Lecture-4
Density estimation

CS 277:
Machine Learning
and
Data Science

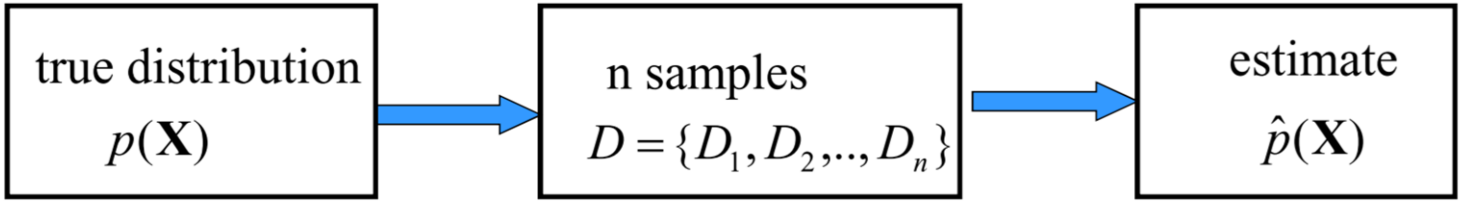
By:
Dr. Joydeep Chandra
Associate Professor
Dept. of CSE, IIT Patna

Density estimation

Data: $D = \{ D_1, D_2, .., D_n \}$

$D_i = \boldsymbol{x}_i$ a vector of attribute values

Objective: estimate the underlying probability distribution over variables **X** , $p(\mathbf{X})$, using examples in D



Standard (iid) assumptions:?

Parametric density estimation

Parametric density estimation:?

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$

Parametric density estimation

Parametric density estimation:?

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in **X** with parameters
- **Data** $D = \{D_1, D_2, \ldots, D_n\}$

Objective: find parameters such that $p(\mathbf{X}|\mathbf{\Theta})$ fits data D the best $\Theta : \hat{p}(\mathbf{X} \mid \Theta)$

Parameter estimation (learning)

- **Maximum likelihood (ML)**

$$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi)$$

- **Bayesian parameter estimation**

keep the posterior density $p(\Theta \mid D, \xi)$

- **Maximum a posteriori probability (MAP)**

$$\Theta_{MAP} = \arg\max_{\Theta} p(\Theta \mid D, \xi)$$

- **Expected value**

$$\Theta_{EXP} = \int_{\Theta} \Theta p(\Theta \mid D, \xi) d\Theta$$

Parameter estimation: Coin example.

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$ from data

Parameter estimation: Example

Assume the unknown and possibly biased coin

- Probability of the head is Θ
- **Data:**

H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What would be your estimate of the probability of a head ?

$$\tilde{\theta} = ?$$

Parameter estimation: Example

Assume the unknown and possibly biased coin

- Probability of the head is Θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What would be your estimate of the probability of a head ?

Solution: use frequencies of occurrences to do the estimate

$$\tilde{\theta} = \frac{15}{25} = 0.6$$

This is **the maximum likelihood estimate** of the parameter Θ

Probability of an outcome

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ

probability of a tail $(1-\theta)$

Assume: we know the probability θ

Probability of an outcome of a coin flip x_i

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$ ⟵ **Bernoulli distribution**

- Combines the probability of a head and a tail
- So that x_i is going to pick its correct probability
- Gives θ for $x_i = 1$
- Gives $(1-\theta)$ for $x_i = 0$

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H (encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = ?$$

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H encoded as D= 110101

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

likelihood of the data

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

$$P(D \mid \theta) = \prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

Can be rewritten using the Bernoulli distribution:

The goodness of fit to the data

Learning: **we do not know the value of the parameter** θ

Our learning goal:

- Find the parameter θ that fits the data D the best?

One solution to the "best": Maximize the likelihood

$$P(D \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Intuition:

- more likely are the data given the model, the better is the fit

Note: Instead of an error function that measures how bad the data fit the model we have a measure that tells us how well the data fit :

$$Error\ (D, \theta) = -P(D \mid \theta)$$

Example: Bernoulli distribution

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ

probability of a tail $(1-\theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$

Probability of an outcome x_i

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$ **Bernoulli distribution**

Maximum likelihood (ML) estimate

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Maximum likelihood estimate

$$\theta_{ML} = \arg\max_{\theta} P(D \mid \theta, \xi)$$

Optimize log-likelihood (the same as maximizing likelihood)

$$l(D,\theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} =$$

$$\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta) = \log\theta \underbrace{\sum_{i=1}^{n} x_i}_{N_1} + \log(1-\theta) \underbrace{\sum_{i=1}^{n} (1-x_i)}_{N_2}$$

N_1 - number of heads seen $\quad$ N_2 - number of tails seen

Maximum likelihood (ML) estimate

Optimize log-likelihood

$$l(D,\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

Set derivative to zero

$$\frac{\partial l(D,\theta)}{\partial\theta} = \frac{N_1}{\theta} - \frac{N_2}{(1-\theta)} = 0$$

Solving $$\theta = \frac{N_1}{N_1 + N_2}$$

ML Solution: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

Maximum likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Maximum likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Head: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$$

Tail: $$(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$

Maximum a posteriori estimate

Maximum a posteriori estimate

– Selects the mode of the **posterior distribution**

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$$

Likelihood of data

prior

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

Normalizing factor

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

$p(\theta \mid \xi)$ - is the prior probability on θ

How to choose the prior probability?

Why posterior?

It provides a natural and principled way of combining prior information with data, within a solid decision theoretical framework.

One can incorporate past information about a parameter and form a prior distribution for future analysis.

Prior distribution

Choice of prior: Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

$\Gamma(x)$ - a Gamma function $\Gamma(x) = (x-1)\Gamma(x-1)$

For integer values of x $\Gamma(n) = (n-1)!$

Why to use Beta distribution?

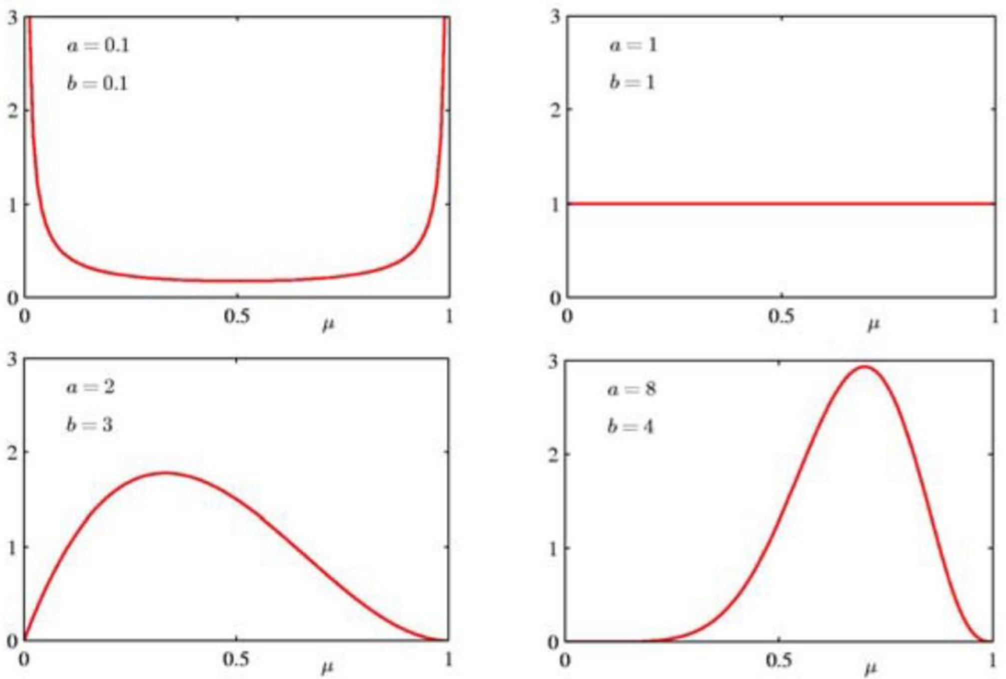
Beta distribution "**fits**" Bernoulli trials - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution

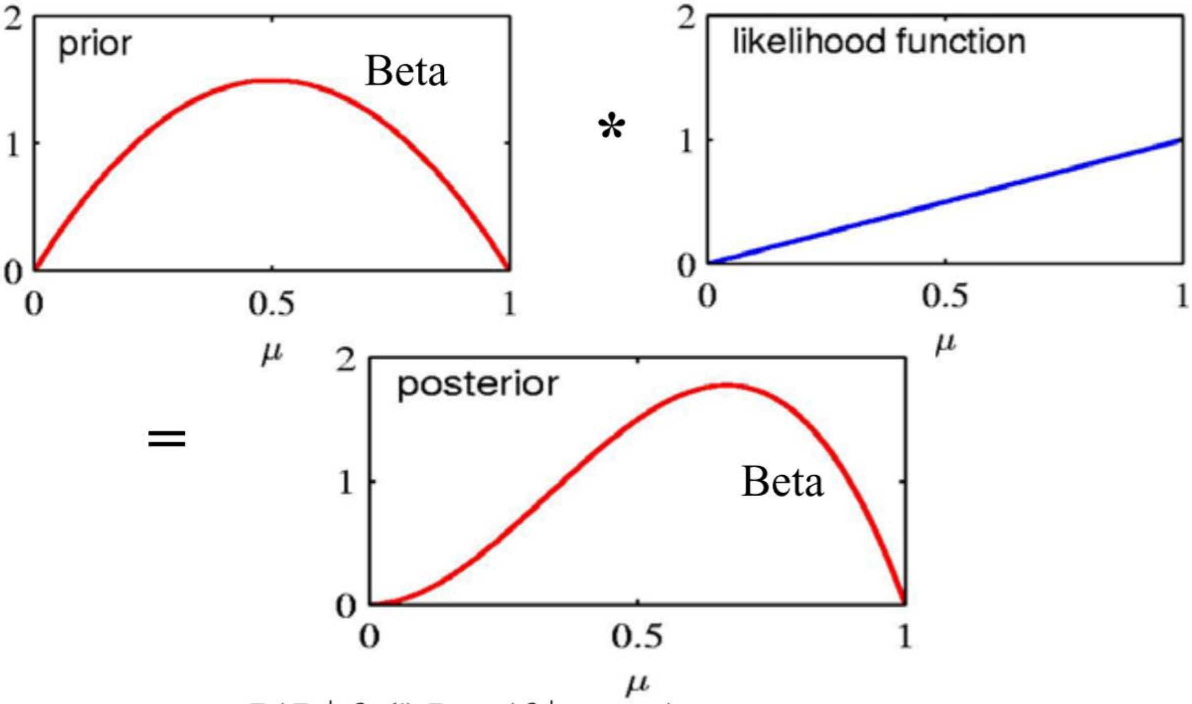
$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Beta distribution



$$p(\theta \mid \xi) = Beta(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

Posterior distribution



$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi) Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Maximum a posterior probability

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi) Beta(\theta \mid \alpha_1,\alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1+\alpha_1-1}(1-\theta)^{N_2+\alpha_2-1}$$

Notice that parameters of the prior
act like counts of heads and tails
(sometimes they are also referred to as **prior counts**)

MAP Solution: $$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

MAP estimate example

Assume the unknown and possibly biased coin

- Probability of the head is
- **Data:**

H H T T H H T H T H T T T H T H H H H T H H H H T

- **Heads:** 15
- **Tails:** 10

• Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5)$$

What is the MAP estimate?

MAP estimate example

Assume the unknown and possibly biased coin

- Probability of the head is
- **Data:**

H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

- Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5)$$

What is the MAP estimate?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- **The MAP can be biased with large prior counts**
- **It is hard to overturn it with a smaller sample size**
- **Data:** H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

Assume:

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5) \qquad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta \mid \xi) = Beta(\theta \mid 5,20) \qquad \theta_{MAP} = \frac{19}{48}$$

Bayesian framework

Both ML or MAP estimates pick one value of the parameter

- **Assume:** there are two different parameter settings that are close in terms of their probability values. Using only one of them may introduce a strong bias, if we use them, for example, for predictions.

Bayesian parameter estimate

- Remedies the limitation of one choice
- Keeps all possible parameter values
- Where $p(\theta \mid D,\xi) \approx Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$

The posterior can be used to define $p(A \mid D)$:

$$p(A \mid D) = \int_{\Theta} p(A \mid \Theta)\, p(\Theta \mid D,\xi)\, d\Theta$$

Bayesian framework

- **Predictive probability of an outcome $x=1$ in the next trial**

$$P(x=1|D,\xi)$$

$$P(x=1|D,\xi)=\int_0^1 P(x=1|\theta,\xi)\overbrace{p(\theta|D,\xi)}^{\text{Posterior density}}d\theta$$

$$=\int_0^1 \theta p(\theta|D,\xi)d\theta=E(\theta)$$

- **Equivalent to the expected value of the parameter**
  - expectation is taken with respect to the posterior distribution

$$p(\theta|D,\xi)=Beta(\theta|\alpha_1+N_1,\alpha_2+N_2)$$

Expected value of the parameter

How to obtain the expected value?

$$E(\theta) = \int_0^1 \theta Beta(\theta \mid \eta_1, \eta_2) d\theta = \int_0^1 \theta \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \theta^{\eta_1-1}(1-\theta)^{\eta_2-1} d\theta$$

$$= \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \int_0^1 \theta^{\eta_1}(1-\theta)^{\eta_2-1} d\theta$$

$$= \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \frac{\Gamma(\eta_1+1)\Gamma(\eta_2)}{\Gamma(\eta_1+\eta_2+1)} \underbrace{\int_0^1 Beta(\eta_1+1, \eta_2) d\theta}_{1}$$

$$= \frac{\eta_1}{\eta_1+\eta_2}$$

Note: $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for integer values of α

Expected value of the parameter

- **Substituting the results for the posterior:**

$$p(\theta \mid D, \xi) = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

- **We get** $$E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

- **Note that the mean of the posterior is yet** another "reasonable" parameter choice:

$$\hat{\theta} = E(\theta)$$

Binomial distribution

Example problem: a biased coin

Outcomes: two possible values -- head or tail

Data: a set of order-independent outcomes for N trials

N1 - number of heads seen *N2* - number of tails seen

Model: probability of a head ***Θ***

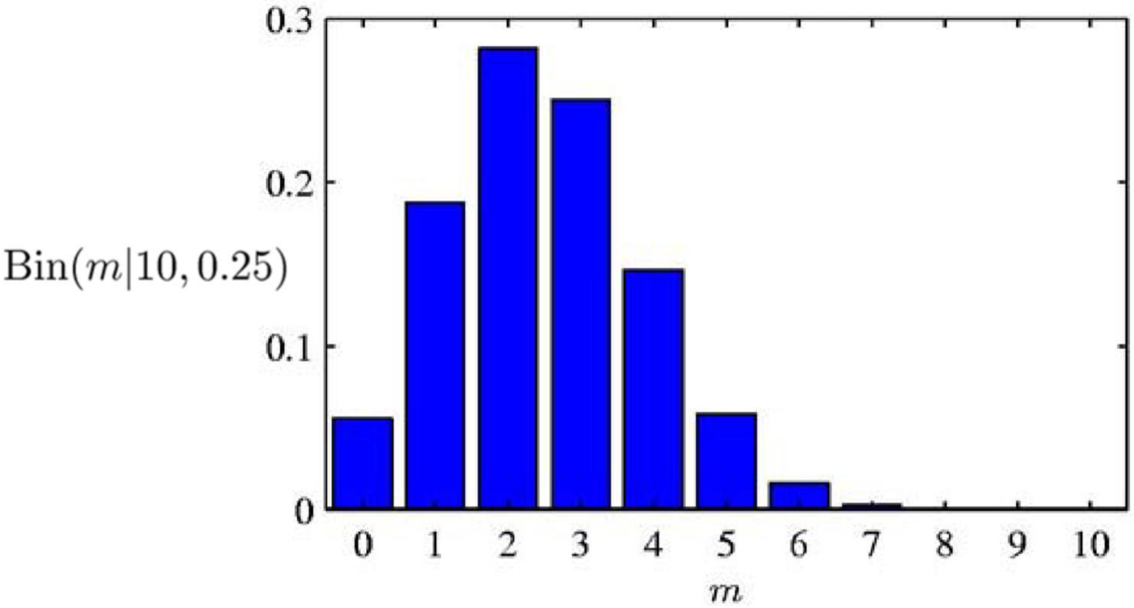
probability of a tail *(1-**Θ**)*

Probability of an outcome $P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$ **Binomial distribution**

Objective: We would like to estimate the probability of a **head** $\hat{\theta}$

Binomial distribution

Binomial distribution:



Maximum likelihood (ML) estimate

Likelihood of data:

$$P(D\mid\theta)=\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_2}=\frac{N!}{N_1!N_2!}\theta^{N_1}(1-\theta)^{N_2}$$

Log-likelihood

$$l(D,\theta)=\log\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_2}=\log\frac{N!}{N_1!N_2!}+N_1\log\theta+N_2\log(1-\theta)$$

Constant from the point of optimization !!!

ML Solution: $\theta_{ML}=\frac{N_1}{N}=\frac{N_1}{N_1+N_2}$

The same as for Bernoulli and D with iid sequence of examples

Posterior density

Posterior density

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)} \quad \textbf{(via Bayes rule)}$$

Prior choice

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

Likelihood

$$P(D \mid \theta) = \frac{\Gamma(N_1 + N_2)}{\Gamma(N_1)\Gamma(N_2)} \theta^{N_1} (1-\theta)^{N_2}$$

Posterior $p(\theta \mid D, \xi) = Beta(\alpha_1 + N_1, \alpha_2 + N_2)$

MAP estimate $\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$

$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

Expected value of the parameter

The result is the same as for Bernoulli distribution

$$E(\theta) = \int_0^1 \theta Beta(\theta \mid \eta_1, \eta_2) d\theta = \frac{\eta_1}{\eta_1 + \eta_2}$$

Expected value of the parameter

$$E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

Predictive probability of event x=1

$$P(x = 1 \mid \theta, \xi) = E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

Multinomial Distribution

Example: Multi-way coin toss, roll of dice

- **Data:** a set of N outcomes (multi-set)

  N_i - a number of times an outcome i has been seen

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$ s.t. $\sum_{i=1}^{k} \theta_i = 1$

θ_i - probability of an outcome i

Probability of data (likelihood)

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

Example: Consider a three-way election for a large country. If 6 voters are selected randomly, what is the probability that there will be exactly one supporter for candidate A, two supporters for candidate B and three supporters for candidate C in the sample, assuming $\theta_1, \theta_2, \theta_3$ are the probabilities of voting for candidates A, B, and C.

Posterior Density and MAP estimate

Choice of the prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1, \dots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for multinomial

$$P(D \mid \boldsymbol{\theta}, \xi) = P(N_1, N_2, \dots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \dots N_k!} \theta_1^{N_1} \theta_2^{N_2} \dots \theta_k^{N_k}$$

Posterior density

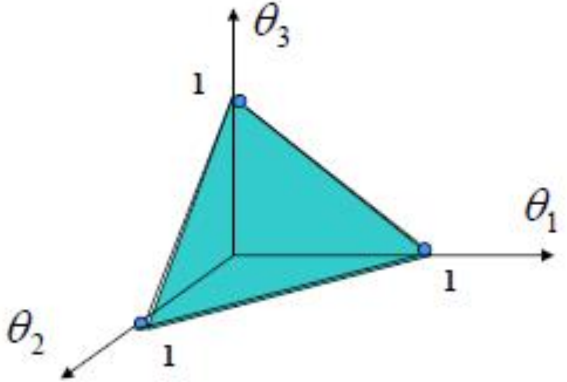
$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, .. \alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, \dots, \alpha_k + N_k)$$

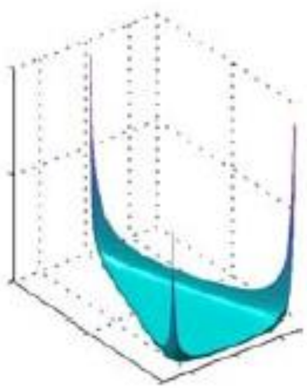
MAP estimate: $\theta_{i,MAP} = \dfrac{\alpha_i + N_i - 1}{\sum_{i=1}^{k} (\alpha_i + N_i) - k}$

Dirichlet Distribution

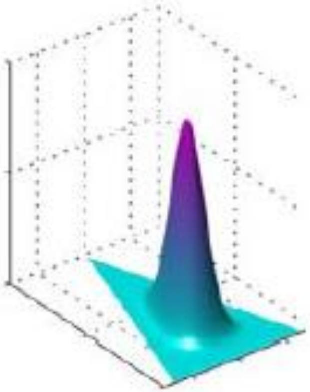
$$Dir(\boldsymbol{\theta} \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Assume: k=3





$\alpha_k = 10^{-1}$



$\alpha_k = 10^{1}$

Expected value

The result is analogous to the result for binomial

$$E(\boldsymbol{\theta}) = \int_{0 \le \theta_i \le 1, \sum \theta_i = 1} \boldsymbol{\theta} \; Dir(\boldsymbol{\theta} \mid \boldsymbol{\eta}) d\boldsymbol{\theta} = \left(\frac{\eta_1}{\eta_1 + \eta_2 + \eta_k}, \ldots \frac{\eta_i}{\eta_1 + \eta_2 + \eta_k}, \ldots \frac{\eta_k}{\eta_1 + \eta_2 + \eta_k} \right)$$

Expectation based parameter estimate

$$E(\boldsymbol{\theta}) = \left(\frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k} \ldots \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k} \ldots \frac{\alpha_k + N_k}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k} \right)$$

Represents the predictive probability of an event x=i

$$P(x = i \mid \boldsymbol{\theta}, \xi) = \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k}$$

Other distributions

The same ideas can be applied to other distributions

- Typically we choose distributions that behave well so that computations lead to a nice solutions

- **Exponential family of distributions**

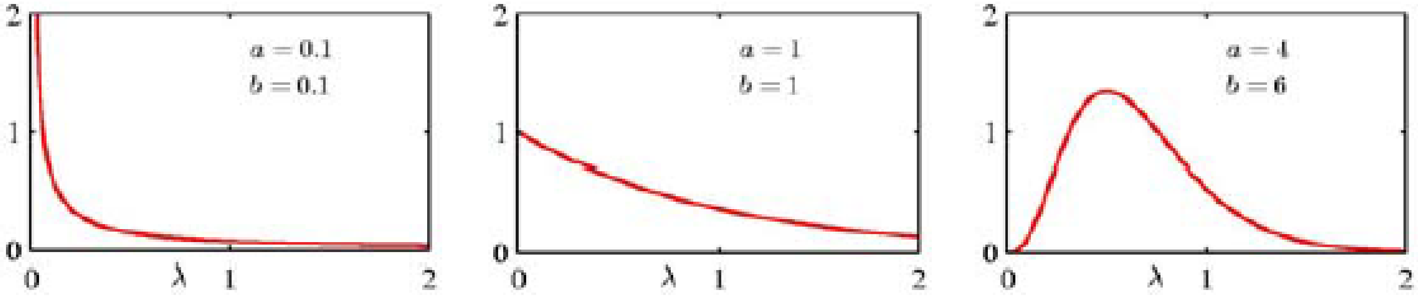
Conjugate choices for some of the distributions from the exponential family:

- **Binomial – Beta**
- **Multinomial - Dirichlet**
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

Gamma Distribution

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad\qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



Exponential and Poisson

Exponential distribution:

- A special case of Gamma for a=1

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}}$$

Poisson distribution:

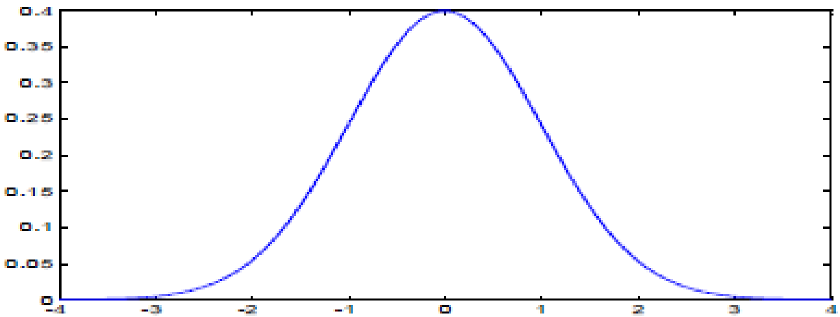
$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^{x}}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$

Gaussian (Normal) Distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean
  σ - standard deviation
- **Density function:**

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

- **Example:**



$N(0,1)$

Parameter estimates

- **Loglikelihood** $$l(D,\mu,\sigma) = \log \prod_{i=1}^{n} p(x_i \mid \mu,\sigma)$$
- **ML estimates of the mean and variance:**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

  – ML variance estimate is biased

$$E_n(\sigma^2) = E_n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- **Unbiased estimate:**

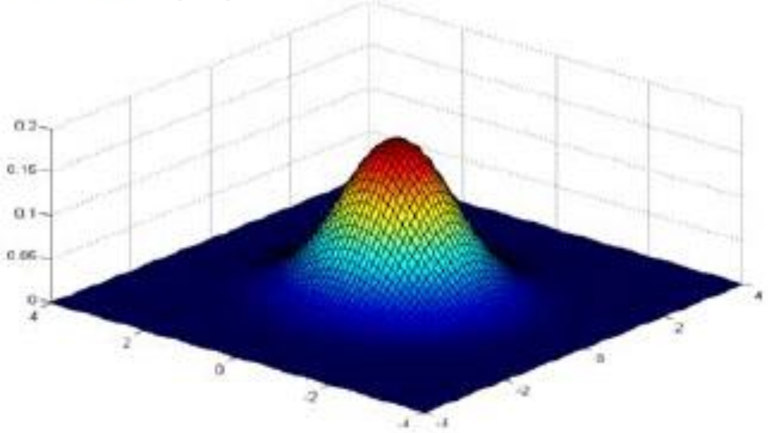
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Multivariate Normal Distribution

- **Multivariate normal:** $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Parameters:** $\boldsymbol{\mu}$ - mean
  $\boldsymbol{\Sigma}$ - covariance matrix
- **Density function:**

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

- **Example:**



Any Questions??