



MACHINE LEARNING

CHAPTER 2: PROBABILITY DISTRIBUTIONS

Learning Objectives

- 1、 What are binary, multinomial and Gaussian distributions and their conjugate prior distributions?
 - 2、 What are the common properties of Gaussian distributions?
 - 3、 What are exponential families and their properties?
 - 4、 How to choose non-informative prior*?
 - 5、 How to use non-parametric methods for learning?
 - 6、 What are KNN based methods?
-

Outlines

- Binary Distributions
 - Multinomial Distributions
 - Gaussian Distributions
 - Exponential Families
 - Non-informative Prior
 - Non-parametric Methods
 - KNN
-

Parametric Distributions

Basic building blocks: $p(\mathbf{x}|\boldsymbol{\theta})$

Need to determine $\boldsymbol{\theta}$ given $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

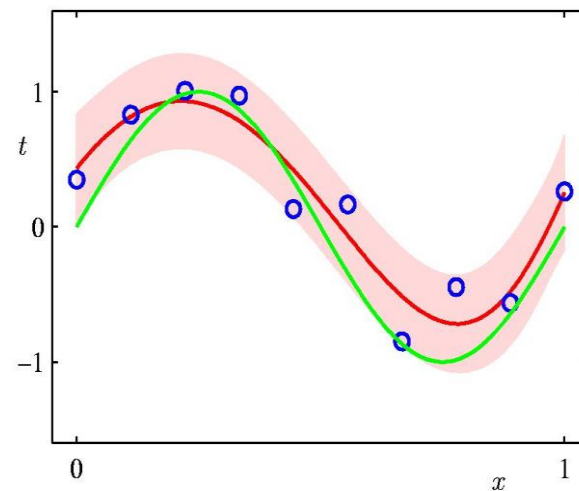
Representation: $\boldsymbol{\theta}^*$ or $p(\boldsymbol{\theta})$?

Model parameter Learning

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Recall Curve Fitting Prediction

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}$$



Binary Variables (1)

Coin flipping: heads=1, tails=0

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

Bernoulli Distribution

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

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

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

Binary Variables (2)

N coin flips:

$$p(m \text{ heads} | N, \mu)$$

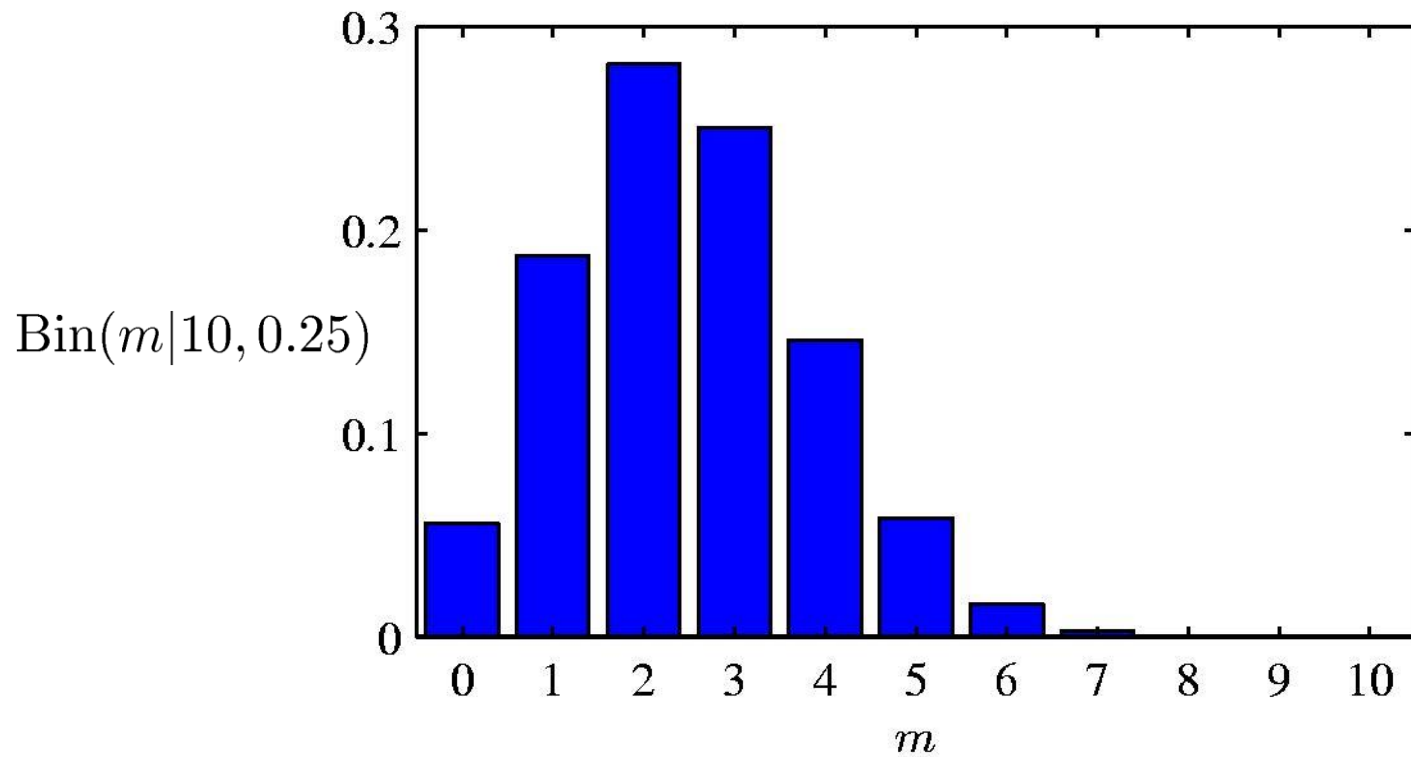
Binomial Distribution

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

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

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

Binomial Distribution



Parameter Estimation (1)

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \dots, x_N\}$, m heads (1), $N - m$ tails (0)

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

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

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

Parameter Estimation (2)

Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\text{ML}} = \frac{3}{3} = 1$

Prediction: *all* future tosses will land heads up

Overfitting to \mathcal{D}

Beta Distribution

Distribution over $\mu \in [0, 1]$.

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

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

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

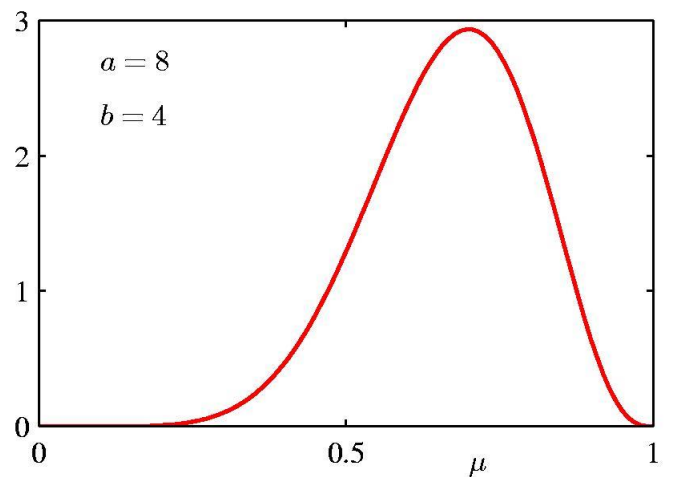
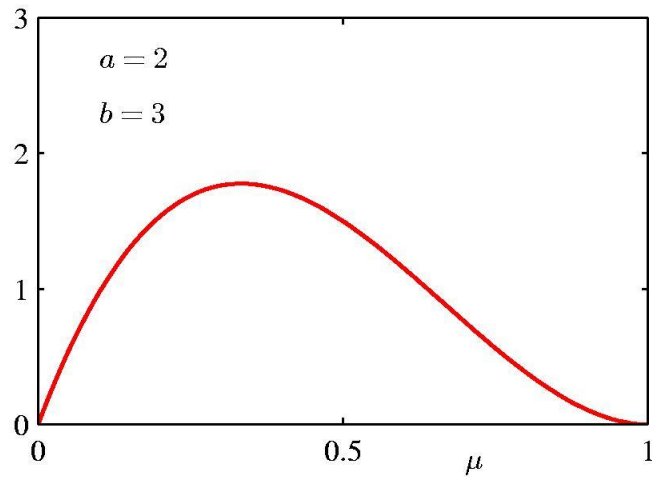
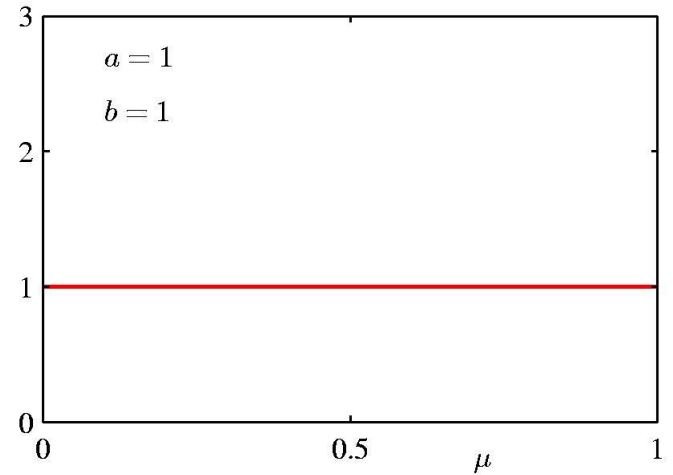
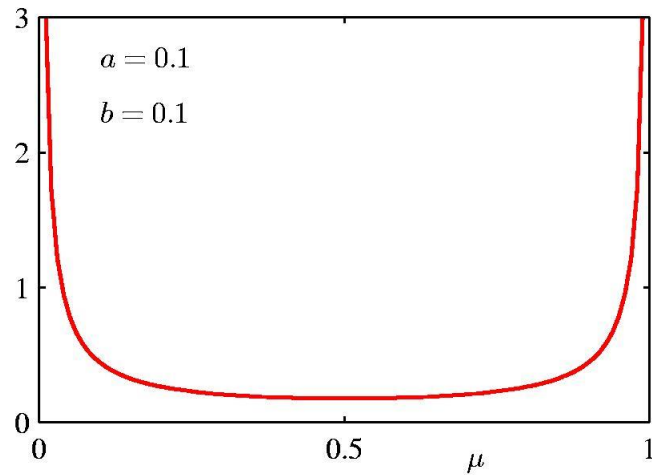
Bayesian Bernoulli

$$\begin{aligned} p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\ &= \left(\prod_{n=1}^N \mu^{x_n} (1 - \mu)^{1-x_n} \right) \text{Beta}(\mu|a_0, b_0) \\ &\propto \mu^{m+a_0-1} (1 - \mu)^{(N-m)+b_0-1} \\ &\propto \text{Beta}(\mu|a_N, b_N) \end{aligned}$$

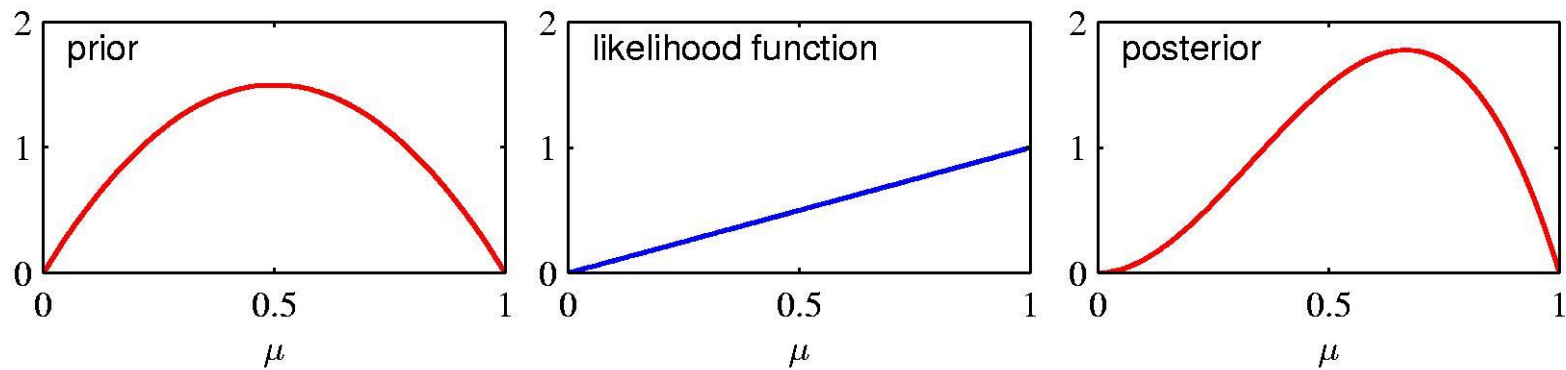
$$a_N = a_0 + m \quad b_N = b_0 + (N - m)$$

The Beta distribution provides the *conjugate* prior for the Bernoulli distribution.

Beta Distribution



Prior · Likelihood = Posterior



Properties of the Posterior

As the size of the data set, N , increase

$$a_N \rightarrow m$$

$$b_N \rightarrow N - m$$

$$\mathbb{E}[\mu] = \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\text{ML}}$$

$$\text{var}[\mu] = \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0$$

Prediction under the Posterior

What is the probability that the next coin toss will land heads up?

$$\begin{aligned} p(x = 1|a_0, b_0, \mathcal{D}) &= \int_0^1 p(x = 1|\mu)p(\mu|a_0, b_0, \mathcal{D}) \, \mathrm{d}\mu \\ &= \int_0^1 \mu p(\mu|a_0, b_0, \mathcal{D}) \, \mathrm{d}\mu \\ &= \mathbb{E}[\mu|a_0, b_0, \mathcal{D}] = \frac{a_N}{a_N + bN} \end{aligned}$$

An Example

	Prior	Data	Posterior
Total #	100	3	103
Head #	50	3	53
Tail #	50		50

The probability that the next coin toss will land heads up is $53/103$ 。

Outlines

- Binary Distributions
 - Multinomial Distributions
 - Gaussian Distributions
 - Exponential Families
 - Non-informative Priors
 - Non-parametric Methods
 - KNN
-

Multinomial Variables

1-of- K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$

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

$$\forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^K \mu_k = 1$$

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

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

ML Parameter estimation

Given: $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$

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

Ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier, λ .

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

$$\mu_k = -m_k / \lambda \quad \mu_k^{\text{ML}} = \frac{m_k}{N}$$

The Multinomial Distribution

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

$$\mathbb{E}[m_k] = N\mu_k$$

$$\text{var}[m_k] = N\mu_k(1 - \mu_k)$$

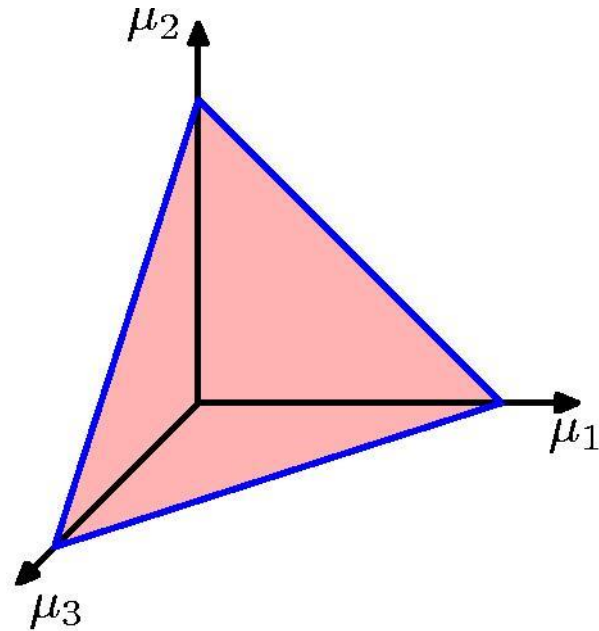
$$\text{cov}[m_j, m_k] = -N\mu_j\mu_k$$

The Dirichlet Distribution

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

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

Conjugate prior for the multinomial distribution.

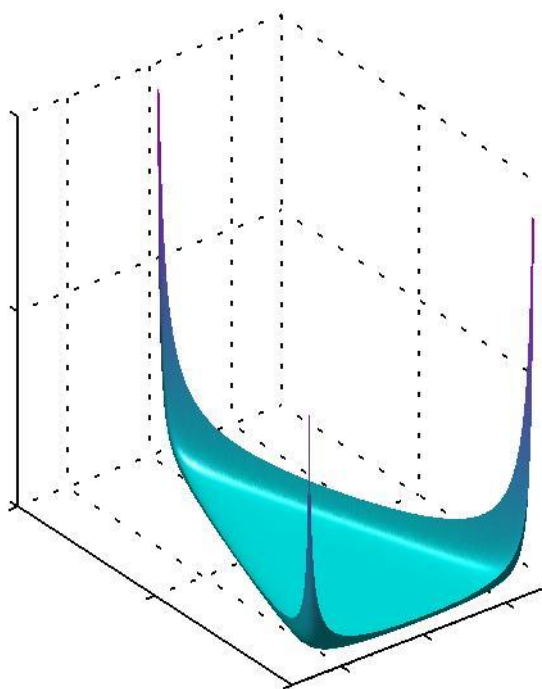


Bayesian Multinomial (1)

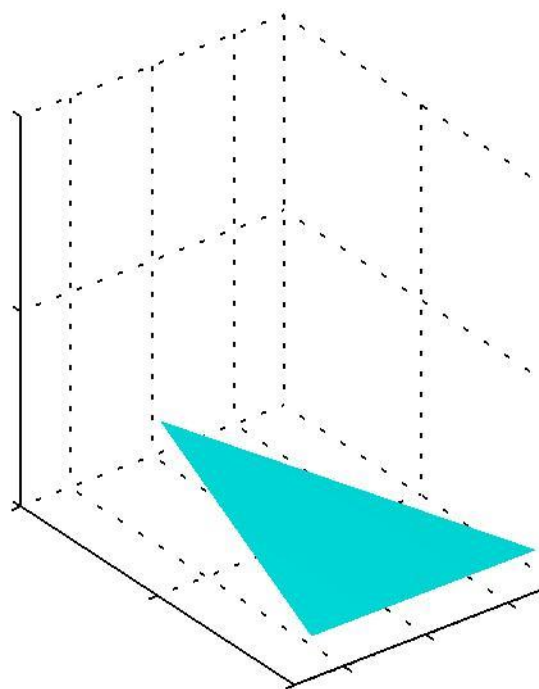
$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k + m_k - 1}$$

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

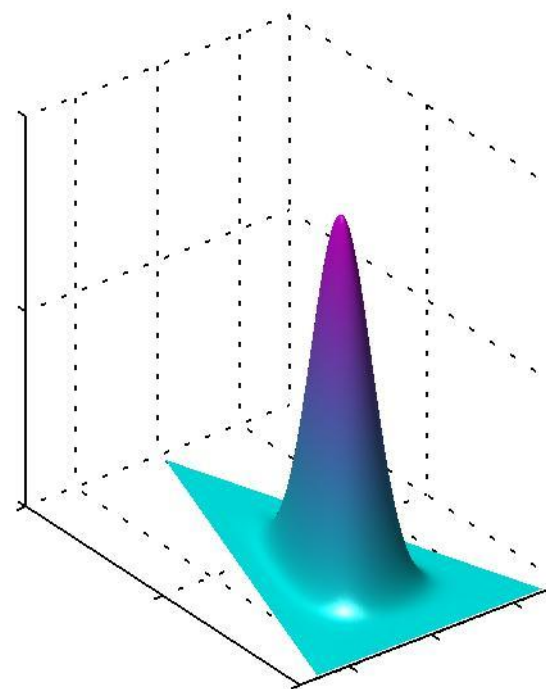
Bayesian Multinomial (2)



$$\alpha_k = 10^{-1}$$



$$\alpha_k = 10^0$$



$$\alpha_k = 10^1$$
