Chapter 2 Probability Distribution

Definition. ***Introduction***

1. ***Density Estimation*** *Model probability distribution $p(x)$ of a random variable x, given a finite set $x_1, \cdots, x_N$ of observations.*
2. ***Parametric Distribution*** *Probability distribution based on a fixed set of parameters. For example, Gaussian, binomial, multinomial are parametric distributions*
3. ***Conjugate Priors*** *Priors that lead to posterior distribution having the same functional form (family) as the prior, and therefore lead to greatly simplified Bayesian analysis. THe conjugate prior for parameters of multinomial distribution is Dirichlet distribuution, while conjugate prior for Gaussian is another Gaussian.*
4. ***Nonparametric Distribution*** *Distribution which typically depends on size of data. For example, nearest-neighbours and kernels*

2.1 Binary Variables

Definition. ***Bernoulli Variable*** *A single binary random variable $x \in \{0, 1\}$ has Bernoulli distribution*

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

for $Bern(x|\mu)$,

$$\mathbb{E}\{x\} = \mu \qquad var\{x\} = \mu(1-\mu)$$

For $\mathcal{D} = \{x_1, \cdots, x_N\}$ where $x_i \overset{i.i.d.}{\sim} Bern(\mu)$

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

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

We obtain maximum likelihood estimator

$$\mu_{mle} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Note $m = \sum x_n$ is a sufficient statistics for Bernoulli distribution.

Definition. ***Binomial Distribution*** *Distribution of m observations of $x = 1$ given a data set size of N. We add a normalizing constant of $\binom{N}{m}$ representing all possible ways of obtaining m distinct observations of $x = 1$ where order does not matter*

$$Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \qquad \binom{N}{m} = \frac{N!}{(N-m)!m!}$$

For $Bin(m|N,\mu)$,

$$\mathbb{E}\{m\} = \sum_{m=0}^{N} mBin(m|N,\mu) = N\mu \qquad var\{m\} = \sum_{m=0}^{N} (m-\mathbb{E}\{m\})^2 Bin(m|N,\mu) = N\mu(1-\mu)$$

Definition. ***Beta Distribution*** *as a conjugate prior of binomial distribution. We want to pick a distirbution such that the probability distribution is proportional to* μ *and* $(1-\mu)$*, the posterior distribution. Beta distribution is given by*

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \qquad \Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

Note

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

The coefficients ensures that beta distribution is normalized such that it's a valid probability density function

$$\int_0^1 Beta(\mu|a,b)d\mu = 1$$

The mean and variance given by

$$\mathbb{E}\{\mu\} = \frac{a}{a+b} \qquad var\{\mu\} = \frac{ab}{(a+b)^2(a+b+1)}$$

We derive posterior distribution of μ *by multiplying the beta prior with binomial likelihood*

$$\begin{aligned} p(\mu|m,l,a,b) &\propto p(m|l,\mu)p(\mu|a,b) \\ &= \binom{N}{m}\mu^m(1-\mu)^{N-m}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ &\propto \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \end{aligned}$$

where $l = N - m$*, is number of observations where* $x = 0$*. We see that* m *observations of* $x = 1$ *and* l *observations of* $x = 0$ *has been to increase value of* a *by* m *and* b *by* l*. Hence, we can interpret* a *and* b *as effective number of observations of* $x = 1$ *and* $x = 0$ *respectively. Note the posterior distribution can act as the prior if we subsequently observe additional data.*

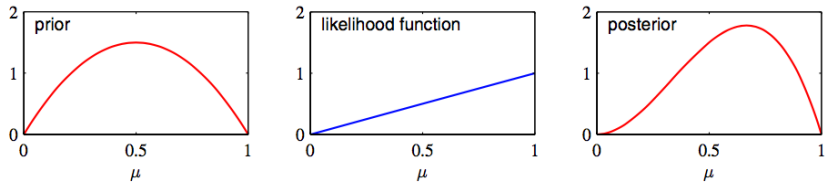


Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2$, $b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution with parameters $a = 3$, $b = 2$.

*The **sequential** approach to learning assumes i.i.d. data and used in real-tiome learning where a stream of data is arriving and that predictions must be made bfore all of the data is seen. Maximum likelihood can be cast into a sequential framework*

2.2 Multinomial Variables

Definition. ***Multinomial Variables*** *A generalization of bernoulli variable, such that we have a discrete variable that can take on one of K possible mutually exclusive states. One representation is 1-of-K scheme in which the variable is represented by a K-dimensional vector $\mathbf{x}$ in which one of $x_k = 1$ and all remaining elements equal 0, such that $\sum_k x_k = 1$. We denote the probability of $x_k = 1$ as μ_k, then the distribution of $\mathbf{x}$ is given by*

$$p(\mathbf{x}|\mu) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \mu = (\mu_1, \cdots, \mu_K)^T \quad \mu_k \geq 0 \quad \sum_k \mu_k = 1$$

The distribution is a genearlization of Bernoulli distribution to more than two outcomes. Notice how the distribution is normalized

$$p(\mathbf{x}|\boldsymbol{\mu}) \geq 0 \qquad \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k^1 = 1$$

$$\mathbb{E}\{x|\mu\} = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = \sum_{k=1}^{K} \mu_k \mathbf{e}_k = (\mu_1, \cdots, \mu_M)^T = \boldsymbol{\mu}$$

Assume a dataset of N observations $\mathbf{x}_1, \cdots, \mathbf{x}_N$, the likelihood is given by

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k} \qquad \ln p(\mathcal{D}|\boldsymbol{\mu}) = \sum_{k=1}^{K} m_k \ln \mu_k$$

where $m_k = \sum_n x_{nk}$, a sufficient statistic representing the number of observations of $x_k = 1$. We maximize $\ln p(\mathcal{D}|\boldsymbol{\mu})$ with respect to μ_k taking account of the constraint the μ_k must sum to one with Lagrange multiplier. We maximize

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

Setting derivative with respect to μ_k to zero, we have $\mu_k = -{}^{m_k}/_{\lambda}$. We solve for λ with $\sum_k \mu_k = 1$ gives $\lambda = -N$. We obtain maximum likelihood solution

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

which is the fraction of N observations for which $x_k = 1$

Definition. ***Multinomial Distribution*** *For n independent trials each of which leads to a success for exactly one of k categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. The joint distribution of $m_1, \cdots, m_K$ conditioned on parameter $\boldsymbol{\mu}$ and on total number of N observations is is known as multinomial distribution*

$$Mult(m_1, m_2, \cdots, m_K|\mu, N) = \binom{N}{m_1 m_2 \cdots m_k} \prod_{k=1}^{K} \mu_k^{m_k}$$

Multinomial distribution is a genearlization of binomial distribution. For example, it models the probability of counts for rolling a k-sided die n times. The normalization coefficients is the number of ways of partitioning N objects into K groups of size $m_1, \cdots, m_K$ and is given by

$$\binom{N}{m_1 m_2 \cdots m_k} = \frac{N!}{m_1! m_2! \cdots m_K!}$$

with constraints $\sum_k m_k = N$.

$$\mathbb{E}\{x\} = N\mu_k \qquad var\{x\} = N\mu_k(1-\mu_k) \qquad cov\{m_j, m_k\} = -N\mu_j\mu_k$$

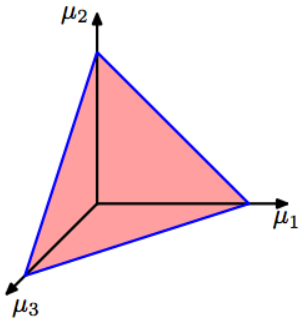
Definition. ***Dirichelet Distribution*** *To find the conjugate prior for the parameter $\{\mu_k\}$ of the multinomial distribution, we see*

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad 0 \le \mu_k \le 1 \quad \sum_k \mu_k = 1$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^T$. The normalized form is called the Dirichlet distribution

$$Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(a_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad \alpha_0 = \sum_k \alpha_k$$

Because of $\sum_k \mu_k = 1$, the distribution over the space of $\{\mu_k\}$ is confined to a simplex of dimensionality $K-1$. For example, ofr $K = 3$, we have



To derive the posterior distribution,

$$p(\boldsymbol{\mu}|\mathcal{D},\alpha) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{a_k+m_k-1}$$

which is also a Dirichlet distribution. Now we determine the normalization coefficients

$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

where $\mathbf{m} = (m_1,\cdots,m_K)^T$ *which are sufficient statistics from the dataset* $\mathcal{D}$*. We interpret* α_k *of Dirichilet prior as an effective number of observations of* $x_k = 1$

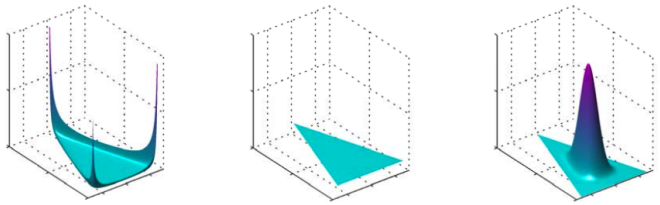


Figure 2.5 Plots of the Dirichlet distribution over three variables, where the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density. Here $\{\alpha_k\} = 0.1$ on the left plot, $\{\alpha_k\} = 1$ in the centre plot, and $\{\alpha_k\} = 10$ in the right plot.

2.3 The Gaussian Distribution

Definition. ***Gaussian***

1. ***Single Variable***

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

where μ *is the mean,* σ^2 *is the variance*

2. ***Multivariate***

$$\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathbf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where D*-dimensional vector* $\boldsymbol{\mu}$ *is called the mean and* $D\times D$ *matrix* $\boldsymbol{\Sigma}$ *is the covariance matrix.*

3. ***Motivation*** *The central limit theorem states that the sum of a set of random variables, which in itself is a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.*

4. ***Quadratic Form***

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

5. ***Mahalanobis distance*** *The quantity* $\triangle$ *is called the Mahalanobis distance from* μ *to* $\mathbf{x}$ *and reduces to the Euclidean distance when* $\boldsymbol{\Sigma}$ *is the identity matrix.*

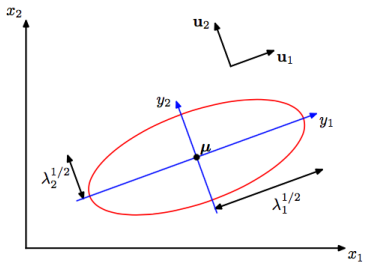
6. ***Covariance Matrix*** *can be expressed as*

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \qquad \boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

where λ_i *and* u_i *are eigenvalue and eigenvector of* $\boldsymbol{\Sigma}$*. The quqdratic form can be written as*

$$\triangle^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \qquad y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu}) \qquad \mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu})$$

where $\{y_i\}$ *can be interpreted as a new coordinate system defined by eigenvectors* u_i*.* $\mathbf{U}$ *is a matrix whose rows are given by* u_i^T



For Gaussian distribution to be well defined, it is necessary and all of λ_i *are strictly positive, i.e.* $\boldsymbol{\Sigma}$ *is a positive definite matrix. For covariance matrix with some zero eigenvalues, distribution is singular and is confined to a subspace of lowe dimensionality.*

7. ***Expectation*** *given by*

$$\mathbb{E}\{x\} = \boldsymbol{\mu} \qquad cov\{x\} = \boldsymbol{\Sigma}$$

8. ***Limitations*** *THe model parameter grows quadratically with* D*, computation task of inverting large matrices can become prohibitive. This can be alleviated by the use of isotropic covariance matrices* $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$*, however restricts the ability to capture interesting correlations in the data. Another limitation is that it is unimodel (has a single maximum) so unable to model multimodal distributions. Introduction of latent variable solves these problems*

2.3.1 Conditional Gaussian Distribution

Definition. ***Conditional Gaussian Distribution*** *The conditional distribution of a set of Gaussian random variables conditioned on the another set is Gaussian. Similarly, the marginal distribution of either set is also Gaussian. Let* $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. We partition* $\mathbf{x}$ *into two sets such that*

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

where $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *has* M *and* $D - M$ *components respectively. Define inverse of covariance matrix as the* ***precision matrix***

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

2.3.4 Maximum Likelihood for the Gaussian

Definition. ***Maximum Likelihood for the Gaussian*** *Given data set* $\mathbf{X} = (\mathbf{x_1}, \cdots, \mathbf{x_N})^{\mathbf{T}}$ *in which* $\mathbf{x}_n \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

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

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) = 0 \qquad \rightarrow \qquad \boldsymbol{\mu}_{mle} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$$

Note $\sum_n x_n$ *and* $\sum_n x_n x_n^T$ *are sufficient statistics. Estimator for covariance matrix is given,*

$$\boldsymbol{\Sigma}_{mle} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{mle})(\mathbf{x}_n - \boldsymbol{\mu}_{mle})^T$$

We check to see if estimators are biased or not

$$\mathbb{E}\{\boldsymbol{\mu}_{mle}\} = \boldsymbol{\mu} \text{ (unbiased)} \qquad \mathbb{E}\{\boldsymbol{\Sigma}_{mle}\} = \frac{N-1}{N} \boldsymbol{\Sigma} \text{ (biased)}$$

2.3.5 Sequential Estimation