Binary and multinomial variables

Probability distributions

Francesco Corona

Outline

Probability distributions

We explore now some probability distributions and their properties

- Of great interest in their own right
- Building blocks for complex models

Definition

One role for these distributions is to model the probability distribution $p(\mathbf{x})$ of a random variable $\mathbf{x}$, given some finite set $\mathbf{x}_1, \dots, \mathbf{x}_N$ of observations

- This problem is known as **density estimation**

A problem that is fundamentally ill-posed, because there are infinitely many probability distributions that could have given rise to the observed finite data

- Any $p(\mathbf{x})$ that is nonzero at each of $\mathbf{x}_1, \dots, \mathbf{x}_N$ is a potential candidate

Probability distributions (cont.)

We begin by considering specific examples of **parametric distributions**

- **Binomial** and **multinomial distribution** for **discrete variables**
- The **Gaussian distribution** for **continuous random variables**

Parametric distributions because governed by a number of parameters

To use such models in density estimation problems, we need a procedure

- Determine the values for the model parameters, given observations

Remark

In a frequentist treatment, we set the parameters by optimising some criterion

- For instance, the likelihood function

In a Bayesian treatment we introduce prior distributions over the parameters

- Bayes' theorem to get the posterior

Probability distributions (cont.)

We introduce the important concept of **conjugate prior**

- It is a prior that leads to a formally special posterior
- A posterior with the same functional form as the prior

The conjugate prior for the parameters of a multinomial distribution

- A Dirichlet distribution

The conjugate prior for the mean of a Gaussian distribution

- A Gaussian distribution

All these distributions are members of the **exponential family**

Probability distributions (cont.)

The parametric approach assumes a specific functional form for the distro

- It may turn out to be inappropriate for a particular application

An alternative approach is given by **nonparametric density estimation**

- the form of the distribution often depends on the size of the data

Such models still contain parameters, but they control model complexity

Nonparametric methods: **Histograms**, **near-neighbours**, and **kernels**

Binary variables
Probability distributions

Binary variables

Consider a single binary variable $x \in \{0, 1\}$

Example

Think of an unfair coin, in which probability of tails and heads is different

- x describes the outcome of flipping the coin
- $x = 1$ represents heads
- $x = 0$ represents tails

The probability of $x = 1$ is denoted by the parameter μ, with $0 \leq \mu \leq 1$

- $p(x = 1|\mu) = \mu$
- $p(x = 0|\mu) = 1 - p(x = 1|\mu) = 1 - \mu$

Binary variables (cont.)

The probability distro over x can be written as $\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \implies \begin{cases} x = 0, & \mu^0(1-\mu)^{1-0} = (1-\mu) \\ x = 1, & \mu^1(1-\mu)^{1-1} = \mu \end{cases} \tag{1}$$

This is the **Bernoulli distribution**, so it is normalised $\sum_x \text{Bern}(x|\mu) = 1$ $(\star)$

- with mean $\mathbb{E}[x] = \sum_x x\text{Bern}(x|\mu)$ and variance $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$

$$\mathbb{E}[x] = \mu \tag{2}$$

$$\text{var}[x] = \mu(1-\mu) \tag{3}$$

Binary variables (cont.)

Now suppose we have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ of observed values of x

We can construct the likelihood function of the data

- It is a function of μ

Under the assumption of iid observations from $p(x|\mu)$

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \quad (4)$$

Remark

We can estimate the value for μ by maximising the likelihood function

- Equivalently, we can maximise the log likelihood function

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \Big(x_n \ln \mu + (1-x_n) \ln (1-\mu) \Big) \quad (5)$$

It depends on the N observations only through their sum $\sum_n x_n$

Binary variables (cont.)

If we set the derivative of $\ln p(\mathcal{D}|\mu)$ with respect to μ equal to zero, we get

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{6}$$

The maximum likelihood estimator of the mean of the Bernoulli distribution

- It is known as the **sample mean**, as always

Binary variables (cont.)

Definition

Denoting the number of observations $x = 1$ (heads) in the data set by m

$$\mu_{ML} = \frac{m}{N} \tag{7}$$

The probability of landing heads is the fraction of heads in the data set

If we toss 3 times and observe heads 3 times, $N = m = 3$ and $\mu_{ML} = 1$

The maximum likelihood result would predict all future observations as heads

- Common sense suggests that this is unreasonable
- It is an extreme case of over-fitting

Setting a prior over μ and using Bayes to get a posterior give sensible results

Binary variables (cont.)

We can work out the distribution of the number m of observations of $x = 1$

- given that the data has size N

This is the **binomial distribution**, it is proportional to $\mu^m (1-\mu)^{N-m}$

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \tag{8}$$

- It considers all possible ways of obtaining m heads out of N flips
- The probability of any dataset of N flips of which m are heads is given, assuming independence, by $\mu^m (1-\mu)^{N-m}$

The term $\binom{N}{m}$ (verbally, '*N choose m*') gives the total number of ways of choosing m objects out of a total of N identical objects and it equals ($\star$)

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \tag{9}$$

Binary variables (cont.)

Example

Consider the case where $N = 3$ and $m = 2$

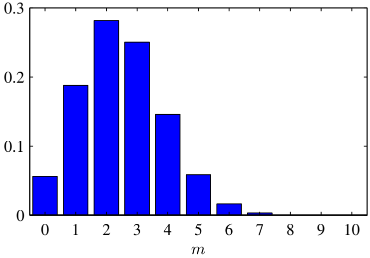
There are $\binom{N=3}{m=2} = 3$ ways of getting two heads out of three coin flips

- $\{\text{HHH}, \text{TTT}, \underbrace{\text{THH}}_{p(\mathcal{D}|\mu)}, \text{TTH}, \text{THT}, \text{HTT}, \underbrace{\text{HHT}}_{p(\mathcal{D}|\mu)}, \underbrace{\text{HTH}}_{p(\mathcal{D}|\mu)}\}$
- $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \text{Bern}(x_n|\mu)$

Each of the three outcomes has the same probability of occurring

$$\begin{aligned} p(\mathcal{D}|\mu) &= \prod_{n=1}^{3} \mu^{x_n}(1-\mu)^{1-x_n} \\ &= \Big[\mu^{x_1}(1-\mu)^{1-x_1}\Big]\Big[\mu^{x_2}(1-\mu)^{1-x_2}\Big]\Big[\mu^{x_3}(1-\mu)^{1-x_3}\Big] \\ &= \mu^{x_1+x_2+x_3}(1-\mu)^{(1+1+1)-(x_1+x_2+x_3)} \\ &= \mu^2(1-\mu)^{3-2} = \mu^m(1-\mu)^{N-m} \end{aligned}$$

Binary variables (cont.)



The binomial distribution

- $N = 10$
- $\mu = 0.25$

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

Binary variables (cont.)

($\star$) For iid events, the mean and variance of the binomial distribution are

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m \text{Bin}(m|N,\mu) = N\mu \quad (10)$$

$$\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N,\mu) = N\mu(1-\mu) \quad (11)$$

$m = x_1 + \cdots + x_N$ and for each x_n the mean is μ and variance is $\mu(1-\mu)$

- The mean of the sum is the sum of the means
- The variance of the sum is the sum of variances

The beta distribution

The maximum likelihood setting for parameter μ in the Bernoulli distribution (and binomial distribution) is the fraction of the observations having $x = 1$

- Severe overfitting for small datasets

To go Bayesian, we need to set a prior distribution $p(\mu)$ over parameter μ

- Here we consider a special form of this prior distribution

The likelihood function takes the form of product of factors $\mu^x(1-\mu)^{1-x}$

- We can choose a prior proportional to powers of μ and $(1-\mu)$

The posterior will be proportional to the product of prior and likelihood

- The posterior will have the same functional form as the prior

Definition

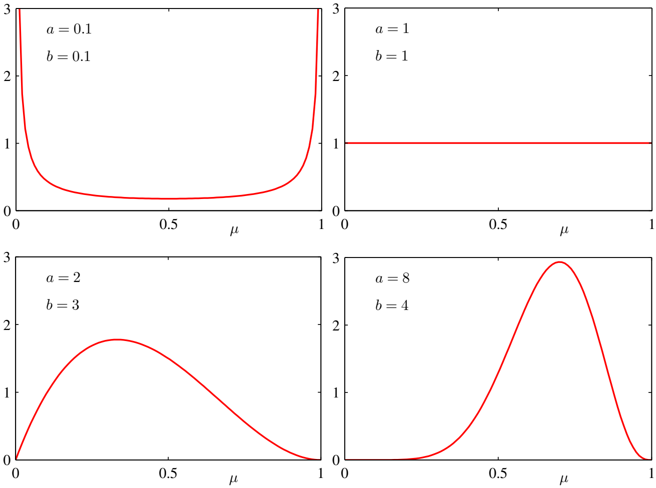
Having a posterior with the same functional form of the prior: **Conjugacy**

The beta distribution (cont.)

We choose a prior distribution called the **beta distribution**

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{12}$$

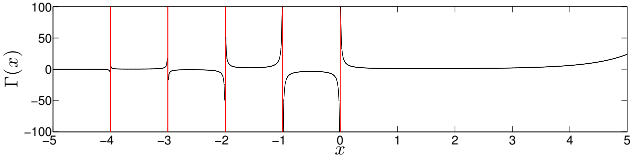
a and b are hyper-parameters controlling the distribution of μ



The beta distribution (cont.)

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

- $\Gamma(\cdot)$ is the gamma function, $\Gamma(x) = \int_0^{+\infty} u^{x-1}e^{-u}du$



- The coefficient ensures normalisation ($\star$)

$$\int_0^1 \text{Beta}(\mu|a,b)d\mu = 1 \tag{13}$$

The beta distribution (cont.)

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

Mean and variance of the beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \quad (14)$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \quad (15)$$

The beta distribution (cont.)

The posterior distribution of μ is obtained by multiplying the beta prior

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

by the binomial likelihood function $\text{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$,

$$p(\mu|m,l,a,b) \propto u^{(m+a)-1}(1-\mu)^{(l+b)-1}, \qquad \text{with } l = N-m \tag{16}$$

where we kept only factors depending on μ to get the expression above

- $l = N - m$ is the number of tails, in the coin example

The beta distribution (cont.)

The posterior distribution over the parameter μ has the same functional form $p(\mu|m, l, a, b) \propto u^{(m+a)-1}(1-\mu)^{(l+b)-1}$ as the beta prior distribution over μ

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

It is in fact another beta distribution with the obvious normalisation coeffcient

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}u^{m+a-1}(1-\mu)^{l+b-1} \quad (17)$$

The beta distribution (cont.)

$$\underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}}_{\text{Beta}(\mu|a,b)} \quad \longrightarrow \quad \underbrace{\frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}u^{m+a-1}(1-\mu)^{l+b-1}}_{p(\mu|m,l,a,b)}$$

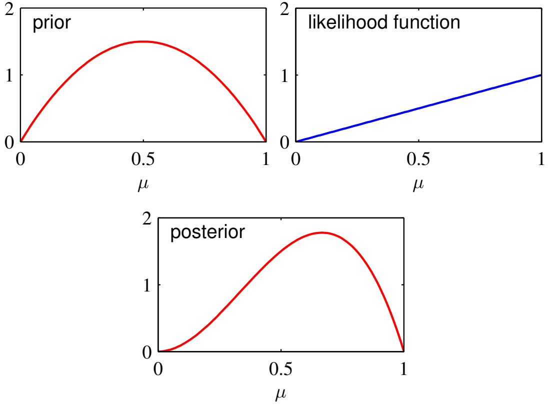
Observing m observations of $x = 1$ and l observations of $x = 0$ has the effect to increase the value of hyper-parameters a and b in the prior over μ

- $a \longrightarrow a + m$
- $b \longrightarrow b + l$

The beta distribution (cont.)

The prior is a beta distribution with parameters $a = 2$ and $b = 2$

The likelihood is for $N = m = 1$, for a single observation $x = 1$ ($l = 0$)



The posterior distribution is another beta distribution with $a = 3$ and $b = 2$

The beta distribution (cont.)

If our goal is to predict the outcome of the next trial, we need the predictive distribution of x, given the observed data set $\mathcal{D}$

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu = \int_0^1 \mu p(\mu|\mathcal{D})d\mu = \mathbb{E}[\mu|\mathcal{D}] \quad (18)$$

Using $p(\mu|\mathcal{D}) = \dfrac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$ and $\mathbb{E}[\mu] = \dfrac{a}{a+b}$

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b} \quad (19)$$

The total fraction of observations (real and fictitious prior) such that $x=1$

Multinomial variables

Probability distributions

Multinomial variables

Binary variables are for quantities that can take one of two possible values

For discrete variables that can take on one of K possible mutually exclusive states there are various alternative ways of representation

- A particularly convenient scheme is called **1-of-K**

The variable is represented by a K-dimensional vector $\mathbf{x}$ in which we have

- only one of the elements x_k equals 1
- all of the other elements $x_{k \not{}}$ equal 0
- $\sum_{k=1}^{K} x_k = 1$

For example, $\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$ with $K = 6$ states and observation $x_3 = 1$

Multinomial variables (cont.)

Denote the probability of $x_k = 1$ by the parameter μ_k with the constraint that $\mu_k \geq 0$ and $\sum_k \mu_k = 1$ because they represent probabilities, we have that

- the distribution of $\mathbf{x}$ is given by

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{20}$$

- where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^T$

The distribution is a generalisation ($K > 2$) of the Bernoulli distribution

- It is normalised

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1 \tag{21}$$

- $$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \dots, \mu_K)^T = \boldsymbol{\mu} \tag{22}$$

Multinomial variables (cont.)

Consider a dataset $\mathcal{D}$ of N iid observations $\mathbf{x}_1, \dots, \mathbf{x}_N$, the likelihood function

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k} \tag{23}$$

depends on the N points only through the K quantities $m_k = \sum_n x_{nk}$[1]

[1] It is the number of observations of $x_k = 1$

Multinomial variables (cont.)

To find the maximum likelihood solution for $\boldsymbol{\mu}$, we maximise $\ln p(\mathcal{D}|\boldsymbol{\mu})$ wrt μ_k

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda\Big(\sum_{k=1}^{K} \mu_k - 1\Big) \qquad (24)$$

where we took into account of the constraint that μ_k must sum up to one

Setting the derivative wrt μ_k to zero, we get

$$\mu_k = -\frac{m_k}{\lambda} \qquad (25)$$

with $\lambda = -N$, by substitution in $\sum_k \mu_k = 1$

$$\mu_k^{ML} = \frac{m_k}{N} \qquad (26)$$

the fraction of $x_k = 1$ cases out of N cases

Multinomial variables (cont.)

Consider the joint distribution of the quantities $m_1, \dots, m_K$ conditioned on the parameters $\boldsymbol{\mu}$ and on the total number N of observations, from Equation 23

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \qquad (27)$$

which is known as the **multinomial distribution**

Remark

The normalisation coefficient is the number of ways of partitioning N objects into K groups of size $m_1, \cdots, m_K$

$$\binom{N}{m_1 m_2 \cdots m_K} = \frac{N!}{m_1! m_2! \cdots m_K!} \qquad (28)$$

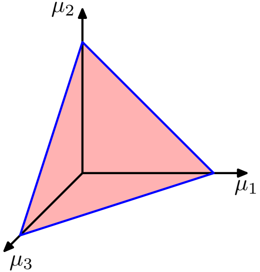
Note that variables m_k are such that $\sum_k m_k = N$

The Dirichlet distribution

A family of priors for the parameters $\{\mu_k\}$ of the multinomial distribution

- Again, by inspection of the form of the multinomial distribution
- Proportional to powers of μ_k

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \quad \text{with } 0 \leq \mu_k \leq 1 \text{ and } \sum_k \mu_k = 1 \qquad (29)$$



$\alpha_1, \ldots, \alpha_k$ are the parameters of the distribution

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$$

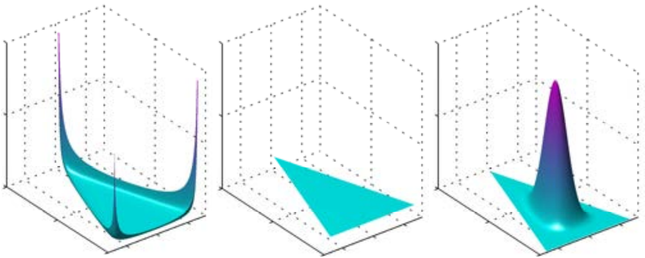
Because of the sum constraint, the distribution over the space of $\{\mu_k\}$ is confined to a simplex

- Bounded $(K-1)$-dimensional linear manifold

The Dirichlet distribution (cont.)

In normalised form, this is known as the **Dirichlet distribution**

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_K^{\alpha_k - 1} \quad \text{with } \alpha_0 = \sum_{k=1}^{K} \alpha_k \qquad (30)$$



The Dirichlet distribution over three variables, for various settings of $\{\alpha_k\}$

The horizontal axes represents coordinates in the plane of the simplex

The vertical axis corresponds to the density

The Dirichlet distribution (cont.)

Multiplying prior $\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \dfrac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_K^{\alpha_k - 1}$ and likelihood function $\text{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$ gives us

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\ldots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} = \text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \quad (31)$$

- The posterior distribution for the parameters $\{\mu_k\}$

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu}) p(\boldsymbol{\mu}, \boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \quad (32)$$

- Again, it takes the form of a Dirichlet distribution
- The normalisation is by comparison with $\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha})$