IFT 6269: Probabilistic Graphical Models **Fall 2017**

Lecture 3 — September 12

Lecturer: Simon Lacoste-Julien *Scribe: Philippe Brouillard and Tristan Deleu*

Disclaimer: These notes have only been lightly proofread.

3.1 Parametric Models

3.1.1 Family of distributions

A parametric model is a family of distributions that is defined by a fixed finite number of parameters.[1]

A family of distributions is formally defined as follows:

$$\mathcal{P}_\Theta = \{p_\theta(\cdot;\theta) \mid \theta \in \Theta\}$$

where $p_\theta(\cdot;\theta)$ is the possible pmf or pdf (understood from context) depending on the parameter θ and Θ is the set of all valid parameters.[2]

The support of distribution Ω_X is usually fixed for all θ. For example, the support of a distribution modelling a coin flip could be $\Omega_X = \{0, 1\}$. Similarly, for the gamma distribution, the support is $\Omega_X = [0, +\infty[$.

3.1.2 Notation

To indicate that a random variable is distributed as a known distribution, we use the symbol "$\sim$". For example, to indicate that the random variable X is distributed as a Bernoulli distribution of parameter θ, we would write:

$$X \sim \text{Bern}(\theta)$$

This notation is a shorthand for:

$$p(x;\theta) = \text{Bern}(x;\theta)$$

where $p(x;\theta)$ represents the pmf for X, and $\text{Bern}(x;\theta)$ indicates that we refer to the pmf (on x) for the Bernouilli distribution.

[1] We will see later in the class non-parametric models, which basically means that the number of parameters is (potentially) infinite. These models are usually fit from data with a number of (effective) parameters growing with the size of training data.

[2] Using $p_\theta(x;\theta)$ instead of $p_\theta(\cdot;\theta)$ would be an abuse of notation since $p_\theta(x;\theta)$ is only a scalar for a specific x and not a pmf/pdf.

To take another example, if the random variable X is distributed as a Normal distribution with parameters μ and σ^2, we would write:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

that is similar to say that it has a pdf (as now X is a continuous R.V.):

$$p(x; (\mu, \sigma^2)) = \mathcal{N}(x \mid \mu, \sigma^2)$$

3.1.3 The Bernoulli distribution

The pmf of a Bernoulli random variable X is given as follows:

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

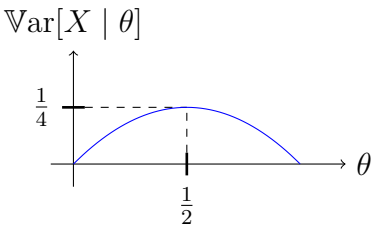
The support of the distribution is $\Omega_X = \{0, 1\}$ and the space of the parameters is $\Theta = [0, 1]$. From the pmf, we can see that $P\{X = 1 \mid \theta\} = \theta$.[3]

The expected value and the variance of a Bernoulli random variable X are:

$$\mathbb{E}[X] = \theta$$

$$\mathbb{V}\text{ar}[X] = \theta(1-\theta)$$

We can see from the figure below that the variance is at its highest point when $\theta = 1/2$.



Intuitively, the Bernoulli distribution models a situation where there are only two possible outcomes: either a success or a failure. The classical example is a coin flip: if getting a head is a success, X will equals 1. In this case, the parameter θ would be the probability to get a head.

3.1.4 The Binomial distribution

A binomial distribution $\text{Bin}(n, \theta)$ can be defined as the sum of n independent and identically distributed (i.i.d.) Bernoulli random variables with parameter θ. Formally:

[3] Note that instead of θ, p is also often used as a parameter for the Bernoulli and the Binomial distribution.

Let $X_i \overset{iid}{\sim} \text{Bern}(\theta)$ [4]

Let $X = \sum_{i=1}^{n} X_i$

then we have that $X \sim \text{Bin}(n, \theta)$

The support of the distribution is $\Omega_X = \{0, 1, ..., n\}$ and the space of the parameters is $\Theta = [0, 1]$.

The pmf is given as follows:

$$p(x; \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

The term $\binom{n}{x}$ is equal to the number of ways to get x successes out of n trials. Formally, this is defined as follow:

$$\binom{n}{x} \triangleq \frac{n!}{x!(n-x)!}$$

As for the term $\theta^x(1-\theta)^{n-x}$, we can notice that it is the product of the pmf of n Bernoulli random variables, since:

$$\theta^x(1-\theta)^{n-x} = \theta^{\Sigma x_i}(1-\theta)^{\Sigma(1-x_i)} = \prod_{i=1}^{n} \text{Bern}(x_i; \theta)$$

The expected value and the variance of a Binomial random variable X can be deduced from the Bernoulli's expected value and variance, since $X = \sum_{i=1}^{n} X_i$:

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] = n\theta$$

$$\mathbb{V}\text{ar}[X] \underset{\text{by indep.}}{=} \sum_i \mathbb{V}\text{ar}[X_i] = n\theta(1-\theta)$$

Intuitively, the Binomial distribution can be seen as a model for n independent coin flips.

3.1.5 Other distributions

- The Poisson distribution is often used to model count data: the pmf is $\text{Poisson}(x|\lambda)$, where λ is the mean parameter. $\Omega_X = \mathbb{N}$.
- The Gaussian distribution is the most common distribution for real numbers. The pdf is denoted $\mathcal{N}(x|\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance parameters. $\Omega_X = \mathbb{R}$ here.

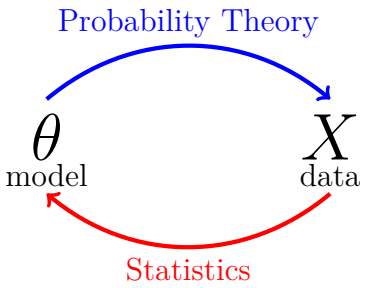
[4] Implicitly, X_i refers to $X_1, ..., X_n$

- The gamma distribution is often used to model positive numbers. The pdf is denoted $\text{Gamma}(x|\alpha, \beta)$, where α is the shape parameter and β is the rate parameter. $\Omega_X = \mathbb{R}_+$.

Here is a list of other common distributions (look them up on Wikipedia): Laplace, Cauchy, exponential, beta, Dirichlet, etc.

3.2 Statistical Concepts

Probability theory can be used as a way to infer or generate data from a model. This is a well defined problem. On the contrary, statistics is a way to infer a model based on observed data. This is an inverse problem that is unfortunately ill-defined.



To illustrate the difference between probability and statistics, suppose we have a model that can generate n independent coin flips. A classical probability theory problem would be to calculate the probability of k heads happening in a row. In this case, the model would be given without data. In the case of statistics, we would only have observed data (e.g. k heads on n trials) and the model wouldn't be accessible. A classical statistics problem would be to infer the parameters of a model that explains the observed data (e.g. what is the bias of the coin flip in this example).

3.2.1 The Frequentist and the Bayesian

As stated earlier, the statistics problem is ill-defined. Furthermore, even the meaning of a probability can differ from different philosophical point of views. Two major schools of thought using different meaning of probability have arisen: The Frequentist and the Bayesian.

1. The traditional Frequentist semantic is the following:

   $P(X = x)$ represents the limiting relative frequency of observing $X = x$ if we could repeat an infinite number of i.i.d. experiments.

2. The Bayesian semantic is the following:

   $P(X = x)$ encodes an agent "belief" that $X = x$.

The laws of probability characterize a "rational" way to combine "beliefs" and "evidence" (i.e. observations). This approach has many motivations in terms of gambling, utility, decision theory, etc.

3.2.2 The Frequentist interpretation of probability

To illustrate the view of the Frequentist interpretation, we will analyze an example. For a discrete random variable, suppose that $P\{X = x\} = \theta$ then $P\{X \neq x\} = 1 - \theta$.

Let $B \triangleq \mathbb{1}\{X = x\} \sim \text{Bern}(\theta)$, which encodes the event that X takes the value x.

Suppose we repeat the i.i.d. experiments a large number of times, i.e. $B_i \overset{iid}{\sim} \text{Bern}(\theta)$.

By the law of large numbers, we have that the empirical average i.i.d. R.V.'s will converge to its expected value:

$$\frac{1}{n}\sum_i B_i \underset{n\to\infty}{\to} \mathbb{E}[B_i] = \theta$$

We can also show that the empirical average will concentrates tightly around the value θ (by the central limit theorem).

Consider the R.V. which represents the sum, it has the distribution:

$$\sum_{i=1}^{n} B_i \sim \text{Bin}(n, \theta)$$

The expected value and the variance of the average are the following:

$$\frac{1}{n}\mathbb{E}[\sum_i B_i] = \frac{n\theta}{n} = \theta$$

$$\mathbb{V}\text{ar}[\frac{1}{n}\sum_i B_i] = \frac{1}{n^2}\mathbb{V}\text{ar}[\text{Bin}(n,\theta)] = \frac{1}{n^2}n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}.$$

We thus see that the variance of the empirical average goes to zero as $n \to \infty$, showing the concentration.

More precisely, we have by the central limit theorem that:

$$\sqrt{n}\left(\frac{1}{n}\text{Bin}(n,\theta) - \theta\right) \xrightarrow{d} \mathcal{N}(0, \theta(1-\theta)))$$

Notice the scaling of the difference by $\sqrt{n}$. For large n, the distribution of the empirical average is close to a Gaussian distribution with mean θ and variance $\theta(1-\theta)/n$.

3.2.3 The Bayesian Approach

The Bayesian approach is very simple philosophically: it treats all uncertain quantities as random variables.

In fact, it encodes all the knowledge about a system (the "beliefs") as "prior" on probabilistic models and then uses laws of probabilities (and Bayes rule) to get answers.

The simplest example to illustrate the Bayesian approach is the result of n coin flips of a biased coin. We believe that $X \sim \text{Bin}(n, \theta)$. Since θ is unknown, we model it as a random variable. Thus, we need a "prior distribution" $p(\theta)$ with a sample space defined as $\Omega_\Theta = [0, 1]$.

Suppose we observe $X = x$ (the result of n flips), then, we can "update" our belief about θ using Bayes rule: [5]

$$p(\theta \mid X = x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}$$

where,

$p(\theta \mid X = x)$ is the posterior belief,

$p(x \mid \theta)$ is the likelihood or the observation model,

$p(\theta)$ is the prior belief and

$p(x)$ is the normalization or "marginal likelihood"

To illustrate the bayesian approach, suppose that $p(\theta)$ is a uniform on $[0, 1]$, i.e. the prior doesn't encode specific preferences.

$$p(\theta \mid x) \propto \theta^x (1 - \theta)^{n-x} \mathbb{1}_{[0,1]}(\theta)$$

(where $x \in 0 : n$) The symbol "$\propto$" means that it is proportional to, i.e. we can drop any term that doesn't contain θ.

The scaling factor is:

$$\int_0^1 \theta^x (1 - \theta)^{n-x} \, d\theta = B(x + 1, n - x + 1)$$

The beta function is defined as:

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

The gamma function is defined as:

$$\Gamma(a) \triangleq \int_0^\infty u^{a-1} e^{-u} \, du$$

[5] Note that if $p(x \mid \theta)$ is a pmf and $p(\theta)$ is a pdf, then the *joint* $p(x, \theta)$ will be a mixed distribution.

$p(\theta \mid x) = \text{Beta}(\theta; x+1, n-x+1)$ is a beta distribution defined as:

$$\text{Beta}(\theta; \alpha, \beta) \triangleq \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \mathbb{1}_{[0,1]}(\theta)$$

As a Bayesian, the posterior distribution $p(\theta \mid X = x)$ contains all the information we need to predict the likelihood of an event.

For example, what is the probability that the next coin flip is $F = 1$? [6]

$$p(F = 1 \mid \theta) = \theta$$

By using the marginalization over θ, we get:

$$P(F = 1 \mid X = x) = \int_\theta P(F = 1, \theta \mid X = x)\, d\theta$$

By using the chain rule, we get:

$$= \int_\theta P(F = 1 \mid \theta, X = x) P(\theta \mid X = x)\, d\theta$$

Now we have $P(F = 1 \mid \theta, X = x) = \theta$ by the definition of our model, and thus we get:

$$\int_\theta \theta P(\theta \mid X = x)\, d\theta = \mathbb{E}_\theta[\theta \mid X = x]$$

Where the conditional expectation is called the posterior mean of θ.
A meaningful Bayesian estimator of θ is $\hat{\theta}_{Bayes}(x) \triangleq \mathbb{E}_\theta[\theta \mid X = x]$. [7]

Since $p(\theta \mid x)$ is a Beta and the expected value of a Beta is:

$$\mathbb{E}[Beta(\theta; \alpha, \beta)] \triangleq \frac{\alpha}{\alpha + \beta}$$

then the Bayes estimator is:

$$\mathbb{E}[\theta \mid X = x] = \frac{x+1}{n+2} = \hat{\theta}_{Bayes}(x)$$

If we compare it to the ML estimator from the Frequentist approach:

$$\hat{\theta}_{MLE}(x) = \frac{x}{n}$$

We can see that while the MLE is unbiased, the Bayesian estimator is biased, but asymptotically unbiased. Furthermore, the Bayesian estimator encodes an uncertainty: even if the data contains only head flips, the estimator gives a small probability to flip a tail. This, however, is not the case with the MLE estimator (which tends to overfit).

[6] By convention, a lowercase θ is used even if it's a random variable because Θ is already used for the parameter space.

[7] Notation: $\hat{\theta}$ is a statistical estimator of θ. Based on the observations, $\hat{\theta}$ is a value included in the valid set of parameters Θ. The Frequentist statistics consider multiple possible estimators: MAP, Bayesian posterior mean, MLE, moment matching. After selecting an estimator, we can analyze their statistical properties: bias, variance, consistency.