Probability Distribution

Mingmin Chi

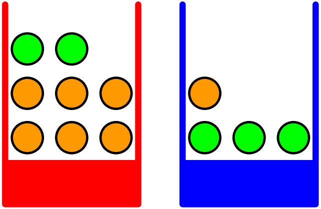
Fudan University, Shanghai, China

Outline

1. Probability Theory
2. Binary Variables
3. Multinomial Variables
4. The Gaussian Distribution

Simple Example

- Uncertainty is a key concept in the fields of pattern recognition and machine learning
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for our study

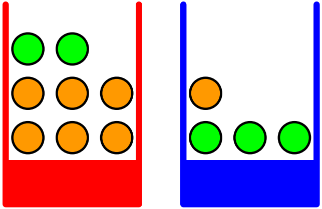


- Example: one red & one blue box
  - 2 apples and 6 oranges in the red box
  - 3 apples and 1 orange in the blue box
- choosing box is random, denoted by B,

- $P(B = r) = 4/10$
- $P(B = b) = 6/10$

Simple Example

- Uncertainty is a key concept in the fields of pattern recognition and machine learning
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for our study

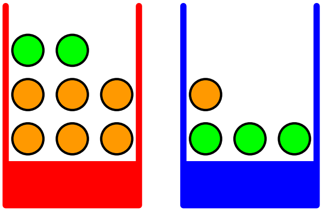


- $P(B = r) = 4/10$
- $P(B = b) = 6/10$

- Example: one red & one blue box
  - 2 apples and 6 oranges in the red box
  - 3 apples and 1 orange in the blue box
- choosing box is random, denoted by B, i.e., $B = r$ or $B = b$
- identity of the fruit is also a random variable, denoted by F,

Simple Example

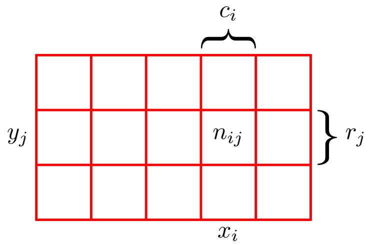
- Uncertainty is a key concept in the fields of pattern recognition and machine learning
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for our study



- $P(B = r) = 4/10$
- $P(B = b) = 6/10$

- Example: one red & one blue box
  - 2 apples and 6 oranges in the red box
  - 3 apples and 1 orange in the blue box
- choosing box is random, denoted by B, i.e., $B = r$ or $B = b$
- identity of the fruit is also a random variable, denoted by F, i.e., $F = a$ or $F = o$

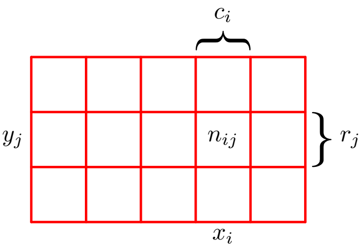
A General Example



- two random variables X and Y
- suppose X can take any of the values $(x_i)_{i=1}^{M}$
- suppose that Y can take the values $(y_j)_{j=1}^{L}$

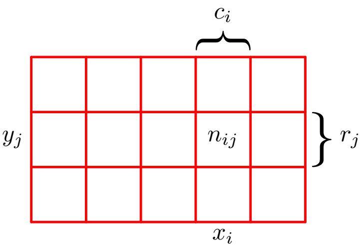
- consider a total of N trials in which we sample both of the variables X and Y
- let n_{ij} be the number of such trials in which $X = x_i$ and $Y = y_j$
- let r_j be the number of trials in which $Y = y_j$
- let c_i be the number of trials in which $X = x_i$

A General Example (cont'd)



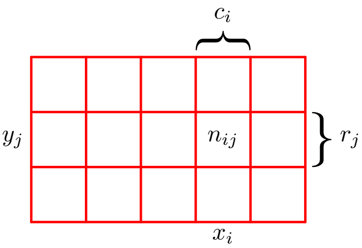
- $P(X = x_i) =$

A General Example (cont'd)



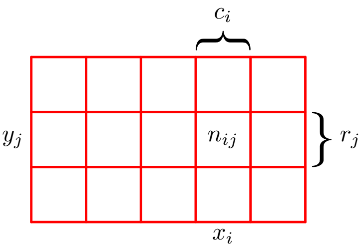
- $P(X = x_i) = c_i/N$
- $P(Y = y_j) = r_j/N$
- joint probability
  $P(X = x_i, Y = y_j) =$

A General Example (cont'd)



- $P(X = x_i) = c_i/N$
- $P(Y = y_j) = r_j/N$
- joint probability
  $P(X = x_i, Y = y_j) = n_{ij}/N$
- conditional probability
  $P(Y = y_j | X = x_i) =$

A General Example (cont'd)

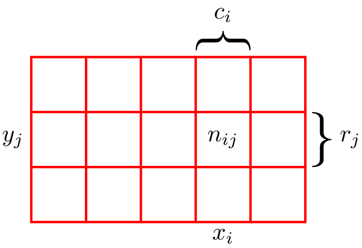


- $P(X = x_i) = c_i / N$
- $P(Y = y_j) = r_j / N$
- joint probability
  $P(X = x_i, Y = y_j) = n_{ij} / N$
- conditional probability
  $P(Y = y_j | X = x_i) = n_{ij} / c_i$

The rules of probability

1. sum rule: $P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$ [a]
2. product rule:
   $P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$

A General Example (cont'd)



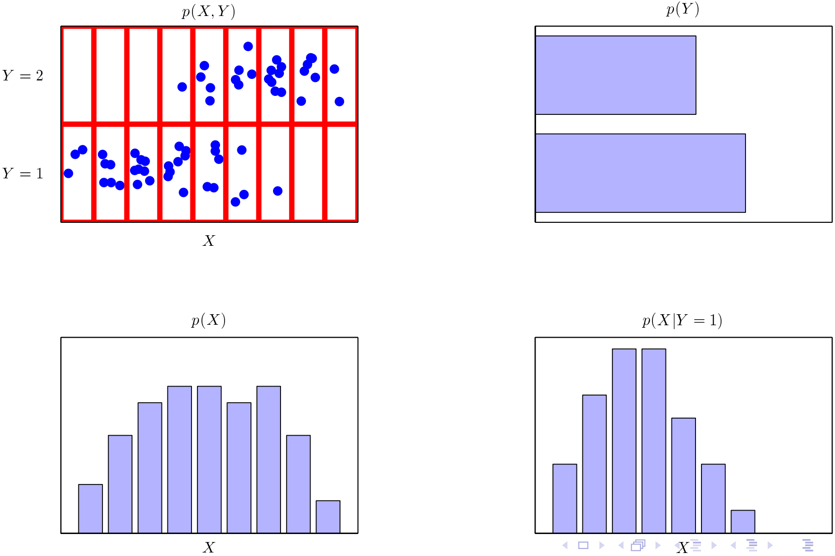
- $P(X = x_i) = c_i/N$
- $P(Y = y_j) = r_j/N$
- joint probability
  $P(X = x_i, Y = y_j) = n_{ij}/N$
- conditional probability
  $P(Y = y_j | X = x_i) = n_{ij}/c_i$

The rules of probability

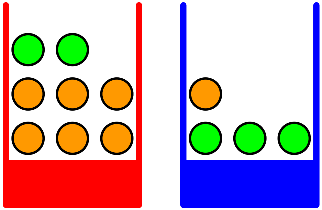
1. sum rule: $P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$ [a]
2. product rule:
   $P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P(Y = y_j | X = x_i) \cdot P(X = x_i)$ [b]

[a] $P(X = x_i)$ is sometimes called the marginal probability
[b] We can derive the Bayes's Theorem.

An Illustration



Example: revisit



$$p(F = a|B = b) = ?$$
$$p(F = a) = ?$$
$$p(B = r|F = o) = ?$$

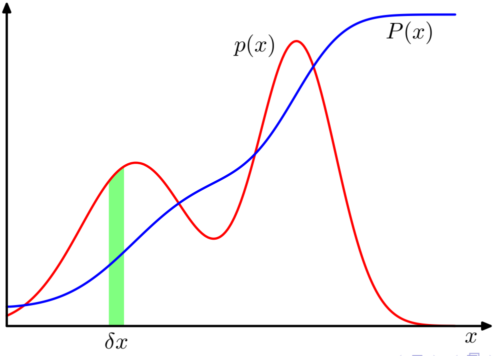
- $P(B = r) = 4/10$
- $P(B = b) = 6/10$

Probability Density

Considering probabilities with respect to continuous variables

Informal definition

If the probability of a real-valued variable x falling in the interval $(x, x+\delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$, then $p(x)$ is called the probability density over x



Probability Density (cont'd)

The probability that x will lie in an interval (a, b) is given by

$$P(x \in (a, b)) = \int_a^b p(x)dx$$

Cumulative distribution function

The probability that x lies in the interval $(-\infty, z)$ is given by

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

Note that If x is a discrete variable, then $p(x)$ is sometimes called a probability mass function

Expectations

The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the expectation of $f(x)$, denoted by $\mathcal{E}[f]$

Expectation

- For a discrete distribution,

$$\mathcal{E}[f] = \sum_x p(x)f(x)$$

- For a continuous distribution,

$$\mathcal{E}[f] = \int p(x)f(x)dx$$

Expectations (cont'd)

In both the continuous and discrete cases, if given a finite number N of points drawn from the probability distribution or probability density, then we can approximate it as a finite sum over these points

$$\mathcal{E}[f] \cong \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

How about expectations of functions of several variables

Expectations (cont'd)

In both the continuous and discrete cases, if given a finite number N of points drawn from the probability distribution or probability density, then we can approximate it as a finite sum over these points

$$\mathcal{E}[f] \cong \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

How about expectations of functions of several variables

$\mathcal{E}_x[f(x, y)]$: the average of the function $f(x, y)$ with respect to the distribution of x

Conditional expectation

$$\mathcal{E}_x[f|y] =$$

Expectations (cont'd)

In both the continuous and discrete cases, if given a finite number N of points drawn from the probability distribution or probability density, then we can approximate it as a finite sum over these points

$$\mathcal{E}[f] \cong \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

How about expectations of functions of several variables

$\mathcal{E}_x[f(x,y)]$: the average of the function $f(x,y)$ with respect to the distribution of x

Conditional expectation

$$\mathcal{E}_x[f|y] = \sum_x p(x|y)f(x)$$

Variance

The variance of $f(x)$

$$var[f] = \mathcal{E}\left[(f(x) - \mathcal{E}[f(x)])^2\right] = \mathcal{E}[f(x)^2] - \mathcal{E}[f(x)]^2$$

Covariance

Variance

The variance of $f(x)$

$$var[f] = \mathcal{E}\left[(f(x) - \mathcal{E}[f(x)])^2\right] = \mathcal{E}[f(x)^2] - \mathcal{E}[f(x)]^2$$

Covariance

- For two random variables x, y

$$cov[x, y] = \mathcal{E}_{x,y}\left[\{x - \mathcal{E}[x]\}\right]\left[\{y - \mathcal{E}[y]\}\right] = \mathcal{E}_{x,y}[xy] - \mathcal{E}[x]\mathcal{E}[y]$$

- For two vectors of random variables **x** and **y**,

$$cov[\mathbf{x}, \mathbf{y}] = \mathcal{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathcal{E}[\mathbf{x}]\}\right]\left[\{\mathbf{y}^\top - \mathcal{E}[\mathbf{y}^\top]\}\right] = \mathcal{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^\top] - \mathcal{E}[\mathbf{x}]\mathcal{E}[\mathbf{y}^\top]$$

Bernoulli Distribution

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

- The probability of $x = 1$ will be denoted by the parameter μ so that $p(x = 1|\mu) = \mu$, where $0 \leq \mu \leq 1$
- easily it follows that $p(x = 0|\mu) =$

Bernoulli Distribution

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

- The probability of $x = 1$ will be denoted by the parameter μ so that $p(x = 1|\mu) = \mu$, where $0 \leq \mu \leq 1$
- easily it follows that $p(x = 0|\mu) = 1 - \mu$
- the probability distribution over x can be written in the form

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

- easily to verify that this distribution is normalized and that it has mean and variance given by

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

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

Bernoulli Distribution (cont'd)

Suppose we have a dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$ of observed values of x

Likelihood function

$$p(\mathcal{D}|\mu) =$$

Bernoulli Distribution (cont'd)

Suppose we have a dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$ of observed values of x

Likelihood function

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

Bernoulli Distribution (cont'd)

Suppose we have a dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$ of observed values of x

Likelihood function

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

We can estimate a value for μ by maximizing the likelihood function, or equivalently by maximizing the logarithm of the likelihood

$$\ln p(\mathcal{D}|\mu) =$$

Bernoulli Distribution (cont'd)

Suppose we have a dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$ of observed values of x

Likelihood function

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

We can estimate a value for μ by maximizing the likelihood function, or equivalently by maximizing the logarithm of the likelihood

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

Bernoulli Distribution (cont'd)

Suppose we have a dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$ of observed values of x

Likelihood function

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

We can estimate a value for μ by maximizing the likelihood function, or equivalently by maximizing the logarithm of the likelihood

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

With $\frac{\partial \ln p(\mathcal{D}|\mu)}{\partial \mu} = 0 \rightarrow \mu_{ML} =$

Bernoulli Distribution (cont'd)

Suppose we have a dataset $\mathcal{D} = \{x_1, \cdots, x_N\}$ of observed values of x

Likelihood function

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

We can estimate a value for μ by maximizing the likelihood function, or equivalently by maximizing the logarithm of the likelihood

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

With $\frac{\partial \ln p(\mathcal{D}|\mu)}{\partial \mu} = 0 \rightarrow \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$

Overfitting of ML

If we denote the number of observations of $x = 1$ (heads) within this data set by m, then

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

Example: Suppose now flip the coin 5 (N=5) times, and happen to observe 5 (m=5) heads. Then, $\mu_{ML} = 1$. What does it mean?

Overfitting of ML

If we denote the number of observations of $x = 1$ (heads) within this data set by m, then

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

Example: Suppose now flip the coin 5 (N=5) times, and happen to observe 5 (m=5) heads. Then, $\mu_{ML} = 1$. What does it mean?

The ML solution would predict that all future observations should give heads.

Binomial Distribution

Consider an extreme example of the over-fitting associated with maximum likelihood

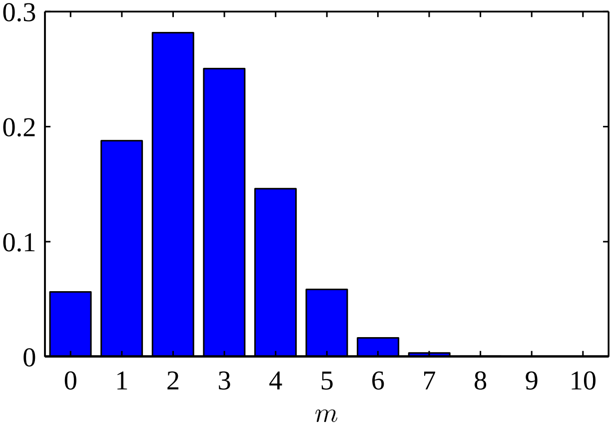
- binomial distribution: the distribution of the number m of observations of $x = 1$, given that the dataset has size N: $\mu^m(1-\mu)^{N-m}$
- If we work the distribution of the number m of observations of $x = 1$ given that the dataset has size N, we can obtain the binomial distribution

$$\mathrm{Bin}(m|N,\mu) = \underbrace{\binom{N}{m}}_{\frac{N!}{(N-m)!m!}}{}^{a} \mu^m(1-\mu)^{N-m}$$

[a] The number of ways of choosing m objects out of a total of N identical objects.

Binomial Distribution (cont'd)

Histogram plot of the binomial distribution as a function of m for $N = 10$ and $\mu = 0.25$



The Beta Distribution

- Problem by maximum likelihood estimation in the binomial distribution -

The Beta Distribution

- Problem by maximum likelihood estimation in the binomial distribution - over-fitted results for small datasets
- Solution -

The Beta Distribution

- Problem by maximum likelihood estimation in the binomial distribution - over-fitted results for small datasets
- Solution - Bayesian treatment -

The Beta Distribution

- Problem by maximum likelihood estimation in the binomial distribution - over-fitted results for small datasets
- Solution - Bayesian treatment - a prior distribution $p(\mu)$ needed

Conjugate prior

- Remember the likelihood function takes the form $\mu^x(1-\mu)^{1-x}$
- If we choose a prior to be proportional to power of μ and $(1-\mu)$, then the posterior distribution,

The Beta Distribution

- Problem by maximum likelihood estimation in the binomial distribution - over-fitted results for small datasets
- Solution - Bayesian treatment - a prior distribution $p(\mu)$ needed

Conjugate prior

- Remember the likelihood function takes the form $\mu^x(1-\mu)^{1-x}$
- If we choose a prior to be proportional to power of μ and $(1-\mu)$, then the posterior distribution,will have the same functional form as the prior
- such kind of priors is called conjugate prior

The Beta Distribution

- Problem by maximum likelihood estimation in the binomial distribution - over-fitted results for small datasets
- Solution - Bayesian treatment - a prior distribution $p(\mu)$ needed

Conjugate prior

- Remember the likelihood function takes the form $\mu^x(1-\mu)^{1-x}$
- If we choose a prior to be proportional to power of μ and $(1-\mu)$, then the posterior distribution,will have the same functional form as the prior
- such kind of priors is called conjugate prior

We therefore choose a prior, called the beta distribution, given by

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

The Beta Distribution (cont'd)

The beta distribution is normalized,

The Beta Distribution (cont'd)

The beta distribution is normalized,

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

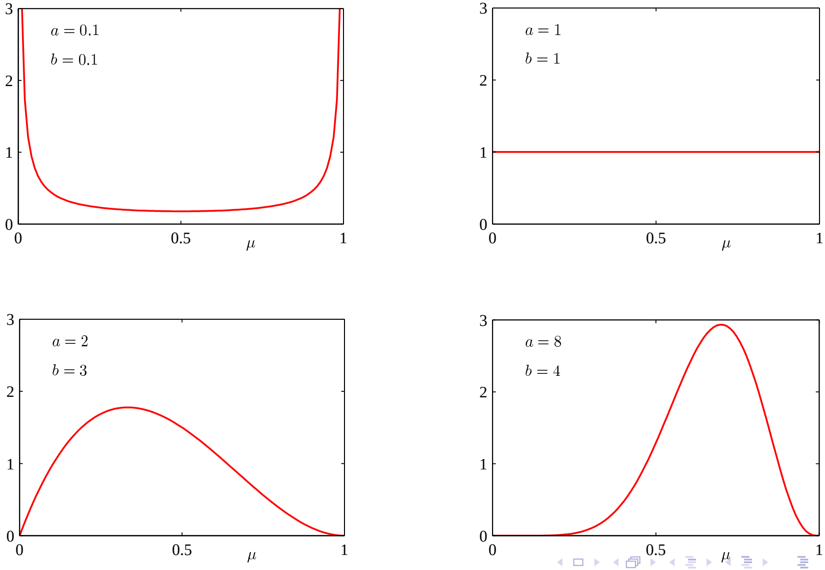
The mean and variance of the beta distribution are given by

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

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

The parameters a and b are often called hyperparameters

The Beta Distribution (cont'd)



The Beta Distribution - Posterior Distribution

The Posterior Distribution of μ is now obtained by multiplying the beta prior by the binomial likelihood function and normalizing.

The Beta Distribution - Posterior Distribution

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, this posterior distribution has the form

$$p(\mu|m, l, a, b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

where $l = N - m$

The Beta Distribution - Posterior Distribution

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, this posterior distribution has the form

$$p(\mu|m, l, a, b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

where $l = N - m$

We can see that the posterior distribution is simply another beta distribution

The Beta Distribution - Posterior Distribution

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, this posterior distribution has the form

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

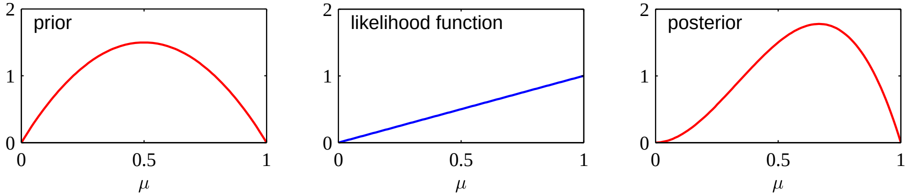
where $l = N - m$

We can see that the posterior distribution is simply another beta distribution

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

Illustration

The prior is given by a beta distribution with parameters $a = 2, b = 2$, and the likelihood function, given by binomial distribution with $N = m = 1$, corresponds to a single observation of $x = 1$



We can see that the posterior is given by a beta distribution with parameters $a = 3, b = 2$

We can interpret a, b in the prior as an effective number of observations of $x = 1$ and $x = 0$, respectively

The Beta Distribution - Prediction

Prediction, given the prior and observations $\mathcal{D}$,

The Beta Distribution - Prediction

Prediction, given the prior and observations $\mathcal{D}$,

$$
\begin{aligned}
P(x=1|\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu \\
&= \int_0^1 \mu p(\mu|\mathcal{D})d\mu \\
&=
\end{aligned}
$$

The Beta Distribution - Prediction

Prediction, given the prior and observations $\mathcal{D}$,

$$\begin{aligned} P(x=1|\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu \\ &= \int_0^1 \mu p(\mu|\mathcal{D})d\mu \\ &= \mathcal{E}[\mu|\mathcal{D}] \\ &= \frac{m+a}{m+a+l+b} \end{aligned}$$

Introduction

- Consider a discrete variable that can take one of possible K values
- Convenient representation with a vector where one element equals 1, others 0, e.g., $\mathbf{x} = (0, 0, 1, 0, 0, 0)^\top$
- If denoting the probability of $x_k = 1$ by the parameter μ_k, then the distribution of $\mathbf{x}$ is given

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

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)^\top$, s.t., $\mu_k \geq 0$ and $\sum_k \mu_k = 1$

Generalization of the Bernoulli Distribution

- the distribution is normalized

Generalization of the Bernoulli Distribution

- the distribution is normalized

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

and that

$$E[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \cdots, \mu_K)^\top = \boldsymbol{\mu}$$

- the likelihood function

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

Generalization of the Bernoulli Distribution

- the distribution is normalized

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

and that

$$E[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \cdots, \mu_K)^\top = \boldsymbol{\mu}$$

- the likelihood function

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

[a] The number of observations of $x_k = 1$, and $m_k = \sum_n x_{nk}$

Maximum Likelihood Estimator

By a Lagrange multiplier λ and maximizing

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

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

Multinomial Distribution

Consider the joint distribution of the quantities $m_1, \cdots, m_K$, the multinomial distribution takes the form

$$\mathrm{Mult}(m_1, \cdots, m_K|\boldsymbol{\mu}, N) = \underbrace{\begin{pmatrix} N \\ m_1 m_2 \cdots m_K \end{pmatrix}}_{\frac{N!}{m_1!m_2!\cdots m_K!}} \prod_{k=1}^{K} \mu^{m_k}$$

The variables m_k are subject to the constraint $\sum_{k=1}^{K} m_k = N$

Dirichlet Distribution

- a family of conjugate prior distributions for the parameters $\{\mu_k\}$
- respected to the multinomial distribution, the conjugate prior is given by

Dirichlet Distribution

- a family of conjugate prior distributions for the parameters $\{\mu_k\}$
- respected to the multinomial distribution, the conjugate prior is given by

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \quad \text{s.t.} \begin{cases} 0 \le \mu_k \le 1 \\ \sum_k \mu_k = 1 \end{cases}$$

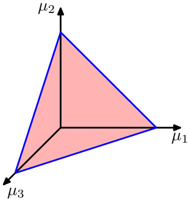
- The normalized form the distribution by

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

This is called the Dirichlet distribution.

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

Dirichlet Distribution (cont'd)

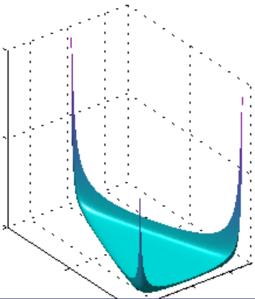


$\Leftarrow$ The domain of the Dirichlet distribution with $K = 3$

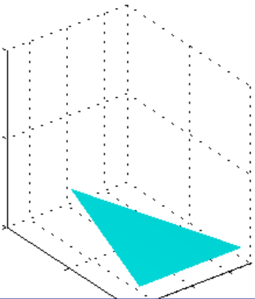
Plots of the Dirichlet distribution ($K = 3$)

$\Downarrow$

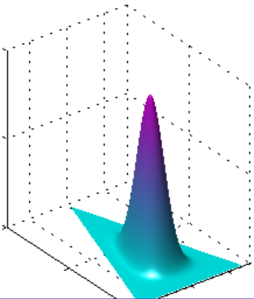
$\alpha_k = 0.1$



$\alpha_k = 1$



$\alpha_k = 10$



Single Variable Gaussian

For a single variable x

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

where μ is the mean and σ^2 is the variance

Multivariable Gaussian

For a d-dimensional vector $\mathbf{x}$

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

where μ is a d-dimensional mean vector and Σ is a $d \times d$ covariance matrix, and $|\Sigma|$ denotes the determinant of Σ

Geometrical Form

Mahalanobis distance

The functional dependence of the Gaussian on $\mathbf{x}$ is through the quadratic form

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

The quantity Δ is called the Mahalanobis distance from $\boldsymbol{\mu}$ to $\mathbf{x}$ and

Geometrical Form

Mahalanobis distance

The functional dependence of the Gaussian on $\mathbf{x}$ is through the quadratic form

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

The quantity Δ is called the Mahalanobis distance from $\boldsymbol{\mu}$ to $\mathbf{x}$ and reduces to the Euclidean distance when Σ is the identity matrix

Consider the eigenvector equation for the covariance matrix

$$\Sigma \boldsymbol{\mu}_i = \lambda_i \boldsymbol{\mu}_i$$

Since Σ is a real, symmetric matrix, its eigenvalues will be real, and its eigenvectors can be chosen to form an orthonormal set, so that,

$$\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = \mathrm{I}_{ij}$$

Geometrical Form (cont'd)

The covariance matrix can be expressed as an expansion in terms of its eigenvectors in the form

$$\Sigma = \sum_{i=1}^{d} \lambda_i \boldsymbol{\mu}_i \boldsymbol{\mu}_j^\top$$

and similarly the inverse covariance matrix Σ^{-1} can be expressed as

Geometrical Form (cont'd)

The covariance matrix can be expressed as an expansion in terms of its eigenvectors in the form

$$\Sigma = \sum_{i=1}^{d} \lambda_i \boldsymbol{\mu}_i \boldsymbol{\mu}_j^\top$$

and similarly the inverse covariance matrix Σ^{-1} can be expressed as

$$\Sigma^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda_i} \boldsymbol{\mu}_i \boldsymbol{\mu}_j^\top$$

Geometrical Form (cont'd)

$$\Sigma^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda_i} \boldsymbol{\mu}_i \boldsymbol{\mu}_j^\top \text{ and } \Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}),$$

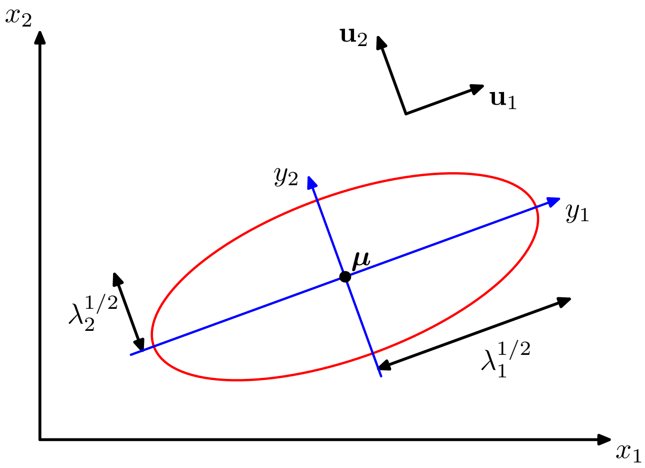
therefore we have,

$$\Delta^2 = \sum_{i=1}^{d} \frac{y_i^2}{\lambda_i}, \; y_i = \boldsymbol{\mu}_i^\top (\mathbf{x} - \boldsymbol{\mu})$$

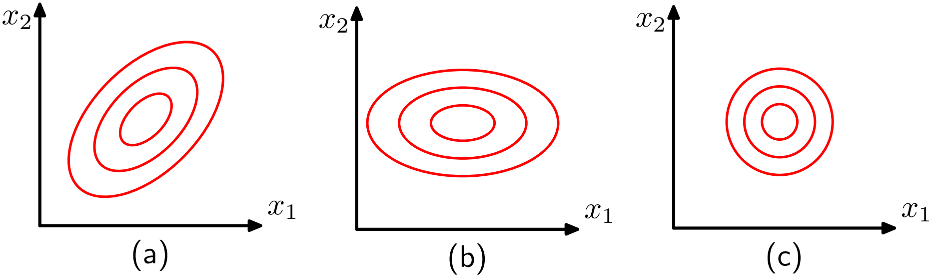
- We can interpret $\{y_i\}$ as a new coordinate system defined by the orthonormal vectors $\boldsymbol{\mu}_i$ that are shifted and rotated with respect to the original x_i coordinates
- Forming the vector $\mathbf{y} = (y_1, \cdots, y_d)^\top$, we have

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

Geometrical Form (cont'd)



One of Limitations of Gaussian Distribution

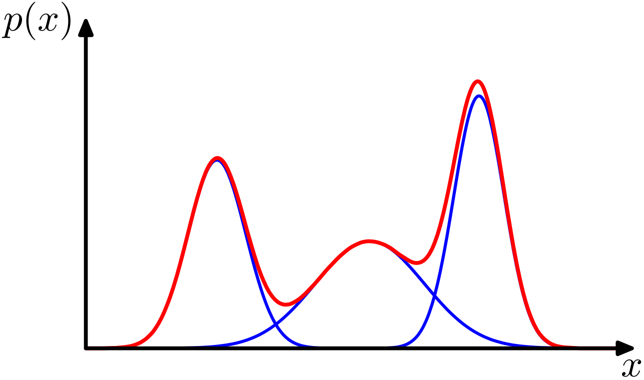


Σ	$\Sigma = \text{diag}(\sigma_i^2)$	$\Sigma = \sigma^2 \mathbf{I}$
$D(D+3)/2$	$2D$	$D+1$

Mixture of Gaussians

Another limitations of Gaussian distribution is that it is uni-modal

The superpositions, formed by taking linear combinations of more basic distributions, can be formulated as probabilistic models known as mixture distribution



Mixture of Gaussians (cont'd)

Consider a superposition of K Gaussian densities of the form

$$p(x) = \sum_{k=1}^{K} \pi_k \mathbf{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

which is called a mixture of Gaussians