PRML Chapter 2. Probability Distribution

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 density estimation, the parameters shall be determined with an observed 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

2.1.1. Bernoulli distribution

2.1.1.1. Definition

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

2.1.1.2. Properties

$$\mathbb{E}[\mathrm{x}] = \mu \qquad (2.2)$$

$$var[x] = \mu(1-\mu) \qquad (2.3)$$

2.1.2. Density estimation

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

2.1.2.1. Frequentist

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

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

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*

2.1.2.2. Bayesian

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

2.1.2.2.1. Binomial distribution

1) Definition

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

2) 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

$$\mathbb{E}[\mathrm{m}] = N\mu \qquad (2.6)$$

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

2.1.2.2.2. 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

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

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

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

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

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

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

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

C. Big Data

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

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

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

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

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

2.2. Multinomial Variables

2.2.1. Multinomial Distribution

2.2.1.1. Definition

$$\mathrm{Mult}(m_1, m_2, \dots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1, m_2, \dots, m_K} \prod_{k=1}^{K} {\mu_k}^{m_k} \tag{2.17}$$

$$m_k = \sum\nolimits_n x_{nk}$$

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

2.2.1.2. Properties

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{\mu}] = \sum\nolimits_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\mu})\boldsymbol{x} = (\mu_1, \dots, \mu_K)^T = \ \boldsymbol{\mu}$$

2.2.2. Density estimation

2.2.2.1. Frequentist

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

$$\mathrm{p}(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} {\mu_k}^{x_{nk}} = \prod_{k=1}^{K} {\mu_k}^{\sum_n x_{nk}} = \prod_{k=1}^{K} {\mu_k}^{m_k}$$

Sufficient statistic for this distribution = $m_k = \sum_n x_{nk}$

$${\mu_k}^{ML} = \frac{m_k}{N}$$

2.2.2.2. Bayesian (Dirichlet distribution)

$$Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)..\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K} {\mu_k}^{\alpha_k-1} \tag{2.18}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$$\mathrm{p}(\mu|\mathcal{D},\alpha) \propto \text{likelihood} \times \text{prior} = \mathrm{p}(\mathcal{D}|\mu)\mathrm{p}(\mu|\alpha) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m})$$

$$= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)..\cdots\Gamma(\alpha_K+m_K)}\prod_{k=1}^{K} {\mu_k}^{\alpha_k+m_k-1}$$

$$*\ \alpha_k = \ effective\ \ number\ \ of\ \ observations\ \ of\ \ x_k = 1$$

2.3. Gaussian Distribution

Gaussian is a widely used model for continuous variable whereas Bernoulli, Binomial, Multinomial are for discrete.

2.3.1. Definition

2.3.1.1. Single variable

$$\mathcal{N}(\mathrm{x}|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.19}$$

2.3.1.2. Multivariate

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}} \tag{2.20}$$

2.3.2. Motivation

1) Gaussian maximizes the entropy for the single, continuous, and real variable.

2) Central Limit Theorem: the sum of a set of random variables becomes more Gaussian as the # of

variable increases (under certain mild condition), e.g., binomial becomes Gaussian as N→∞.

2.3.3. Geometrical form (Transformation)

1) Mahalanobis distance:

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \qquad (2.21)$$

2) $\boldsymbol{\Sigma}$ can be symmetric = its eigenvalue is real & eigenvectors can be an orthonormal set

2.3.3.1. Transform

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \boldsymbol{u_i u_i}^{\boldsymbol{T}} \qquad (2.22)$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} {\lambda_i}^{-1} \boldsymbol{u_i u_i}^{\boldsymbol{T}} \qquad (2.23)$$

Therefore, equation 2.21 becomes,

$$\Delta^2 = \sum_{i=1}^{D} \frac{{y_i}^2}{\lambda_i} \qquad (2.24)$$

$$y_i = \boldsymbol{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}) \qquad (2.25)$$

$$\mathbf{y} = \mathbf{U}(\mathbf{x}-\boldsymbol{\mu}) \qquad (2.26)$$

2.3.3.2. Requirements for normalization

1) Positive definite: Properly normalized with elliptical shape

   A. Center: @ $\boldsymbol{\mu}$

   B. Axes: along $\mathbf{u}$

   C. Scaling factor: ${\lambda_i}^{1/2}$

2) Positive semi-definite: subspace of lower dimensionality (singular)

3) Negative eigenvalue: Probability cannot be defined

2.3.3.3. Normalization

$$|\mathbf{J}|^2 = 1$$

By transforming (shift, rotate) and using a new coordinate system, Multivariate Gaussian becomes the product of D independent univariate Gaussian distribution: $\prod_{j=1}^{D} 1 = 1 \therefore \text{Normalized}$

2.3.3.4. Properties

2.3.3.4.1. First moment

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$$

2.3.3.4.2. Second moment (Covariance)

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\mathrm{T}}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

2.3.4. Limitation and Alternatives

2.3.4.1. Too many parameters

Total of D(D+3)/2 parameters determining Gaussian distribution (grows quadratically with D)

1) Just use Gaussian: D(D+3)/2
2) $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma_i}^{\boldsymbol{2}})$: 2D (axis-aligned ellipsoid)
3) $\boldsymbol{\Sigma} = \sigma^{\boldsymbol{2}}\boldsymbol{I}$: D+1 (isotropic covariance)

2.3.4.2. Unimodal

Introduce latent variable as a solution

1) Discrete latent variable
   A. Mixture of Gaussian
2) Continuous latent variable
   A. Markov random field: to model pixel intensities of an image considering spatial organization
   B. Linear dynamical system: to model time series data for applications such tracking
   C. Probabilistic graphical model

2.3.5. Conditional & Marginal Gaussian distribution

If two sets of distributions are jointly Gaussian,

$$\mathrm{x} = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

2.3.5.1. Conditional distribution

$$\mathrm{p}(x_a|x_b) = \mathcal{N}\left(\boldsymbol{x_a}\middle|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda_{aa}}^{-1}\right) \tag{2.27}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda_{aa}}^{-1}\boldsymbol{\Lambda_{ab}}(\boldsymbol{x_b} - \boldsymbol{\mu}_b) \tag{2.28}$$

*The mean of the conditional distribution, given by (2.27), is a linear function of x_b and that the covariance is independent of x_b.

2.3.5.2. Marginal distribution

$$\mathrm{p}(x_a) = \mathcal{N}(\boldsymbol{x_a}|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{2.29}$$

*the mean and covariance of a marginal distribution are most simply expressed in terms of the partitioned covariance matrix whereas those of conditional distribution are well expressed by the partitioned precision matrix.

2.3.6. Bayes' theorem for Gaussian variables

2.3.6.1. Given distributions

1) Linear Gaussian model, i.e., $\mathbb{E}$(y|x) is linear function of x.
2) Gaussian p(x)

$$\mathrm{p}(\mathbf{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.30}$$

3) Gaussian p(y|x)

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{Ax} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \tag{2.31}$$

2.3.6.2. Target distributions

1) p(y)
2) p(x|y)

2.3.6.3. Note

1) We are given a prior and likelihood instead of the joint distribution as in chapter 2.3.5. We will derive the equation for the target distribution with the prior and likelihood.
2) Find the mean, covariance, and the precision matrix for the joint distribution **z**.

$$\mathbf{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$$

3) Derive the target equations using 2.28 and 2.29.

2.3.6.4. Results

1) Marginal (normalization term)

$$\mathrm{p}(\mathbf{y}) = \mathcal{N}(\boldsymbol{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\ \boldsymbol{L}^{-\mathbf{1}} + \boldsymbol{A}\boldsymbol{\Lambda}^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}) \qquad (2.32)$$

2) Posterior distribution

$$\mathrm{p}(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\boldsymbol{x}|(\boldsymbol{\Lambda} + \boldsymbol{A}^{\boldsymbol{T}}\mathbf{LA})^{-\mathbf{1}}\{\boldsymbol{A}^{\boldsymbol{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\},\ (\boldsymbol{\Lambda} + \boldsymbol{A}^{\boldsymbol{T}}\mathbf{LA})^{-\mathbf{1}}) \qquad (2.33)$$

2.3.7. Maximum likelihood for the Gaussian

2.3.7.1. Given observation

A data set **X** = ($\mathbf{x}_1,\ldots,\mathbf{X}_N)^T$

2.3.7.2. Log likelihood function

$$\ln p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\mathit{ln}|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_{\boldsymbol{n}} - \boldsymbol{\mu})$$

2.3.7.3. Sufficient statistics

$$\sum_{n=1}^{N}\boldsymbol{x}_n, \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}^T$$

2.3.7.4. ML solution for mean and variance

1) Differentiate the log likelihood by $\boldsymbol{\mu}$ and then by $\boldsymbol{\Sigma}$

   i. $\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n$

   ii. $\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_{\boldsymbol{n}} - \boldsymbol{\mu}_{\boldsymbol{ML}})\,(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T$

2) Evaluate ML solutions under the true distribution

   i. $\mathbb{E}[\boldsymbol{\mu}_{ML}] = \frac{1}{N}\sum_{n=1}^{N}x_n$

   ii. $\mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \frac{N-1}{N}\boldsymbol{\Sigma}$

3) Correct the bias for unbiased estimator of variance

   i. $\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}_{\boldsymbol{n}} - \boldsymbol{\mu}_{\boldsymbol{ML}})\,(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T$

2.3.8. Sequential estimation

2.3.8.1. Particular version: mean estimator

Maximum likelihood estimator of the mean based on N observations, i.e. ${\boldsymbol{\mu}_{ML}}^{(N)}$

Is obtained by moving the old estimate a small amount, proportional to 1/N, in the direction of the ‘error’ signal, i.e. $X_N - {\boldsymbol{\mu}_{ML}}^{(N-1)}$

$${\boldsymbol{\mu}_{ML}}^{(N)} = \ {\boldsymbol{\mu}_{ML}}^{(N-1)} + \frac{1}{N}\left(X_N - {\boldsymbol{\mu}_{ML}}^{(N-1)}\right) \qquad (2.34)$$

2.3.8.2. General version: Robbins-Monro algorithm

$$\theta^{(N)} = \ \theta^{(N-1)} - a_{N-1} z(\theta^{(N-1)})$$

Where z is an output of a function that takes θ as its argument.

Three conditional should be satisfied:

1) To converge the process to a limiting value:

$$\lim_{N\to\infty} a_N = 0$$

2) So that the convergence does not short earlier:

$$\sum_{N=1}^{\infty} a_N = \ \infty$$

3) so that the accumulated noise has finite variance and thus does not spoil convergence

$$\sum_{N=1}^{\infty} {a^2}_N < \infty$$

General ML solution for N observations using log likelihood and finding a stationary point:

$$\frac{\partial}{\partial\theta}\{-\frac{1}{N}\sum_{n=1}^{N}\ln p(x_n|\theta)\} \ = \ \mathbb{E}_x[-\frac{\partial}{\partial\theta}\ln p(x|\,\theta)] = 0$$

Therefore, ML solution equals the root of a ~~derivative~~ of regression function. As a result,

$$\mu^{(N)} = \mu^{(N-1)} + a_{N-1}\frac{1}{\sigma^2}\left(x_N - \mu_{ML}{}^{(N-1)}\right) \qquad (2.35)$$

If we choose a_{N-1} as $\frac{\sigma^2}{N}$, then 2.34 becomes equal to 2.35.

*Regression function is

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta)dz$$

2.3.9. Bayesian inference for the Gaussian