can find the most reasonable parameter by solving the following optimization problem:

$$\widehat{\theta} = \arg\max_\theta L_n(\theta)(\omega) = \arg\max_\theta \prod_{i=1}^{n} f_\theta(x_i).$$

Unlike the above example, many statistical models are suggested to estimate *joint probability distribution* of several random variables. The joint probability distrbution contains data about correlations among the random variables. For example, suppose that we are asked 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

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

Example 1.2. 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, which will be usually written as θ or β.

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.

Be cautious that the following examples are not statistical models because it is not designed for estimation from sampled data.

Example 1.3 (Markov chain)**.** Define a graph $G = (V, E)$ and a set S such that:

$$\begin{aligned} V &= \mathbb{Z}_{t\geq 0}, \\ E &= \{(t, t+1)\}_{t\in V}, \\ S &= \text{a finite set of states.} \end{aligned}$$

An element $t \in V$ denotes the time t. Then, the set of 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 S is finite, alternatively we may write it by

$$\Pr(X_t = x_t \mid X_{t-1} = x_{t-1}) = \Pr(X_t = x_t \mid X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)$$

for all $x_i \in S$ with $i = 0, \cdots, t$. With abusing notation, it is often written as

$$p(x_t|x_{t-1}) = p(x_t|x_{t-1}, \cdots, x_0).$$

Example 1.4 (Maxwell-Boltzmann statistics for ideal gas)**.** The goal is to describe the energy distributions of noninteracting ideal gas particles that are parametrized by reciprocal temperature $\beta > 0$. Fix β.

Let us give definitions. Let X be a S-valued random field on a graph $G = (V, E)$ such that:

$$\begin{aligned} V &= \{j\}_{j=1}^{N}, \\ E &= \varnothing, \\ S &= \mathbb{R}^3_x \times \mathbb{R}^3_p, \quad \text{phase space} \end{aligned}$$

for a large natural number N. Define the energy function $H : S \to \mathbb{R}$ such that

$$H(x, p) := \frac{\|p\|^2}{2m} \quad (m : \text{ mass of particle}),$$

and parametrized family of functions $\phi : S \to \mathbb{R}$ such that

$$\phi(s_j) := e^{-\beta H(s_j)}.$$

The functions ϕ are called *Boltzmann factors*.

In this model, the set V is the set of ideal gas particles. At each particle $j \in V$ is attached an S-valued random variable X_j with distribution $f_j : S \to \mathbb{R}_{\geq 0}$. Our primary goal is to describe the joint probability distribution of X_j's; equivalently, the distribution f of a S^N-valued random variable $X = (X_1, \cdots, X_N)$. Elements of S^N are called *microstates*.

First we give the probability distribution of an individual random variable X_j. The assumption for Boltzmann factors states that the probability for a particle j to be in a state $s_j \in S$ is proportional to the Boltzmann factor:

$$f_j(s_j) \propto_\beta \phi(s_j) = e^{-\beta H(s_j)}$$

for each state $s_j \in S$ and a fixed particle j. Thus, we can write

$$f_j(s_j) = \frac{\phi(s_j)}{\int_S \phi}.$$

Next, consider the entire random field $X : \Omega \to S^N$. If we assume the independency of X_j's, then we get the disjoint probability distribution

$$f(s) = \frac{\phi(s)}{\int_{S^N} \phi} =: \frac{\phi(s)}{Z(\beta)},$$

where the definitions of $\phi, H : S^N \to \mathbb{R}$ are extended such that

$$\phi(s) = \prod_{j=1}^{N} \phi(s_j), \quad H(s) = \sum_{j=1}^{N} H(s_j).$$

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

Finally, we will compute the distribution of $H \circ X : \Omega \to \mathbb{R}$.

1.3. **Notations.** We review the definition of probability distribution. In this subsection, let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and $(S, \mathcal{S})$ a measurable space.

Definition 1.4. Let $X : \Omega \to S$ be a random variable. The *probability distribution* of X is the pushforward measure $X_* \Pr$ on S defined as $X_* \Pr(A) = \Pr(X^{-1}(A))$ for measurable $A \in \mathcal{S}$. We often also write $\Pr(X \in A)$ for $X_* \Pr(A)$.

Definition 1.5. Let S be a Lebesgue measurable subset of a Euclidean space. Let $X : \Omega \to S$ be a random variable. If the probability distribution of X is absolutely continuous with respect to the Lebesgue measure, then we have the Radon-Nikodym derivative. We call the derivative *probability distribution function* or shortly *pdf* of X.

Definition 1.6. Furthermore, let S be finite or countable with discrete measurable structure. Let $X : \Omega \to S$ be a random variable. We call a function $p(s) = \Pr(X = s)$ the *probability mass function* or shortly *pmf* of X.

We will not consider a random variable that has neither pdf nor pmf. If we do not restrict the regularity of distributions, we must resolve intractable technical issues. Also note that pdf is more general notion than pmf.

We mainly deal with several random variables and their joint distribution. The notations then become dirty and equations get longer due to the number of random variables, so we introduce alternative notations. Suppose $X = \{X_n\}_{n=1}^N$ is a set of random variables and let $(S_n, \mathcal{S}_n)$ be the codomain of X_n. The pdf and pmf of an individual random variable X_n will be denoted by $f(x_n) : S \to \mathbb{R}_{\geq 0}$ and $p(x_n) : S \to [0, 1]$ using lower case alphabets for the random variables. Moreover, we will admit the abuse of notations more generally without precise definitions, but for examples, just note that

- $f(x_1)$ is a pdf on S_1,
- $f(x_1, x_2) = f(x_{\{1,2\}})$ is a pdf on $S_1 \times S_2$,
- $f(x)$ is a pdf on $S = \prod_{n=1}^{N} S_n$ (joint distribution),
- $f(x_2|x_1) : S_2 \times S_1 \to \mathbb{R}_{\geq 0}$ is a pdf on S_2 parametrized by S_1 (or more generally by the σ-algebra $\mathcal{S}_1$),
- $f(x_A|x_B)$ is a pdf on $\prod_{n\in A} S_n$ parametrized by $\prod_{n\in B} S_n$.

Every notation above in fact can be recognized as a real-valued function on S via projections. For instance, the function $f(x_1) : S_1 \to \mathbb{R}_{\geq 0}$ can have extended domain like $f(x_1) : S \twoheadrightarrow S_1 \to \mathbb{R}_{\geq 0}$ that is constant for all variables except X_1.

According to this notation, we can say that our interest is always the joint distribution $f(x)$ of a random field.

2. Bayesian networks

2.1. **Factorization of probability.** Recall the definition of descendants/ancestors and children/parents of nodes.

Definition 2.1 (Bayesian network)**.** Let X be a random field on a directed acyclic graph G. We say X is a *Bayesian network* or satisfies the *local Markov property* if

$$X_v \perp\!\!\!\perp X_{V\setminus de(v)} \mid X_{pa(v)}.$$

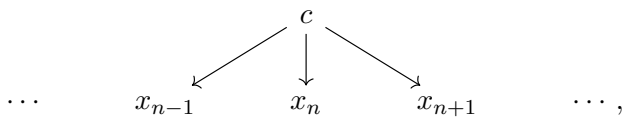
Definition 2.2 (Factorization)**.** Let X be a random field on a directed acyclic graph G. We say X *factorizes over* G if

$$f(x) = \prod_{v\in V} f(x_v|x_{pa(v)}).$$

Theorem 2.1. *A random field X on a directed acyclic graph G is a bayesian network if and only if it factorizes over G.*

Proof. □

Example 2.1 (NBC, Naive Bayesian Classifier)**.** Consider a random field on the following directed acyclic graph(DAG):



where $n = 1, \cdots, N$. The joint distribution is

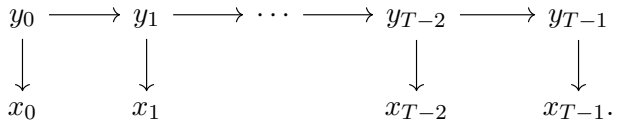
$$p(c, x) = p(c) \prod_{n=1}^{N} p(x_n|c).$$

It is represented by factors $p(c)$ and $p(x_n|c)$'s; the conditional probability $p(c|x) = p(c, x)/p(x)$ can be computed if we assume that the input feature x has been given and the probability $p(c)$, $p(x_n|c)$ had been estimated.

If we view $p(c) \in [0, 1]^{S_c}$ and $p(x_n|c) \in [0, 1]^{S_x\times S_c}$ as *parameters* which parametrize the joint distribution $p(c, x) \in [0, 1]^{S_c\times S_x^N}$, then we can say the NBC defines a family of distribution functions for approximating $p(x, c)$ that is parametrized by $(|S_c| - 1 +$

$N(|S_x|-1)|S_c|$)-dimensional parameter, which is much smaller than $|S_c||S_x|^N - 1$. The NBC provides with a statistical model via "conditional parametrization".

Example 2.2 (HMM, Hidden Markov Model)**.** Consider a random field on the following DAG:

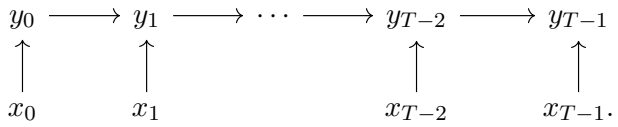


The joint distribution is

$$p(y,x) = p(x_0|y_0)p(y_0)\prod_{t=1}^{T-1} p(x_t|y_t)p(y_t|y_{t-1}).$$

The terms $p(y_0)$, $p(y_t|y_{t-1})$, and $p(x_t|y_t)$ are called *start probability*, *transition probability*, and *emmision probability* respectively, and they play a role of parameters.

Example 2.3 (MEMM, Maximaum entropy Markov Model)**.** Consider a random field on the following DAG:



The joint distribution is

$$p(y,x) = p(x_0)p(y_0|x_0)\prod_{t=1}^{T-1} p(x_t)p(y_t|y_{t-1},x_t).$$

The MEMM introduces exponential modeling

$$p(y_i|y_{i-1},x_i) \propto_{y_{i-1},x,\beta} e^{-\beta\cdot H(y_i,y_{i-1},x_i)}$$

with a appropriately designed energy function $H : S_y^2 \times S_x \to \mathbb{R}^d$. The normalizing constant Z depends only on y_{i-1} and x_i, hence

$$p(y_i|y_{i-1},x_i) = \frac{e^{-\beta\cdot H(y_i,y_{i-1},x_i)}}{Z(y_{i-1},x_i;\beta)}.$$

Estimate of the joint distribution is done by adjusting the d-dimensional parameter β.

3. Markov networks

In this section, we discuss random fields on undirected graphs. As in Bayesian networks, we expect the whole joint probability to be factorized into a computable form with smaller dimensional parameters. In this factorization, a modified version of Markov property on the graph that is different from the classical Markov process is required.

It would seem that the initial and general definitions are too abstract that they are hardly applied in practical problems..... But...

Definition 3.1 (Markov properties on a graph)**.** There are three Markov properties:

(1) *Pairwise Markov property*

Definition 3.2 (Markov network)**.** Let X be a random field on a undirected graph G.

Markov networks are sometimes called MRF, Markov random field.

Example 3.1 (Ising model)**.**

Example 3.2 (CRF, Conditional Random Field)**.** Consider a network with a graph G such that vertices are divided into two classes.

$$p(y|x) \propto_{x,\beta} e^{-\beta \cdot H(y,x)}.$$

For the Viterbi algorithm, the energy function H is frequently taken to have the form

$$H(y,x) = \sum_i E(y_i, y_{i-1}, x).$$

4. Neural networks

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

Definition 4.1 (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 4.1 (MLP, Multi-layer Perceptron)**.**

Example 4.2 (RNN, Recurrent Neural Network)**.**

4.1. Maximum likelihood estimate.

Definition 4.2. 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*.