Probability Models,Binomial,Beta,Multinomial,and Dirichlet Distributions

Chunpai Wang

There are three reasons to learn probability distribution for solving pattern recognition problems.

- can be used to model the probability distribution $p(X = x)$ of a random variable X, given a finite set $x_1, ..., x_n$ of observations, and this problem is known as *density estimation*.
- can be used to introduce some key statistical concepts, such as *Bayesian Inference*.
- can be used to form buildings blocks for more complex models.

Note, there are infinitely many probability distribution that could have given rise to the observed finite data set. The issue is how to choose an appropriate distribution, and this is very similar to curve fitting problem.

1 Probability Models

Parametric Models: distributions governed by a small number of adaptive parameters Θ, such as mean and variance in the case of a Gaussian. To apply such models to the problem of density estimation, we need to assume a specific functional form for the distribution $P(X|\Theta)$, and determine suitable values $\hat{\Theta}$ for the parameters given an observed data set D.

1. In a *frequentist* treatment: we choose specific values for the parameters by optimizing some criterion, such as the likelihood function.
   Maximum Likelihood (ML):
   $$\textbf{maximize} \;\; p(D|\Theta, \xi) \tag{1}$$
   where ξ represents the prior (background) knowledge.
2. In a *Bayesian* treatment: we introduce prior distributions over the parameters and then use Bayes' theorem to compute the corresponding posterior distribution given the observed data.
   Maximim A Posterior (MAP):
   $$\textbf{maximize} \;\; p(\Theta|D, \xi) = \frac{p(D|\Theta, \xi)p(\Theta|\xi)}{p(D|\xi)} \tag{2}$$

Nonparametric Models: the form of the distribution typically depends on the size of the data set, and such models still contain parameters to control the model complexity rather than the form of the distribution.

2 Binary Variables, Binomial Distribution, and Beta Distribution

Random variable $X \in \{0, 1\}$ has

$$p(X = 1 \mid \mu) = \mu \quad \text{and} \quad p(X = 0 \mid \mu) = 1 - \mu \tag{1}$$

The probability distribution over X can therefore be written in the form

$$Bern(X = x \mid \mu) = \mu^x (1 - \mu)^{1-x} \tag{2}$$

which is known as Bernoulli distribution, and it has mean and variance given by

$$E[X] = \mu \tag{3}$$
$$Var[X] = \mu(1 - \mu) \tag{4}$$

Now given a data set $D = \{x_1, ..., x_n\}$ of observed values of X, we can construct the **likelihood function**, which is a function of μ,on the assumption that the observations are drawn independently from $p(X = x \mid \mu)$, so that

$$p(D \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n}(1 - \mu)^{1-x_n} \tag{5}$$

In the frequentist setting, we can estimate a value for μ by maximizing the (log) likelihood function

$$\ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n \mid \mu) \tag{6}$$

$$= \sum_{n=1}^{N} \ln(\mu^{x_n}) + \ln((1 - \mu)^{1-x_n}) \tag{7}$$

$$= \sum_{n=1}^{N} x_n \ln(\mu) + \sum_{n=1}^{N} (1 - x_n) \ln(1 - \mu) \tag{8}$$

$$= N_1 \ln(\mu) + N_2 \ln(1 - \mu). \tag{9}$$

where N_1 and N_2 are the observations of 1_s and 0_s respectively. Since we have already known the observed values, it is a function respect to μ. If we can set the derivative of $\ln p(D|\mu)$ with respect to μ equal to 0, the maximum likelihood estimator can be obtained

$$\frac{N_1}{\mu} - \frac{N_2}{1 - \mu} = 0 \tag{10}$$

$$\mu_{ML} = \frac{N_1}{N_1 + N_2} = \frac{N_1}{N} \tag{11}$$

2.1 The Binomial Distribution

The distribution of the number N_1 of observations of $X = 1$, given that the data set has size N

$$Bin(N_1 \mid N, \mu) = \binom{N}{N_1} \mu^{N_1}(1 - \mu)^{N-N_1} \tag{12}$$

The mean and variance are

$$E[N_1] = \sum_{N_1=0}^{N} N_1 \; Bin(N_1 \mid N, \mu) = N\mu \tag{13}$$

$$Var[N_1] = \sum_{N_1=0}^{N} (N_1 - E[N_1])^2 Bin(N_1 \mid N, \mu) = N\mu(1 - \mu) \tag{14}$$

2.2 The Beta Distribution

For maximum likelihood estimator, we find that the frequentist treatment will give severely **over-fitted** results for small data set. Hence, we are looking for the Bayesian treatment for this problem, and introduce the maximize a posterior estimate approach and a prior distribution $p(\mu)$ over the parameter μ.
Posterior Distribution: is the probability distribution of an unknown quantity μ conditioned on the evidence or observations D.

$$p(\mu \mid D) = \frac{p(D \mid \mu)p(\mu)}{p(D)} = \frac{p(D \mid \mu)p(\mu)}{\int_\mu p(D \mid \mu')p(\mu')d\mu'} \tag{15}$$

where we can see that

$$posterior\ distribution \propto likelihood\ *\ prior\ distribution \tag{16}$$

In addition, in order to develop a Bayesian treatment for this problem, we need to introduce a prior distribution $p(\mu)$ over the parameter μ. The question here is how to choose a good prior distribution ? What is a good prior distribution ? Here we consider a form of prior distribution that has a simple interpretation as well as some

useful analytical properties. For example, the likelihood function (equation (5)) has the form $\mu^x(1-\mu)^{1-x}$. If we can choose prior distribution in this form,then the posterior distribution will have the same functional form as the prior.

Maximize A Posteriori Estimate: is to select a mode of the posterior distribution:

$$\mu_{MAP} = \mathbf{argmax}_\mu \; p(\mu \mid D, \theta) \tag{17}$$

where θ is the hyperparameter that control the prior distribution of the parameter μ.

Conjugate Prior and Purpose: In Bayesian probability theory, if the posterior distribution are in the same family as the prior distribution, the prior and posterior are then called conjugate distribution, and the prior is called a conjugate prior for the likelihood function. A conjugate prior is an algebraic convenience, giving a *closed-form* expression for the posterior; otherwise numerical integration may be necessary. Further, conjugate priors may give intuitiion, by more transparently showing how a likelihood function updates a prior distribution.

Now, you can find that the conjugate prior of the binomial distribution is the beta distribution

$$Beta(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{18}$$

where $\Gamma(n) = (n-1)!$, and the fraction coefficient ensures that the beta distribution is normalized, so that

$$\int_0^1 Beta(\mu \mid a, b)d\mu = 1 \tag{19}$$

The mean and variance of the beta distribution are given by

$$E[\mu] = \frac{a}{a+b} \tag{20}$$

$$Var[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{21}$$

The posterior distribution of μ is now obtained by multiplying the beta prior by the binomial likelihood function and **normalizing**. Keeping only the factors that depend on μ, we see that this posterior distribution has the form

$$p(\mu \mid N_1, N_2, a, b) \propto \mu^{N_1+a-1}(1-\mu)^{N_2+b-1} \tag{22}$$

where $N_1 + N_2 = N$. The complete representation is given by

$$p(\mu \mid N_1, N_2, a, b) = \frac{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{N_1+a-1}(1-\mu)^{N_2+b-1}}{\int_\mu p(N_1, N_2 \mid \mu')p(\mu')d\mu'} \tag{23}$$

where the purpose of denominator is normalization. If we integrate over the posterior distribution, it will get 1. Therefore, the posterior is simply another beta distribution, which reflecting the conjugacy properties of the prior with respect to the likelihood function. Therefore, its normalization coefficient can be obtained by comparison with beta distribution definition to give

$$p(\mu \mid N_1, N_2, a, b) = \frac{\Gamma(N_1+a+N_2+b)}{\Gamma(N_1+a)\Gamma(N_2+b)}\mu^{N_1+a-1}(1-\mu)^{N_2+b-1} \tag{24}$$

3 Multinomial Variables, Multinomial Distribution, and Dirichlet Distribution

Binary variable can be used to describe quantities that can take one of two possible values. When we need to describe the discrete variable that can take one of k values, we need multinomial variable. Now we use vector $\boldsymbol{x}$ to represent a random variable X, for example

$$\boldsymbol{x} = (0, 0, 0, 1, 0, 0)^\top \tag{1}$$

means the random variable X equals to fourth category, which we denote as $x_4 = 1$, and $\sum_{k=1}^K x_k = 1$. If we denote the probability of $x_k = 1$ by the parameter μ_k, that is

$$p(x_k = 1|\boldsymbol{\mu}) = \mu_k \tag{2}$$

then the distribution of $\boldsymbol{x}$ is given

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k} \tag{3}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)^\top$, and the parameters μ_k are constrained to be $\mu_k > 0$ and $\sum \mu_k = 1$, because they represent probabilities. The distribution above can be regarded as a generalization of the Bernoulli distribution to more than two outcomes.

We have

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1 \tag{4}$$

and that

$$\begin{aligned} E(\boldsymbol{x}|\boldsymbol{\mu}) &= \sum_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{\mu})\boldsymbol{x} \\ &= (\mu_1, 0, ..., 0)^\top + (0, \mu_2, 0, ..., 0)^\top + ... + (0, ..., 0, \mu_K)^\top \\ &= (\mu_1, ..., \mu_K)^\top \\ &= \boldsymbol{\mu} \end{aligned}$$

Now consider N observations $D = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3, ..., \boldsymbol{x}^N\}$, we can compute the likelihood function

$$p(D|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{(\boldsymbol{x}_k^n)} = \prod_{k=1}^{K} \mu_k^{(\sum_n \boldsymbol{x}_k^n)} = \prod_{k=1}^{K} \mu_k^{m_k} \tag{5}$$

where m_k is just the number of observations of $x_k = 1$.

Take logarithm of the likelihood function, we have

$$\begin{aligned} \ln p(D|\boldsymbol{\mu}) &= \ln \prod_{k=1}^{K} \mu_k^{m_k} \\ &= \sum_{k=1}^{K} \ln \mu_k^{m_k} \\ &= \sum_{k=1}^{K} m_k \ln \mu_k \end{aligned}$$

However, there is a constraint $\sum_{k=1}^{K} \mu_k = 1$, and we use Lagrangian multiplier λ and maximize the function

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1) \tag{6}$$

Setting the derivative with respect to μ_k to zero, we obtain

$$\mu_k = -m_k/\lambda \tag{7}$$

and

$$\sum_{k=1}^{K} \mu_k = \sum_{k=1}^{K} \frac{-m_k}{\lambda} = 1 \tag{8}$$

$$\lambda = -N \tag{9}$$

then we can

$$\mu_k^{ML} = \frac{m_k}{N} \tag{10}$$

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

3.1 Multinomial Distribution

We can consider the joint probability of quantities $m_1, ...m_k$ conditioned on the parameters $\boldsymbol{\mu}$ and on the total number N of observations.

$$Mult(m_1, ..., m_K|N, \boldsymbol{\mu}) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k} = \frac{N!}{m_1! m_2! \cdots m_K!} \prod_{k=1}^{K} \mu_k^{m_k} \tag{11}$$

where the normalization coefficient is the number of way of partitioning N objects into K groups of size $m_1, ..., m_K$.

3.2 The Dirichlet Distribution

Similar to beta distribution to binomial distribution, we are looking for a conjugacy prior of multinomial distribution such that

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{12}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^\top$ are parameters of the prior distribution, which is proved to be the Dirichlet distribution:

$$Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{13}$$

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

We can find that posterior takes the form of a Dirichlet distribution, because

$$\textit{posterior distribution} \propto \textit{multinomial likelihood} * \textit{Dirichlet distribution}$$

$$p(\boldsymbol{\mu}|D, \boldsymbol{\alpha}) \propto p(D|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{14}$$

confirming that the Dirichlet is indeed a conjugate prior for the multinomial.

We can also determine the normalization coefficient by

$$p(\boldsymbol{\mu}|D, \boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha} + \boldsymbol{m}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)...\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{15}$$

References

[1] (Chaprter 2) Bishop, Christopher M. Pattern recognition and Machine Learning. springer, 2006.