Joohyung Lee

2. Probability Distribution

2.0. Introduction

1) Chapter Scope
   A. Examples of probability distributions
   B. Their properties
2) Purpose of Introducing Distributions
   A. a building blocks for more complex models
   B. a recipe to discuss some essential statistical concept, e.g., Bayesian inference
   C. to model the probability distribution p(**x**), i.e., density estimation

      * Model Selection becomes an issue since density estimation is fundamentally ill-posed problem in that infinitely many distributions can fit the observed data set.
3) Parametric distribution vs. Non-Parametric distribution
   A. Parametric distribution
      i. binomial distribution, multinomial distribution, Gaussian distribution (continuous R.V.)
      ii. For <u>density estimation</u>, the parameters shall be determined with an <u>observed</u> data set.
         1. Frequentist: specific values for parameters (earned by optimizing some criterion, e.g., likelihood function)
         2. Bayesian: estimate posterior distribution with introduced prior distributions over the parameters as well as the observed data
      iii. Conjugate Priors: To simplify the Bayesian analysis, use conjugate prior which let posterior distribution be in the same form of prior distribution.
         1. Exponential family of distributions is presented as it possesses a number of important properties.
   B. Non-Parametric distribution
      i. Distribution form is not forced by a user but typically depends on the size of the data set
      ii. Still has the parameters but they do not determine the distribution form but the complexity
      iii. Histogram, nearest-neighbors, kernels

Table. 1 Conjugate prior with posterior distribution in exponential family

Conjugate Prior	**Posterior Distribution**
Dirichlet distribution	Multinomial distribution
Gaussian distribution	Gaussian distribution

2.1. Binary Variables

1) Bernoulli distribution

A. Definition

$$\mathrm{Bern}(\mathrm{x}|\mu) = \mu^x(1-\mu)^{1-x}, where\ \ 0 \le \mu < 1\ \ and\ \ x \in \{0,1\} \tag{2.1}$$

B. Properties

$$\mathbb{E}[\mathrm{x}] = \mu \tag{2.2}$$

$$var[x] = \mu(1-\mu) \tag{2.3}$$

C. Density estimation

$$\mathcal{D} = \{x_1, \dots, x_N\}$$

i. Frequentist

1. Estimate μ by maximizing the likelihood function, i.e., maximize the log of likelihood

$$\ln\big(p(\boldsymbol{\mathcal{D}}|\mu)\big) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \ln p(x_n|\mu) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \{x_n \ln\mu + (1-\mu)\ln(1-\mu)\} \tag{2.4}$$

2. The above log likelihood function depends on the N observations only through their sum, i.e., *sufficient statistics*: $\sum_n x_n$.
3. $\mu_{ML} = \frac{m}{N}$ = *sample mean*

ii. Bayesian

1. Flip a coin 3 times resulting all heads → what is the reasonable prediction? (overfitting)

2) Binomial distribution

A. Definition

$$\mathrm{Bin}(\mathrm{m}|\mathrm{N},\mu) = \binom{N}{m}\mu^x(1-\mu)^{1-x} \tag{2.5}$$

B. Properties

For independent events, 1) the means of the sum is the sum of the mean and 2) the variance of the sum is the sum of the variance

$$\mathbb{E}[\mathrm{m}] = N\mu \tag{2.6}$$

$$\mathrm{var}[\mathrm{m}] = N\mu(1-\mu) \tag{2.7}$$

1.1 The beta distribution (conjugate prior for the binomial distribution)

1) Motivation for the conjugate prior distribution

A. Prior distribution is required in order to develop a Bayesian treatment.

B. Make posterior distribution have the same functional form as the prior (conjugacy).

2) Definition

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}, \text{gamma coefficient for the normalization purpose}$$

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.8}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.9}$$

3) Gamma function

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}\,du \tag{2.10}$$

$$\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1, \Gamma(x+1) = x! \tag{2.11}$$

a and b controls the distribution of the parameter μ, and thus called *hyperparameters*

4) Posterior distribution

The posterior distribution of μ: prior distribution(beta) × likelihood function(binomial)

→ Normalization

$$p(\mu|m,l,a,b) = \mathrm{Beta}(\mu|a,b) \times \mathrm{Bin}(\mathrm{m}|\mathrm{N},\mu) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.12}$$

$$l = N - m = \#\ of\ \ tails$$

$$\mathrm{m} = \#\ \text{ of heads}$$

A. Sequential nature

i. a and b (in the prior): *effective number of observations* of x = 1 and x = 0, respectively

ii. The posterior can act as the prior if subsequent observation is followed. If following observation is x=1(x=0), a(b) will increase by 1.

iii. Bayesian viewpoint raises such sequential approach of learning

B. Prediction of the future outcome

$$\mathrm{p}(\mathrm{x}=1|\boldsymbol{\mathcal{D}}) = \int_0^1 p(x=1|\,\mu)p(\mu|\boldsymbol{\mathcal{D}})d\mu = \int_0^1 \mu p(\mu|\boldsymbol{\mathcal{D}})d\mu = \mathbb{E}[\mu|\boldsymbol{\mathcal{D}}] \tag{2.13}$$

$$p(x=1|\boldsymbol{\mathcal{D}}) = \frac{m+a}{m+a+l+b} \tag{2.14}$$

$\rightarrow Total\ \ fraction\ \ of\ \ an\ \ effective\ \ number\ \ for\ \ x=1\ \ (including\ \ both\ \ real\&fictituous)$

C. Big Data

i. Bayesian result converges to ML (general phenomenon):

$$\mathrm{p}(\mathrm{x}=1|\boldsymbol{\mathcal{D}}) = \mathbb{E}[\mu|\boldsymbol{\mathcal{D}}] = \mu_{ML} = sample\ \ mean = \frac{m}{N} \tag{2.15}$$

ii. $\mathrm{var}[\mu|\boldsymbol{\mathcal{D}}]$ is approaching zero (Eq 2.9):

In general, the posterior variance is smaller than the prior variance <u>on average</u> (not for every observation)

$$\mathbb{E}_{\boldsymbol{\mathcal{D}}}\left[var_\theta[\theta|\boldsymbol{\mathcal{D}}]\right] = \boldsymbol{var_\theta[\theta]} - \boldsymbol{var_{\mathcal{D}}}\left[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}]\right] \tag{2.16}$$

2.2. Multinomial Variables

2.2.1. Multinomial Distribution

Binary variable can represent the quantity with two possible states. How about the quantity with K possible states? Let's use K-dimensional vector **x** in which one of the elements x_k equals 1, and all the other elements 0 (mutually exclusive, one-hot-code)

$$\mathbf{x} = (x_1, x_2, \dots, x_K)^T$$

$$\mathrm{p}(x_k=1) = \mu_k \tag{2.17}$$

$$\mathrm{p}(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} {\mu_k}^{x_k} \tag{2.18}$$

It is the generalization of the Bernoulli distribution to more than two outcomes(states)

$$\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$$