Bayesian Methods for Machine Learning

Simon Leglaive - CentraleSupélec

Bayesian inference for the Gaussian

Let $\mathbf{x} = \{x_i \in \mathbb{R}\}_{i=1}^N$ denote a set of N independent and identically distributed (i.i.d) observations following a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+^*$.

Part 1: We first consider the mean μ and the variance σ^2 as deterministic parameters.

Question 1 Why can we factorize the likelihood as in equation (1)?

$$p(\mathbf{x}; \mu, \sigma^2) = \prod_{i=1}^N p(x_i; \mu, \sigma^2), \qquad \text{where } p(x_i; \mu, \sigma^2) = \mathcal{N}(x_i; \mu, \sigma^2). \tag{1}$$

Question 2 Using the probability density function (pdf) of the Gaussian distribution defined in equation (6) of the appendix, show that the maximum-likelihood estimates of the mean and variance are given by:

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i; \tag{2}$$

$$\sigma^2_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{\text{ML}})^2. \tag{3}$$

Part 2: We now consider the mean μ as a latent random variable following a Gaussian prior distribution $p(\mu) = \mathcal{N}(\mu; \mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are considered as deterministic hyper-parameters.[1]

The likelihood model is unchanged, i.e. $p(\mathbf{x} \mid \mu; \sigma^2) = \prod_{i=1}^N \mathcal{N}(x_i; \mu, \sigma^2)$, where the conditioning bar '|' indicates that μ is now a random variable. The variance σ^2 is still considered as a deterministic parameter.

Question 3 Show that the posterior distribution of μ is given by equation (4).

$$p(\mu \mid \mathbf{x}; \sigma^2) = \mathcal{N}(\mu; \mu_\star, \sigma_\star^2), \qquad \text{where } \begin{cases} \mu_\star &= \dfrac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \dfrac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}} \\ \dfrac{1}{\sigma_\star^2} &= \dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma^2} \end{cases}, \tag{4}$$

[1] To simplify notations, we omit to denote the hyper-parameters in the prior, i.e. we simply write $p(\mu)$ instead of $p(\mu; \mu_0, \sigma_0^2)$.

where μ_{ML} is defined in (2).

Question 5 Give the limit of $\mu_\star$ and $\sigma_\star^2$ when the number of observations N goes to zero and interpret the result.

Question 6 Give the limit of $\mu_\star$ and $\sigma_\star^2$ when the number of observations N goes to infinity and interpret the result.

Part 3: We consider now that the mean μ is again a deterministic parameter while the variance is a latent random variable following an inverse-gamma prior distribution $p(\sigma^2) = \mathcal{IG}(\sigma^2; \alpha, \beta)$ where α and β are deterministic hyper-parameters.

The likelihood model is unchanged, i.e. $p(\mathbf{x} \mid \sigma^2; \mu) = \prod_{i=1}^N \mathcal{N}(x_i; \mu, \sigma^2)$, where the conditioning bar '|' indicates that σ^2 is now a random variable, while again, the mean μ is still considered as a deterministic parameter.

Question 7 Show that the posterior distribution of σ^2 is given by equation (5).

$$p(\sigma^2 \mid \mathbf{x}; \mu) = \mathcal{IG}(\sigma^2; \alpha_\star, \beta_\star), \qquad \text{where} \begin{cases} \alpha_\star &= \alpha + \dfrac{N}{2} \\ \beta_\star &= \beta + \dfrac{1}{2}\displaystyle\sum_{i=1}^N (x_i - \mu)^2 \end{cases}. \tag{5}$$

Note: Use the probability density functions defined in the appendix.

Question 8 Explain how we could estimate the deterministic model parameters μ, α and β?

Appendix

Gaussian distribution The probability density function (pdf) of the Gaussian distribution is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{6}$$

where $x \in \mathbb{R}$ is the Gaussian random variable, $\mu = \mathbb{E}[x] \in \mathbb{R}$ is the mean and $\sigma^2 = \mathbb{E}[(x-\mu)^2] \in \mathbb{R}_+^*$ is the variance.

Inverse-Gamma distribution The probability density function (pdf) of the inverse-gamma distribution is given by

$$\mathcal{IG}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right), \tag{7}$$

where $x \in \mathbb{R}_+^*$ is the inverse-gamma random variable, $\alpha \in \mathbb{R}_+^*$ and $\beta \in \mathbb{R}_+^*$ are the shape and scale parameters, respectively, and $\Gamma(\cdot)$ is the Gamma function (you do not need its definition).

Moreover, we have the following properties:

$$\mathbb{E}[x^{-1}] = \alpha/\beta, \tag{8}$$

$$\mathbb{E}[\ln(x)] = \ln(\beta) - \psi(\alpha), \tag{9}$$

$$\tag{10}$$

where $\psi(\cdot)$ is the digamma function (you do not need its definition).