maximum a posteriori

maximum likelihood

Machine Learning

More on distributions, models, MAP, ML

SLIDES BY Stefan Harmeling

~~27~~. October 202~~1~~

19. 22

Last time:

Gaussian distribution

Univariate Gaussian distribution

see MLPP 2.4.1 (Murphy: Machine Learning: a Probabilistic Perspective)

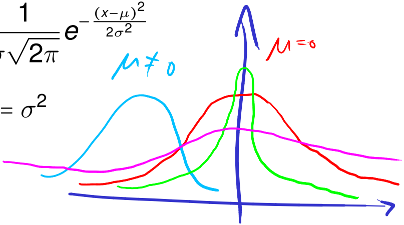
- random variable X is real-valued
- parameters μ called mean, $\sigma^2 > 0$ called variance
- X has univariate Gaussian distribution, written

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

- probability density function

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- one can show: $\mathrm{E}\,X = \mu$ and $\mathrm{Var}\,X = \sigma^2$



Multivariate Gaussian distribution

see MLPP 2.5.2

- random vector X has real-valued components
- parameters μ called mean vector, pos-def symmetric matrix Σ called covariance matrix
- X has multivariate Gaussian distribution, written

$$X \sim \mathcal{N}(\mu, \Sigma)$$

- probability density function

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

- special case: $\mathcal{N}(\mu, \sigma^2)$
- one can show: $\mathrm{E}\, X = \mu$ and $\mathrm{Var}\, X = \Sigma$

Closed under sum- and product rule:

$\Omega = A \cup \bar{A}$ $\quad$ $P(B) = P(B, A) + P(B, \bar{A})$

how to compute marginals

A Gaussian joint distribution

$$p(x, y) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} \mu \\ \nu \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \right)$$

has Gaussian marginals (sum rule)

$$p(x) = \int p(x, y)\, dy = \mathcal{N}(x, \mu, A)$$

$$p(y) = \int p(x, y)\, dx = \mathcal{N}(y, \nu, C)$$

and Gaussian conditionals

product rule

$$p(x \mid y) = p(x, y)/p(y) = \mathcal{N}(x, \mu + BC^{-1}(y - \nu), A - BC^{-1}B^T)$$

$$p(y \mid x) = p(x, y)/p(x) = \mathcal{N}(y, \nu + B^T A^{-1}(x - \mu), C - B^T A^{-1} B)$$

Main justification for using Gaussians

Central Limit theorem

If you add lots of i.i.d independent random var.s $X_1, \dots, X_n$
then their sum $X_1 + \dots + X_n$ is roughly Gaussian

$\rightsquigarrow$ model noise by Gaussians.

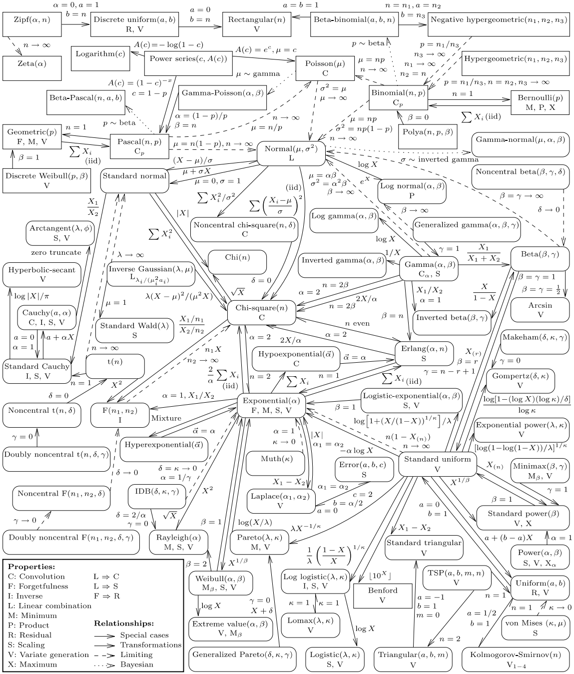


Figure 1. Univariate distribution relationships.

previous graphics from: “Univariate Distribution Relationships”, Lawrence M. Leemis and Jacquelyn T. McQueston, The American Statistician, February 2008, Vol. 62, No. 1, page 47

A zoo of probability distributions

Distribution for waiting times

Poisson distribution

see MLPP 2.3.3

- counts of rare events
- let random variable $X \in \{0, 1, \ldots\}$ be the number of events in some time interval
- let $\lambda > 0$ be the parameter (the rate)
- X has Poisson distribution, written

$$X \sim \mathrm{Poi}(\lambda)$$

- probability mass function

$$\mathrm{Poi}(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

- $\mathrm{E}\, X = \mathrm{Var}\, X = \lambda$
- e.g. number of emails you receive every days is Poisson distributed
- e.g. the waiting time between events

Distributions for tossing dice

Binomial distribution

see MLPP 2.3.1

- toss a coin n times
- let random variable $X \in \{0, \dots, n\}$ be number of heads
- let θ be the probability of heads
- X has binomial distribution, written

$$X \sim \mathrm{Bin}(n, \theta)$$

- probability mass function

$$\mathrm{Bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

- $\mathrm{E}\, X = n\theta$, $\mathrm{Var}\, X = n\theta(1 - \theta)$

Bernoulli distribution

see MLPP 2.3.1

- toss a coin once
- let random variable $X \in \{0, 1\}$ be a binary variable
- let θ be the probability of heads
- X has Bernoulli distribution, written

$$X \sim \mathrm{Ber}(\theta)$$

- probability mass function

$$\mathrm{Ber}(x \mid \theta) = \theta^{[x=1]}(1-\theta)^{[x=0]} = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

  using Iverson brackets $[A] = 1$ if A is true, $[A] = 0$ if A is false
- $\mathrm{E}\,X = \theta$, $\mathrm{Var}\,X = \theta(1-\theta)$
- special case: $\mathrm{Ber}(\theta) = \mathrm{Bin}(1, \theta)$

Multinomial distribution

MURPHY

see MLPP 2.3.2 ← Machine Learning, A probabilistic perspective

- toss a K-sided dice n times
- let $X = [x_1, \ldots, x_K]^T$ be a random (column) vector, with x_j being the number of times side j occurs, $\sum_j x_j = n$
- let $\theta = [\theta_1, \ldots, \theta_K]^T$ be the parameter (column) vector, with $\sum_j \theta_j = 1$ and $\theta_j \geq 0$
- let θ_j be the probability of side j of the dice
- X has multinomial distribution, written

$$X \sim \mathrm{Mu}(n, \theta)$$

- probability mass function

$$\mathrm{Mu}(x \mid n, \theta) = \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

with multinomial coefficient $\binom{n}{x_1 \ldots x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$

Multinoulli distribution

see MLPP 2.3.2

- toss a K-sided dice once
- let $X = (x_1, \ldots, x_K)$ be a random vector, with x_j being binary, such that only one is non-zero (aka one-hot encoding)
- let $\theta = (\theta_1, \ldots, \theta_K)$ be the parameter vector, with $\sum_j \theta_j = 1$ and $\theta_j \geq 0$
- let θ_j be the probability of side j of the dice
- X has multinoulli distribution, written

$$X \sim \mathrm{Cat}(\theta) = \mathrm{Mu}(1, \theta)$$

- probability mass function

$$\mathrm{Cat}(x \mid \theta) = \prod_{j=1}^{K} \theta_j^{x_j}$$

- aka categorical distribution or discrete distribution

Tossing dice (1)

- tossing n times a K sided dice
- let X be random vector of number of times side j appeared
- distribution of X: Multinomial

$$X \sim \mathrm{Mu}(n, \theta)$$

  with parameter vector θ
- assume $n = 1$: Multinoulli

$$\mathrm{Cat}(\theta) = \mathrm{Mu}(1, \theta)$$

- assume case $K = 2$: Binomial

$$\mathrm{Bin}(n, \theta) = \mathrm{Mu}(n, (\theta, 1 - \theta))$$

  with $\theta \in [0, 1]$
- assume $n = 1$ and $K = 2$: Bernoulli

$$\mathrm{Ber}(\theta) = \mathrm{Bin}(1, \theta) = \mathrm{Mu}(1, (\theta, 1 - \theta)) = \mathrm{Cat}((\theta, 1 - \theta))$$

  with $\theta \in [0, 1]$

Tossing dice (2)

- tossing n times a K sided dice

	$n = 1$	$n > 1$
$K = 2$	Bernoulli	Binomial
$K > 2$	Multinoulli	Multinomial

Probability Theory:

Describe mathematically how random processes generate data

Statistics:

Given data, try to find the probability distribution that best explains it.

Maximum likelihood estimation

Typically we understand data as having been obtained from repetitions of the same experiment

[" samples from a single probability distribution"]

Each single result is recorded in a random variable.

E.g. n coin tosses: $X_1, \dots, X_n \in \{0,1\}$

Standard assumption:

The different instances of the experiment don't influence each other.

independence

Each time the experiment is set up in precisely the same way.

ident. distributed

Formally:

$X_1, \dots, X_n$ are i.i.d.

[independent & identically distributed]

$X_1, \dots, X_n$ are i.i.d.

$\Rightarrow$ Their joint distribution is

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$

$X_1, \dots, X_n$ are i.i.d.

$\Longrightarrow$ Their joint distribution is

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$

Usually we have a family of candidate distributions, parametrized by some parameter θ [= a „statistical model"]

$X_1, \ldots, X_n$ are i.i.d.

$\Rightarrow$ Their joint distribution is

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$

Usually we have a family of candidate distributions, parametrized by some parameter θ [= a „statistical model"]

$\rightsquigarrow$ For each θ we have $P_\theta(X_1 = x_1, \ldots, X_n = x_n)$; we also write $P(X_1 = x_1, \ldots, X_n = x_n \mid \theta)$.

$X_1, \dots, X_n$ are i.i.d.

$\Longrightarrow$ Their joint distribution is

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$

Usually we have a family of candidate distributions, parametrized by some parameter θ [= a „statistical model"]

a function of $x_1, \dots, x_n$ and θ

$\rightsquigarrow$ For each θ we have $P_\theta(X_1 = x_1, \dots, X_n = x_n)$; we also write $P(X_1 = x_1, \dots, X_n = x_n \mid \theta)$.

$X_1, \dots, X_n$ are i.i.d.

$\Rightarrow$ Their joint distribution is

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$

Usually we have a family of candidate distributions, parametrized by some parameter θ [= a „statistical model"]

a function of $x_1, \dots, x_n$ and θ

$\rightsquigarrow$ For each θ we have $P_\theta(X_1 = x_1, \dots, X_n = x_n)$; we also write $P(X_1 = x_1, \dots, X_n = x_n \mid \theta)$.

Which θ explains the observed data best?

Max. likelihood estimation: The θ that maximizes $P(X_1 = x_1, \dots, X_n = x_n \mid \theta)$.

Max. likelihood estimation: The θ that maximizes

$$P(X_1 = x_1, \dots, X_n = x_n \mid \theta).$$

Fixed θ, varying $x_1, \dots, x_n$: „Probability distribution"

(„prob. mass function" or „density function")
discrete continuous

Fixed $x_1, \dots, x_n$, varying θ: „Likelihood function"

$\rightsquigarrow$ want to maximize likelihood function

Example:

Thumbtack falls on pin $(X=0)$
or on head $(X=1)$

thumbtack

We want to find out the probability that the thumbtack falls on its pin.

Statistical model: $P(X=0\mid\theta) = \theta$
$P(X=1\mid\theta) = 1-\theta$

$$\theta \in [0,1]$$

__Example__:

Thumbtack falls on pin $(X=0)$
or on head $(X=1)$

thumbtack

We want to find out the probability that the thumbtack falls on its pin.

Statistical model: $P(X=0|\theta)=\theta$
$P(X=1|\theta)=1-\theta$

Observe: $8\times$ head, $2\times$ pin

$$P(X_1=0, X_2=1, \dots, X_9=1, X_{10}=0 \,|\, \theta)$$
$$= P(X_1=0|\theta)\cdot \dots \cdot P(X_{10}=0|\theta)$$
$$= \theta^2\cdot(1-\theta)^8$$

Maximize

$$L(\theta; x_1, \dots, x_n) = P(X_1 = 0, X_2 = 1, \dots, X_9 = 1, X_{10} = 0 \mid \theta)$$

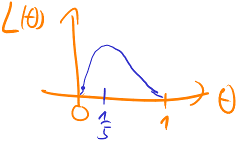
$$= P(X_1 = 0 \mid \theta) \cdot \dots \cdot P(X_{10} = 0 \mid \theta)$$

$$= \theta^2 \cdot (1-\theta)^8$$

$$0 \overset{!}{=} \frac{\partial}{\partial \theta} L(\theta, x_1, \dots, x_n) = 2 \cdot \theta (1-\theta)^8 - \theta^2 \cdot (1-\theta)^7 \cdot 8$$

$$= \theta \cdot (1-\theta)^7 \cdot (2 \cdot (1-\theta) - 8 \cdot \theta)$$

$$\Rightarrow \theta = 0, \text{ or } \theta = 1, \text{ or } \underbrace{2 - 2\theta - 8\theta}_{= \, 2 - 10\theta} = 0 \Rightarrow \theta = \frac{1}{5}$$



$$\hat{\theta}_{ML} := \text{"} \operatorname*{argmax}_{\theta} L(\theta; x_1, \dots, x_n) \text{"}$$

Example: Univariate Gaussian, mean μ known, family parametrized by variance σ:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

observations: $x_1, \dots, x_m$

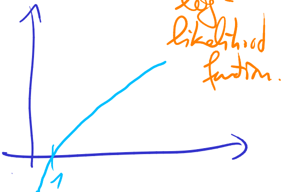
$$L(\sigma \mid x_i, \mu) = \prod_{x_i} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

maximizing $L(\sigma)$ is hard but maximizing $\log(L(\sigma)) = \ell(\sigma)$

$$\ell(\sigma) = \log(L(\sigma))$$

$$= \sum_{x_i} \overbrace{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}^{-\log(\sigma) - \log(\sqrt{2\pi})} - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \underbrace{m \cdot (-\log(\sqrt{2\pi}))} + m \cdot (-\log(\sigma)) - \sum_{x_i} \frac{1}{\sigma^2}(x_i-\mu)^2 \frac{1}{2}$$



$$0 \overset{!}{=} \frac{\partial}{\partial \sigma} l(\sigma) = -m \cdot \frac{1}{\sigma} - \sum_{x_i} (x_i - \mu)^2 \cdot \frac{1}{2} \frac{1}{\sigma^3} \cdot (-2)$$

$$= -m \cdot \frac{1}{\sigma} + \sum_{x_i} (x_i - \mu)^2 \frac{1}{\sigma^3}$$

$$\Rightarrow \quad \sigma^2 \cdot m = \sum_{x_i} (x_i - \mu)^2$$

$$\Rightarrow \quad \sigma^2 = \frac{1}{m} \sum_{x_i} (x_i - \mu)^2$$

$$Var(X) = E((X - EX)^2)$$

Maximum a posteriori estimation

Idea: Use Bayes rule

$$\underbrace{P(\Theta \mid data)}_{\text{posterior}} = \frac{P(data \mid \theta) \cdot \overbrace{P(\Theta)}^{\text{prior}}}{P(data)}$$

What is the most probable Θ, given the seen data?

That one is called max. a posteriori estimator

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} P(\Theta \mid data)$$

$$= \arg\max_{\Theta} \underbrace{P(data \mid \theta) \cdot P(\theta)}_{\text{known}}$$

Observe: Maximizing numerator means maximizing everything

In a bag there are 5 coins. Two of the coins are of type A and 3 of type B. A coin of type A shows heads with probability 0.5 and tails with probability 0.5. A coin of type B shows heads with probability 0.3 and tails with probability 0.7.

A coin is randomly drawn from the bag and then flipped two times: It shows heads both times.

→ Use Bayes' formula to compute the probability that the coin is of type A.

→ The coin is thrown one more time and shows tails. Given this additional data, what is now the probability that the coin is of type A? [You can either reuse your results from (a) or start a whole new computation. Or better, do both to convince yourself that both ways give the same result]

Here $\Theta = \{A, B\}$ prior distr on $\{A, B\}$: $\frac{2}{5}$, $\frac{3}{5}$

$$P(X=H \mid A) = 0.5$$

$$P(X=H \mid B) = 0.3$$

$$P(H,H \mid A) = \frac{1}{2} \cdot \frac{1}{2}, \quad P(H,H \mid B) = \frac{3}{10} \cdot \frac{3}{10}$$

$$P(A \mid H,H) = \frac{P(H,H \mid A) \cdot P(A)}{P(H,H)} = \frac{1}{P(H,H)} \cdot \underbrace{\frac{1}{4} \cdot \frac{2}{5}}_{\text{bigger}}$$

$$P(B \mid H,H) = \frac{1}{P(H,H)} \cdot \frac{9}{100} \cdot \frac{3}{5}$$

$$\Rightarrow \hat{\theta}_{MAP} = A$$

Uses of MAP estimation:

- online learning
- Suppose in the thumbtack experiment we got 30 times head.

  max. lik. $\Rightarrow$ $P(\text{head}) = 1$

  Common sense: pin is possible

  $\rightsquigarrow$ take as prior $P(\text{head}) = \frac{1}{2}$

  $P(\text{pin}) = \frac{1}{2}$

  i.e. $\theta = \frac{1}{2}$
- if you know „nothing", choose max. entropy prior

What distribution should we choose for the parameters?

(i.e. for the prior)

- incorporating knowledge
- expressing lack of knowledge
- "good" expressing posterior

[conjugate priors]

Bayesian updating involves two families of probability distributions:

1. The parametrized family in which we look for the model for our data.

2. Another family of distributions on the set of parameters (each member tells us how probable a certain parameter, or family of parameters, is)

One has to choose a prior from family 2. This family is called a conjugate family for family 1. if it is closed under Bayesian updates, i.e. if starting in family 2. and updating with a member of family 1. results in a new member of family 2.

Beta-binomial model

MLPP 3.3

The beta family is a conjugate family for the binomial distributions.

Data

- flip repeatedly a coin with unknown heads probability θ
- k number of heads, n total number of throws
- k is the data $\mathcal{D}$
- same as wearing glasses example (Section 05)

Specify

$$
\begin{aligned}
\theta &\sim \mathrm{Beta}(a, b) \qquad & p(\theta) &= \mathrm{Beta}(\theta \mid a, b) \qquad & &\text{prior} \\
k \mid \theta &\sim \mathrm{Bin}(n, \theta) \qquad & p(k \mid \theta) &= \mathrm{Bin}(k \mid n, \theta) \qquad & &\text{likelihood}
\end{aligned}
$$

Infer

$$
\begin{aligned}
\theta \mid k &\sim \mathrm{Beta}(a + k, b + n - k) \qquad & &\text{posterior} \\
p(\theta \mid k) &= \mathrm{Beta}(\theta \mid a + k, b + n - k) \qquad & &\text{posterior}
\end{aligned}
$$

- both notations are fine: $\theta \sim \mathrm{Beta}(a, b)$ and $p(\theta) = \mathrm{Beta}(\theta \mid a, b)$

Beta distribution

see MLPP 2.4.6

- random variable $\theta \in [0, 1]$ (interval between zero and one)
- parameters $a > 0$ and $b > 0$
- θ has beta distribution, written

$$\theta \sim \mathrm{Beta}(a, b)$$

- probability density function

$$\mathrm{Beta}(\theta \mid a, b) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

  with $B(a, b)$ being the beta function

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

- $\mathrm{E}\, X = \frac{a}{a+b}$, $\mathrm{Var}\, X = \frac{ab}{(a+b)^2(a+b+1)}$, mode $= \frac{a-1}{a+b-2}$ (max of the PDF)

Gamma function, Beta function, and all that

from `http://en.wikipedia.org/wiki/Gamma_function`
and `http://en.wikipedia.org/wiki/Beta_function`

Gamma function (extension of factorial function)

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1}\, dt \qquad \text{for } z \in \mathbb{C}$$

$$\Gamma(n) = (n-1)! = n!/n \qquad \text{for } n \in \mathbb{N}$$

Beta function (extension of ... ?)

$$\begin{aligned} B(x,y) &= \int_0^1 t^{x-1}(1-t)^{y-1}\, dt \\ &= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} && \text{for } x, y \in \mathbb{C} \text{ with } x+\bar{x}, y+\bar{y} > 0 \\ B(m,n) &= \frac{(m-1)!\,(n-1)!}{(m+n-1)!} && \text{for } m, n \in \mathbb{N} \\ &= \binom{m+n}{n}^{-1} \frac{m+n}{m\,n} && \text{binomial coefficient} \end{aligned}$$

Dirichlet distribution

The Dirichlet family is a conjugate family for the multinomial distributions.

see MLPP 2.5.4

- random vector $\theta = [\theta_1, \dots, \theta_K]^T$ with values in probability simplex, i.e. $\sum_j \theta_j = 1$, $\theta_j \geq 0$.
- parameter vector $\alpha = [\alpha_1, \dots, \alpha_K]^T$, with $\alpha_j > 0$
- θ has Dirichlet distribution, written

$$\theta \sim \mathrm{Dir}(\alpha)$$

- probability density function

$$\mathrm{Dir}(\theta \,|\, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

with $B(\alpha)$ generalizing the beta function

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

- special case: $\mathrm{Beta}(a, b) = \mathrm{Dir}([a, b]^T)$

Beta-binomial model

MLPP 3.3

Data

- flip repeatedly a coin with unknown heads probability θ
- k number of heads, n total number of throws
- k is the data $\mathcal{D}$
- same as wearing glasses example (Section 05)

Specify

$$\begin{aligned} p(\theta) &= \mathrm{Beta}(\theta \,|\, a, b) && \text{prior} \\ p(\mathcal{D} \,|\, \theta) &= \mathrm{Bin}(k \,|\, n, \theta) && \text{likelihood} \end{aligned}$$

Infer

$$p(\theta \,|\, \mathcal{D}) = \mathrm{Beta}(\theta \,|\, a + k, b + n - k) \qquad \text{posterior}$$

Since the prior and posterior have the same distribution, we say that Beta distribution is the conjugate prior for the binomial likelihood.

Dirichlet-multinomial model

MLPP 3.4

Data

- throw n times a dice with unknown probabilities $\theta = [\theta_1, \ldots, \theta_K]^T$
- data $\mathcal{D} = [x_1, \ldots, x_K]^T$, with x_j number of times side j

Specify

$$p(\theta) = \mathrm{Dir}(\theta \mid \alpha) \qquad \text{prior}$$
$$p(\mathcal{D} \mid \theta) = \mathrm{Mu}(x \mid n, \theta) \qquad \text{likelihood}$$

Infer

$$p(\theta \mid \mathcal{D}) = \mathrm{Dir}(\theta \mid \alpha + x) \qquad \text{posterior}$$

Since the prior and posterior have the same distribution, we say that Dirichlet distribution is the conjugate prior for the multinomial likelihood.

Digression: Gaussian-Gaussian model

Data

- sample n times from a univariate Gaussian distribution with unknown mean μ and fixed variance σ^2
- data are n samples $x_1, \ldots, x_n$

The family of Gaussians is conjugate to itself.

Specify

$$p(\mu) = \mathcal{N}(\mu \,|\, 0, \tau^2) \qquad \text{prior}$$

$$p(x_1, \ldots, x_n \,|\, \mu) = \prod_{i=1}^{n} \mathcal{N}(x_i \,|\, \mu, \sigma^2) \qquad \text{likelihood}$$

Infer

$$p(\mu \,|\, x_1, \ldots, x_n) = \mathcal{N}(\mu \,|\, \nu, \xi^2) \qquad \text{posterior}$$

with

$$\nu = \frac{\sigma^{-2} \sum_{i=1}^{n} x_i}{\tau^{-2} + n\sigma^{-2}} \qquad \xi^2 = \frac{1}{\tau^{-2} + n\sigma^{-2}}$$

Since the prior and posterior have the same distribution, we say that Gaussian distribution is the conjugate prior for the Gaussian likelihood.

For a long list of conjugate prior and their likelihood, see
`https://en.wikipedia.org/wiki/Conjugate_prior`.

Summary: distributions for tossing coins and dice

Throw a coin ($K = 2$) or a dice ($K > 2$).

Distributions for the outcome

- coin ($K = 2$): $X \sim \mathtt{Ber}(\theta)$ with θ being scalar
- dice ($K > 2$): $X \sim \mathtt{Mu}(\theta)$ with θ being vector (length K)

Distributions for the parameter (conjugate priors!)

- coin ($K = 2$): $\theta \sim \mathtt{Beta}(a, b)$ with a and b being scalar
- dice ($K > 2$): $\theta \sim \mathtt{Dir}(\alpha)$ with α being vector (length K)