Maximum-likelihood and Bayesian parameter estimation

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

Parameter estimation

Setting

- Data are sampled from a probability distribution $p(x, y)$
- The form of the probability distribution p is known but its parameters are unknown
- There is a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of examples sampled i.i.d. according to $p(x, y)$

Task

Estimate the unknown parameters of p from training data $\mathcal{D}$.

Note: i.i.d. sampling

- *independent*: each example is sampled independently from the others
- *identically* distributed: all examples are sampled from the same distribution

Parameter estimation

Multiclass classification setting

- The training set can be divided into $\mathcal{D}_1, \dots, \mathcal{D}_c$ subsets, one for each class ($\mathcal{D}_i = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ contains i.i.d examples for target class y_i)
- For any *new* example $\mathbf{x}$ (not in training set), we compute the posterior probability of the class given the example and the full training set $\mathcal{D}$:

$$P(y_i|\mathbf{x}, \mathcal{D}) = \frac{p(\mathbf{x}|y_i, \mathcal{D})p(y_i|\mathcal{D})}{p(\mathbf{x}|\mathcal{D})}$$

Note

- same as Bayesian decision theory (compute posterior probability of class given example)
- except that parameters of distributions are unknown
- a training set $\mathcal{D}$ is provided instead

Parameter estimation

Multiclass classification setting: simplifications

$$P(y_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|y_i,\mathcal{D}_i)p(y_i|\mathcal{D})}{p(\mathbf{x}|\mathcal{D})}$$

- we assume **x** is independent of $\mathcal{D}_j$ ($j \neq i$) given y_i and $\mathcal{D}_i$
- without additional knowledge, $p(y_i|\mathcal{D})$ can be computed as the fraction of examples with that class in the dataset
- the normalizing factor $p(\mathbf{x}|\mathcal{D})$ can be computed marginalizing $p(\mathbf{x}|y_i,\mathcal{D}_i)p(y_i|\mathcal{D})$ over possible classes

Note

- We must estimate class-dependent parameters $\boldsymbol{\theta}_i$ for:

$$p(\mathbf{x}|y_i,\mathcal{D}_i)$$

Maximum Likelihood vs Bayesian estimation

Maxiumum likelihood/Maximum a-posteriori estimation

- Assumes parameters $\boldsymbol{\theta}_i$ have *fixed* but *unknown* values
- Values are computed as those maximizing the probability of the observed examples $\mathcal{D}_i$ (the training set for the class)
- Obtained values are used to compute probability for new examples:

$$p(\mathbf{x}|y_i, \mathcal{D}_i) \approx p(\mathbf{x}|\boldsymbol{\theta}_i)$$

Maximum Likelihood vs Bayesian estimation

Bayesian estimation

- Assumes parameters $\boldsymbol{\theta}_i$ are *random variables* with some known *prior* distribution
- Observing examples turns prior distribution over parameters into a *posterior* distribution
- Predictions for new examples are obtained *integrating* over all possible values for the parameters:

$$p(\mathbf{x}|y_i, \mathcal{D}_i) = \int_{\boldsymbol{\theta}_i} p(\mathbf{x}, \boldsymbol{\theta}_i | y_i, \mathcal{D}_i) d\boldsymbol{\theta}_i$$

Maxiumum likelihood/Maximum a-posteriori estimation

Maximum a-posteriori estimation

$$\boldsymbol{\theta}_i^* = \text{argmax}_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_i | \mathcal{D}_i, y_i) = \text{argmax}_{\boldsymbol{\theta}_i} p(\mathcal{D}_i, y_i | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i)$$

- Assumes a prior distribution for the parameters $p(\boldsymbol{\theta}_i)$ is available

Maximum likelihood estimation (most common)

$$\boldsymbol{\theta}_i^* = \text{argmax}_{\boldsymbol{\theta}_i} p(\mathcal{D}_i, y_i | \boldsymbol{\theta}_i)$$

- maximizes the **likelihood** of the parameters with respect to the training samples
- no assumption about prior distributions for parameters

Note

- Each class y_i is treated independently: replace $y_i, \mathcal{D}_i \rightarrow \mathcal{D}$ for simplicity

Maximum-likelihood (ML) estimation

Setting (again)

- A training data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of i.i.d. examples for the target class y is available
- We assume the parameter vector $\boldsymbol{\theta}$ has a fixed but unknown value
- We estimate such value maximizing its **likelihood** with respect to the training data:

$$\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{\theta}} \prod_{j=1}^{n} p(\mathbf{x}_j|\boldsymbol{\theta})$$

- The joint probability over $\mathcal{D}$ decomposes into a product as examples are i.i.d (thus independent of each other given the distribution)

Maximum-likelihood estimation

Maximizing log-likelihood

- It is usually simpler to maximize the logarithm of the likelihood (monotonic):

$$\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \ln p(\mathcal{D}|\boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{\theta}} \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\boldsymbol{\theta})$$

- Necessary conditions for the maximum can be obtained zeroing the gradient wrt to $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\boldsymbol{\theta}) = \mathbf{0}$$

- Points zeroing the gradient can be local or global maxima depending on the form of the distribution

Maximum-likelihood estimation

Univariate Gaussian case: unknown μ and σ^2

- the log-likelihood is:

$$\mathcal{L} = \sum_{j=1}^{n} -\frac{1}{2\sigma^2}(x_j - \mu)^2 - \frac{1}{2}\ln 2\pi\sigma^2$$

- The gradient wrt μ is:

$$\frac{\partial \mathcal{L}}{\partial \mu} = 2\sum_{j=1}^{n} -\frac{1}{2\sigma^2}(x_j - \mu)(-1) = \sum_{j=1}^{n} \frac{1}{\sigma^2}(x_j - \mu)$$

Maximum-likelihood estimation

Univariate Gaussian case: unknown μ and σ^2

- Setting the gradient to zero gives mean:

$$\sum_{j=1}^{n} \frac{1}{\sigma^2}(x_j - \mu) = 0 = \sum_{j=1}^{n}(x_j - \mu)$$

$$\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} \mu$$

$$\sum_{j=1}^{n} x_j = n\mu$$

$$\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Maximum-likelihood estimation

Univariate Gaussian case: unknown μ and σ^2

- the log-likelihood is:

$$\mathcal{L} = \sum_{j=1}^{n} -\frac{1}{2\sigma^2}(x_j - \mu)^2 - \frac{1}{2}\ln 2\pi\sigma^2$$

- The gradient wrt σ^2 is:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \sum_{j=1}^{n} -(x_j - \mu)^2 \frac{\partial}{\partial \sigma^2}\frac{1}{2\sigma^2} - \frac{1}{2}\frac{1}{2\pi\sigma^2}2\pi$$

$$= \sum_{j=1}^{n} -(x_j - \mu)^2 \frac{1}{2}(-1)\frac{1}{\sigma^4} - \frac{1}{2\sigma^2}$$

Maximum-likelihood estimation

Univariate Gaussian case: unknown μ and σ^2

- Setting the gradient to zero gives variance:

$$\sum_{j=1}^{n} \frac{1}{2\sigma^2} = \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^4}$$

$$\sum_{j=1}^{n} \sigma^2 = \sum_{j=1}^{n} (x_j - \mu)^2$$

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \mu)^2$$

Maximum-likelihood estimation

Multivariate Gaussian case: unknown μ and Σ

- the log-likelihood is:

$$\sum_{j=1}^{n} -\frac{1}{2}(\mathbf{x}_j - \mu)^t \Sigma^{-1} (\mathbf{x}_j - \mu) - \frac{1}{2} \ln (2\pi)^d |\Sigma|$$

- The maximum-likelihood estimates are:

$$\mu = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j$$

- and:

$$\Sigma = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{x}_j - \mu)(\mathbf{x}_j - \mu)^t$$

Maximum-likelihood estimation

general Gaussian case:

- Maximum likelihood estimates for Gaussian parameters are simply their empirical estimates over the samples:
  - Gaussian mean is the sample mean
  - Gaussian covariance matrix is the mean of the sample covariances

Bayesian estimation

setting (again)

- Assumes parameters $\boldsymbol{\theta}_i$ are *random variables* with some known *prior* distribution
- Predictions for new examples are obtained *integrating* over all possible values for the parameters:

$$p(\mathbf{x}|y_i, \mathcal{D}_i) = \int_{\boldsymbol{\theta}_i} p(\mathbf{x}, \boldsymbol{\theta}_i | y_i, \mathcal{D}_i) d\boldsymbol{\theta}_i$$

- probability of $\mathbf{x}$ given each class y_i is independent of the other classes y_j, for simplicity we can again write:

$$p(\mathbf{x}|y_i, \mathcal{D}_i) \rightarrow p(\mathbf{x}|\mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

- where $\mathcal{D}$ is a dataset for a certain class y and $\boldsymbol{\theta}$ the parameters of the distribution

Bayesian estimation

setting

$$p(\mathbf{x}|\mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

- $p(\mathbf{x}|\boldsymbol{\theta})$ can be easily computed (we have both form and parameters of distribution, e.g. Gaussian)
- need to estimate the parameter posterior density given the training set:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

Bayesian estimation

denominator

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- $p(\mathcal{D})$ is a constant independent of $\boldsymbol{\theta}$ (i.e. it will no influence final Bayesian decision)
- if final *probability* (not only decision) is needed we can compute:

$$p(\mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Bayesian estimation

Univariate normal case: unknown μ, known σ^2

- Examples are drawn from:

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

- The Gaussian mean prior distribution is itself normal:

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

- The Gaussian mean posterior given the dataset is computed as:

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} = \alpha \prod_{j=1}^{n} p(x_j|\mu)p(\mu)$$

where $\alpha = 1/p(\mathcal{D})$ is independent of μ

Univariate normal case: unknown μ, known σ^2

a posteriori parameter density

$$p(\mu|\mathcal{D}) = \alpha \prod_{j=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma}\right)^2\right]}^{p(x_j|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{j=1}^{n}\left(\frac{\mu-x_j}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{j=1}^{n} x_j + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

Normal distribution

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

Univariate normal case: unknown μ, known σ^2

recovering mean and variance

$$\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{j=1}^{n}x_j+\frac{\mu_0}{\sigma_0^2}\right)\mu+\alpha''' = \left(\frac{\mu-\mu_n}{\sigma_n}\right)^2$$

$$= \frac{1}{\sigma_n^2}\mu^2 - 2\frac{\mu_n}{\sigma_n^2}\mu + \frac{\mu_n^2}{\sigma_n^2}$$

- Solving for μ_n and σ_n^2 we obtain:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

- where $\hat{\mu}_n$ is the sample mean:

$$\hat{\mu}_n = \frac{1}{n}\sum_{j=1}^{n}x_j$$

Univariate normal case: unknown μ, known σ^2

Interpreting the posterior

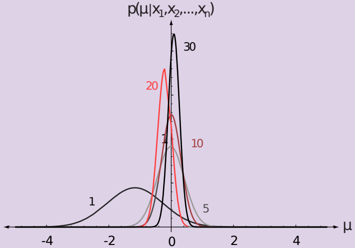
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

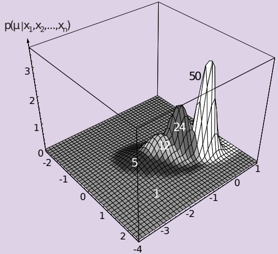
- The mean is a linear combination of the prior (μ_0) and sample means ($\hat{\mu}_n$)
- The more training examples (n) are seen, the more sample mean (unless $\sigma_0^2 = 0$) dominates over prior mean.
- The more training examples (n) are seen, the more variance decreases making the distribution sharply peaked over its mean:

$$\lim_{n\to\infty} \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

Univariate normal case: unknown μ, known σ^2

Mean posterior distribution varying sample size





Univariate normal case: unknown μ, known σ^2

Computing the class conditional density

$$\begin{aligned} p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})d\mu \\ &= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\ &\sim N(\mu_n, \sigma^2 + \sigma_n^2) \end{aligned}$$

Note (proof omitted)

- the probability of x given the dataset for the class is a Gaussian with:
  - mean equal to the posterior mean
  - variance equal to the sum of the known variance (σ^2) and an additional variance (σ_n^2) due to the uncertainty on the mean

Multivariate normal case: unknown μ, known Σ

Generalization of univariate case

- $p(\mathbf{x}|\mu) \sim N(\mu, \Sigma)$
- $p(\mu) \sim N(\mu_0, \Sigma_0)$

$$\Downarrow$$

- $p(\mu|\mathcal{D}) \sim N(\mu_n, \Sigma_n)$

$$\Downarrow$$

- $p(\mathbf{x}|\mathcal{D}) \sim N(\mu_n, \Sigma + \Sigma_n)$

Sufficient statistics

Definition

- Any function on a set of samples $\mathcal{D}$ is a *statistic*
- A statistic $\mathbf{s} = \phi(\mathcal{D})$ is *sufficient* for some parameters $\boldsymbol{\theta}$ if:

$$P(\mathcal{D}|\mathbf{s}, \boldsymbol{\theta}) = P(\mathcal{D}|\mathbf{s})$$

- If $\boldsymbol{\theta}$ is a random variable, a sufficient statistic contains all relevant information $\mathcal{D}$ has for estimating it:

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathbf{s}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathbf{s})p(\boldsymbol{\theta}|\mathbf{s})}{p(\mathcal{D}|\mathbf{s})} = p(\boldsymbol{\theta}|\mathbf{s})$$

Use

- A sufficient statistic allows to compress a sample $\mathcal{D}$ into (possibly few) values
- Sample mean and covariance are sufficient statistics for true mean and covariance of the Gaussian distribution

Conjugate priors

Definition

- Given a likelihood function $p(x|\theta)$
- Given a prior distribution $p(\theta)$
- $p(\theta)$ is a *conjugate prior* for $p(x|\theta)$ if the posterior distribution $p(\theta|x)$ is in the same family as the prior $p(\theta)$

Examples

Likelihood	Parameters	Conjugate prior
Binomial	p (probability)	Beta
Multinomial	$\boldsymbol{p}$ (probability vector)	Dirichlet
Normal	μ (mean)	Normal
Multivariate normal	$\boldsymbol{\mu}_i$ (mean vector)	Normal

Bernoulli distribution

Setting

- Boolean event: $x = 1$ for success, $x = 0$ for failure (e.g. tossing a coin)
- Parameters: θ = probability of success (e.g. head)
- Probability mass function

$$P(x|\theta) = \theta^x (1-\theta)^{1-x}$$

- Beta conjugate prior:

$$P(\theta|\psi) = P(\theta|\alpha_h, \alpha_t) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_h)\Gamma(\alpha_t)} \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

Bernoulli distribution

Maximum likelihood estimation: example

- Dataset $\mathcal{D} = \{H, H, T, T, T, H, H\}$ of N realizations (e.g. head/tail coin toss results)
- Likelihood function:

$$p(\mathcal{D}|\theta) = \theta \cdot \theta \cdot (1-\theta) \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta = \theta^h (1-\theta)^t$$

- Maximum likelihood parameter:

$$\frac{\partial}{\partial\theta} \ln p(\mathcal{D}|\theta) = 0 \quad \Rightarrow \quad \begin{aligned} \frac{\partial}{\partial\theta} h \ln\theta + t \ln(1-\theta) &= 0 \\ h\frac{1}{\theta} - t\frac{1}{1-\theta} &= 0 \\ h(1-\theta) &= t\theta \\ \theta &= \frac{h}{h+t} \end{aligned}$$

- h, t are the sufficient statistics

Bernoulli distribution

Bayesian estimation: example

- Parameter posterior is proportional to:

$$P(\theta|\mathcal{D},\psi) \propto P(\mathcal{D}|\theta)P(\theta|\psi) \propto \theta^h(1-\theta)^t\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$$

- i.e. the posterior has a beta distribution with parameters $h+\alpha_h, t+\alpha_t$:

$$P(\theta|\mathcal{D},\psi) \propto \theta^{h+\alpha_h-1}(1-\theta)^{t+\alpha_t-1}$$

- The prediction for a new event is the expected value of the posterior beta:

$$P(x|\mathcal{D}) = \int P(x|\theta)P(\theta|\mathcal{D},\psi)d\theta = \int \theta P(\theta|\mathcal{D},\psi)d\theta$$
$$= \mathrm{E}_{P(\theta|\mathcal{D},\psi)}[\theta] = \frac{h+\alpha_h}{h+t+\alpha_h+\alpha_t}$$

Bernoulli distribution

Interpreting priors

- Our prior knowledge is encoded as a number $\alpha = \alpha_h + \alpha_t$ of imaginary experiments
- we assume α_h times we observed heads
- α is called *equivalent sample size*
- $\alpha \rightarrow 0$ reduces estimation to the classical ML approach (frequentist)

Multinomial distribution

Setting

- Categorical event with r states $x \in \{x^1, \ldots, x^r\}$ (e.g. tossing a six-faced dice)
- One-hot encoding $\mathbf{z}(x) = [z_1(x), \ldots, z_r(x)]$ with $z_k(x) = 1$ if $x = x^k$, 0 otherwise.
- Parameters: $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_r]$ probability of each state
- Probability mass function

$$P(x|\boldsymbol{\theta}) = \prod_{k=1}^{r} \theta_k^{z_k(x)}$$

- Dirichlet conjugate prior:

$$P(\boldsymbol{\theta}|\psi) = P(\boldsymbol{\theta}|\alpha_1, \ldots, \alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1}$$

Multinomial distribution

Maximum likelihood estimation: example

- Dataset $\mathcal{D}$ of N realizations (e.g. results of tossing a dice)
- Likelihood function:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{j=1}^{N} \prod_{k=1}^{r} \theta_k^{z_k(x_j)} = \prod_{k=1}^{r} \theta_k^{N_k}$$

- Maximum likelihood parameter:

$$\theta_k = \frac{N_k}{N}$$

- $N_1, \ldots, N_r$ are the sufficient statistics

Multinomial distribution

Bayesian estimation: example

- Parameter posterior is proportional to:

$$P(\boldsymbol{\theta}|\mathcal{D},\psi) \propto P(\mathcal{D}|\boldsymbol{\theta})P(\theta|\psi) \propto \prod_{k=1}^{r} \theta_k^{N_k} \theta_k^{\alpha_k - 1}$$

- i.e. the posterior has a Dirichlet distribution with parameters $N_k + \alpha_k, k = 1, \ldots, r$:

$$P(\boldsymbol{\theta}|\mathcal{D},\psi) \propto \prod_{k=1}^{r} \theta_k^{N_k + \alpha_k - 1}$$

- The prediction for a new event is the expected value of the posterior Dirichlet:

$$P(x_k|\mathcal{D}) = \int \theta_k P(\boldsymbol{\theta}|\mathcal{D},\psi) d\boldsymbol{\theta} = \mathrm{E}_{P(\boldsymbol{\theta}|\mathcal{D},\psi)}[\theta_k] = \frac{N_k + \alpha_k}{N + \alpha}$$

Appendix

Additional reference material

Maximum-likelihood estimation

Multivariate Gaussian case: proof (mean)

- The gradient wrt to the mean is:

$$\nabla_{\boldsymbol{\mu}} \sum_{j=1}^{n} -\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) - \frac{1}{2} \ln{(2\pi)^d |\Sigma|} = \sum_{j=1}^{n} \Sigma^{-1} (\mathbf{x}_j - \boldsymbol{\mu})$$

Note

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T A \mathbf{x} &= A^T \mathbf{x} + A \mathbf{x} \\ &= 2A\mathbf{x} \quad \text{for symmetric } A \end{aligned}$$

Maximum-likelihood estimation

Multivariate Gaussian case: proof (mean)

- Setting the gradient to zero gives:

$$\sum_{j=1}^{n} \Sigma^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) = \mathbf{0}$$

$$\sum_{j=1}^{n}(\mathbf{x}_j - \boldsymbol{\mu}) = \Sigma\,\mathbf{0} = \mathbf{0}$$

$$\sum_{j=1}^{n} \mathbf{x}_j = \sum_{j=1}^{n} \boldsymbol{\mu} = n\,\boldsymbol{\mu}$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j$$

Maximum-likelihood estimation

Multivariate Gaussian case: proof (covariance)

- The gradient wrt to the covariance is:

$$\frac{\partial}{\partial\Sigma}\sum_{j=1}^{n}-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}_j-\boldsymbol{\mu})-\frac{1}{2}\ln(2\pi)^d|\Sigma| =$$

$$-\frac{1}{2}\left(\sum_{j=1}^{n}\frac{\partial}{\partial\Sigma}(\mathbf{x}_j-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}_j-\boldsymbol{\mu})+\sum_{j=1}^{n}\frac{\partial}{\partial\Sigma}\ln(2\pi)^d|\Sigma|\right)$$

Maximum-likelihood estimation

Multivariate Gaussian case: proof (covariance)

$$\frac{\partial}{\partial\Sigma}(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) =$$

$$(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \frac{\partial}{\partial\Sigma}\Sigma^{-1} =$$

$$-(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2}$$

Note

Use matrix derivative rule:

$$\frac{\partial}{\partial B} tr(ABC) = CA$$

Where $A = (\mathbf{x}_j - \boldsymbol{\mu})^t$, $B = \Sigma^{-1}$, $C = (\mathbf{x}_j - \boldsymbol{\mu})$ and $tr(ABC) = ABC$ as ABC is a scalar.

Maximum-likelihood estimation

Multivariate Gaussian case: proof (covariance)

$$\frac{\partial}{\partial \Sigma} \ln (2\pi)^d |\Sigma| = \frac{1}{(2\pi)^d} |\Sigma|^{-1} \frac{\partial}{\partial \Sigma} (2\pi)^d |\Sigma| =$$

$$\frac{1}{(2\pi)^d} |\Sigma|^{-1} (2\pi)^d \frac{\partial}{\partial \Sigma} |\Sigma| = |\Sigma|^{-1} |\Sigma| \Sigma^{-1} = \Sigma^{-1}$$

Note

Use matrix derivative rule:

$$\frac{\partial}{\partial A} |A| = |A| A^{-1}$$

Maximum-likelihood estimation

Multivariate Gaussian case: proof (covariance)

- Combining and putting equal to zero:

$$-\frac{1}{2}\left(\sum_{j=1}^{n}\overbrace{-(\mathbf{x}_j-\boldsymbol{\mu})(\mathbf{x}_j-\boldsymbol{\mu})^t\Sigma^{-2}}^{\frac{\partial}{\partial\Sigma}(\mathbf{x}_j-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}_j-\boldsymbol{\mu})}+\sum_{j=1}^{n}\overbrace{\Sigma^{-1}}^{\frac{\partial}{\partial\Sigma}\ln(2\pi)^d|\Sigma|}\right)=0$$

Maximum-likelihood estimation

Multivariate Gaussian case: proof (covariance)

$$\sum_{j=1}^{n} \Sigma^{-1} = \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2}$$

$$\Sigma^2 \sum_{j=1}^{n} \Sigma^{-1} = \Sigma^2 \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2}$$

$$\sum_{j=1}^{n} \Sigma = \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t$$

$$n\Sigma = \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t$$

$$\Sigma = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t$$

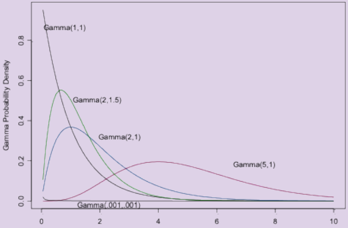
Bayesian estimation

Gamma distribution

- Defined in the interval $[0, \infty]$
- Parameters: $\alpha > 0$ (shape) $\beta > 0$ (rate)
- Probability density function:

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- $\mathrm{E}[x] = \frac{\alpha}{\beta}$
- $\mathrm{Var}[x] = \frac{\alpha}{\beta^2}$



Note

Used to model the prior distribution of the *precision* (inverse variance, i.e. $\lambda = 1/\sigma^2$).

Bayesian estimation

Univariate normal case: unknown μ and $\lambda = 1/\sigma^2$

- Examples are drawn from:

$$p(x|\mu, \lambda) \sim N(\mu, 1/\lambda)$$

- The Prior of mean and precision is the NormalGamma distribution:

$$\begin{aligned} p(\mu, \lambda) = p(\mu|\lambda)p(\lambda) &= N(\mu|\mu_0, \frac{1}{\kappa_0 \lambda})Ga(\lambda|\alpha_0, \beta_0) \\ &= NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) \end{aligned}$$

Univariate normal case: unknown μ and $\lambda = 1/\sigma^2$

a posteriori parameter density

$$p(\mu,\lambda|\mathcal{D}) = \frac{1}{\mathcal{D}} \prod_{j=1}^{n} \overbrace{\frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda}{2}(x_j-\mu)^2\right]}^{p(x_j|\mu,\lambda)} \overbrace{\frac{(\kappa_0\lambda)^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2\right]}^{p(\mu|\lambda)}$$

$$\underbrace{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\lambda^{\alpha_0-1}\exp(-\beta_0\lambda)}_{p(\lambda)}$$

$$\propto \lambda^{\alpha_0+n/2-1}\exp(-\beta_0\lambda)\lambda^{1/2}\exp -\frac{\lambda}{2}\left[\sum_{j=1}^{n}(x_j-\mu)^2 - \kappa_0(\mu-\mu_0)^2\right]$$

a posteriori parameter density is still NormalGamma

$$p(\mu,\lambda|\mathcal{D}) = NG(\mu,\lambda|\mu_n,\kappa_n,\alpha_n,\beta_n)$$

Univariate normal case: unknown μ and $\lambda = 1/\sigma^2$

a posteriori parameter density is still NormalGamma

$$p(\mu, \lambda | \mathcal{D}) = NG(\mu, \lambda | \mu_n, \kappa_n, \alpha_n, \beta_n)$$

where

$$\mu_n = \frac{\kappa_0 \mu_0 + n\hat{\mu}_n}{k_0 + n}$$

$$\kappa_n = k_0 + n$$

$$\alpha_n = \alpha_0 + n/2$$

$$\beta_n = \beta_0 + \frac{1}{2}\sum_{j=1}^{n}(x_j - \hat{\mu}_n)^2 + \frac{\kappa_0 n(\hat{\mu}_n - \mu_0)^2}{2(\kappa_0 + n)}$$

Univariate normal case: unknown μ and $\lambda = 1/\sigma^2$

Interpreting the posterior

- Posterior mean is weighted average of prior (μ_0) and sample (μ_n) means, weighted by κ_0 and n respectively

$$\mu_n = \frac{\kappa_0 \mu_0 + n\hat{\mu}_n}{k_0 + n}$$

- Posterior κ_n is increased by the number of samples n

$$\kappa_n = k_0 + n$$

- Posterior α_n is increased by half the number of samples n

$$\alpha_n = \alpha_0 + n/2$$

Univariate normal case: unknown μ and $\lambda = 1/\sigma^2$

Interpreting the posterior

- Posterior sum of squares (β_n) is sum of prior sum of squares (β_0) and sample sum of squares $\frac{1}{2}\sum_{j=1}^{n}(x_j - \hat{\mu}_n)^2$ and a term due to the discrepancy between the sample mean and the prior mean.

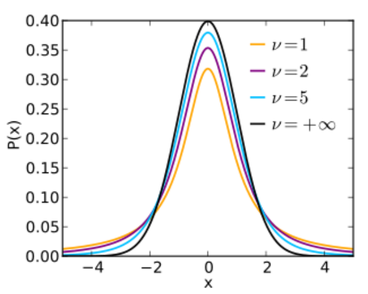
$$\beta_n = \beta_0 + \frac{1}{2}\sum_{j=1}^{n}(x_j - \hat{\mu}_n)^2 + \frac{\kappa_0 n(\hat{\mu}_n - \mu_0)^2}{2(\kappa_0 + n)}$$

Univariate normal case: unknown μ and $\lambda = 1/\sigma^2$

Computing the posterior predictive

$$p(x|\mathcal{D}) = \int_\mu \int_\lambda p(x|\mu, \lambda) p(\mu, \lambda|\mathcal{D}) d\mu d\lambda$$
$$= \frac{P(x, \mathcal{D})}{P(\mathcal{D})} = t_{2\alpha_n}\left(x|\mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n \kappa_n}\right)$$

- It is a T-distribution with mean μ_n and precision $\frac{\beta_n(\kappa_n+1)}{\alpha_n\kappa_n}$ (proof omitted)



Bayesian estimation

Wishart distribution

- Defined over $d \times d$ positive semi-definite matrix
- Parameters: $\nu > d - 1$ (degree of freedom) $T > 0$ ($d \times d$ scale matrix)
- Probability density function:

$$p(X; \nu, T) = \frac{1}{2^{\nu d/2} |T|^{\nu/2} \Gamma_d(\nu/2)} |X|^{\frac{\nu - d - 1}{2}} \exp -\frac{1}{2} tr(T^{-1} X)$$

- $\mathrm{E}[X] = \nu T$
- $\mathrm{Var}[X_{ij}] = \nu (T_i i^2 + T_{ii} T_{jj})$

Note

Used to model the prior distribution of the *precision* matrix (inverse covariance matrix, i.e. $\Lambda = \Sigma^{-1}$). T is the prior covariance

Multivariate normal case: unknown μ and Σ

- Examples are drawn from:

$$p(x|\boldsymbol{\mu}, \Lambda) \sim N(\boldsymbol{\mu}, \Lambda^{-1})$$

- The Prior of mean and precision is the NormalWishart distribution:

$$p(\boldsymbol{\mu}, \Lambda) = p(\boldsymbol{\mu}|\Lambda)p(\Lambda) = N(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\kappa_0\Lambda)^{-1})Wi(\Lambda|\nu, T)$$

Multivariate normal case: unknown μ and Σ

a posteriori parameter density

$$p(\boldsymbol{\mu}, \Lambda|\mathcal{D}) = N(\mu|\mu_n(\kappa_n\Lambda)^{-1})Wi(\Lambda|\nu_n, T_n)$$

where

$$\mu_n = \frac{\kappa_0\mu_0 + n\hat{\mu}_n}{k_0 + n}$$

$$T_n = T + \sum_{i=1}^{n}(x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^T + \frac{\kappa n}{\kappa + n}(\mu_0 - \hat{\mu}_n)(\mu_0 - \hat{\mu}_n)^T$$

$$\nu_n = \nu + n \qquad \kappa_n = \kappa + n$$

Computing the posterior predictive

$$p(x|\mathcal{D}) = t_{\nu_n - d+1}\left(x|\mu_n, \frac{T_n(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)}\right)$$