PRML Chapter 2. Probability Distribution

Joohyung Lee

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($\mathbf{x}$), 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

1. Binary Variables

1) Bernoulli distribution

A. Definition

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

B. Properties

$$\mathrm{E}[\mathrm{x}] = \mu$$

$$var[x] = \mu(1-\mu)$$

C. Density estimation

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

i. Frequentist

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

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

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

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

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

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

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

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**

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

$$\text{gamma function: } \Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}\,du$$

$$\Gamma(x+1) = x\Gamma(x), \Gamma(1) = 1, \Gamma(x+1) = x!$$

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

3) **Posterior distribution**

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

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

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