Bayesian Methods for Machine Learning

Exam - January 2021

This exam contains 5 pages, 3 exercises and 1 appendix. The duration and the scoring scale for each exercise is given on an indicative basis only.

All documents are forbidden.

You have to give back this copy of the exam along with your answers.

You can answer in French or in English.

The duration of the exam is two hours.

Good luck!

Exercise 1 (45 minutes - 7.5 points)

Let $\mathcal{X}$ denote a set of observed random variables and $\mathcal{Z}$ a set of latent random variables. Defining a probabilistic model in this context consists in defining the joint distribution of the latent and observed variables, also called the complete-data likelihood:

$$p(\mathcal{X}, \mathcal{Z}; \theta) = p(\mathcal{X}|\mathcal{Z}; \theta)p(\mathcal{Z}; \theta), \tag{1}$$

where θ denotes a set of deterministic model parameters.

Question 1 (0.5 point) What "rule" of probabilities is used to write the factorization in equation (1)?

Question 2 (0.5 point) From equation (1), indicate which distribution corresponds to the prior and which one corresponds to the likelihood.

Question 3 (0.5 point) The posterior distribution of the latent variables is obtained using Bayes' theorem:

$$p(\mathcal{Z}|\mathcal{X}; \theta) = \frac{p(\mathcal{X}|\mathcal{Z}; \theta)p(\mathcal{Z}; \theta)}{p(\mathcal{X}; \theta)}, \tag{2}$$

where $p(\mathcal{X}; \theta)$ is called the marginal likelihood, or the evidence. How can the marginal likelihood be computed from the joint distribution $p(\mathcal{X}, \mathcal{Z}; \theta)$?

Question 4 (0.5 point) Briefly explain what do the prior, the likelihood and the posterior characterize in a Bayesian model.

Question 5 (1 point) For some models, the exact posterior distribution cannot be computed analytically. Explain why and describe the techniques that can be used to approximate the posterior.

Question 6 (0.5 point) Explain what does it mean to take a decision in Bayesian statistics.

Question 7 **(0.5 point)** If the model parameters θ are unknown, how can we estimate them?

Question 8 **(0.5 point)** Define the posterior predictive distribution of some possible new observations $\mathcal{X}_{\text{new}}$ given the already observed variables $\mathcal{X}$ as an expectation with respect to the posterior distribution of the latent variables $\mathcal{Z}$ (you can assume that $\mathcal{X}_{\text{new}}$ is conditionally independent of $\mathcal{X}$ given $\mathcal{Z}$).

Question 9 **(2 point)** Show that for any probability density function $q(\mathcal{Z})$, the following decomposition of the log-marginal likelihood holds:

$$\ln p(\mathcal{X};\theta) = \mathcal{L}\big(q(\mathcal{Z}),\theta\big) + D_{\text{KL}}\big(q(\mathcal{Z}) \parallel p(\mathcal{Z}|\mathcal{X};\theta)\big), \tag{3}$$

where the evidence lower-bound (ELBO) is defined by

$$\mathcal{L}\big(q(\mathcal{Z}),\theta\big) = \mathbb{E}_{q(\mathcal{Z})}[\ln p(\mathcal{X},\mathcal{Z};\theta) - \ln q(\mathcal{Z})], \tag{4}$$

and the Kullback-Leibler divergence is defined by

$$D_{\text{KL}}\big(q(\mathcal{Z}) \parallel p(\mathcal{Z}|\mathcal{X};\theta)\big) = \mathbb{E}_{q(\mathcal{Z})}[\ln q(\mathcal{Z}) - \ln p(\mathcal{Z}|\mathcal{X};\theta)]. \tag{5}$$

Question 10 **(0.5 point)** Give the expression of the distribution $q(\mathcal{Z})$ that maximizes the ELBO defined in (4).

Question 11 **(0.5 point)** Both equations (4) and (5) require computing an expectation of the form:

$$\mathbb{E}_{q(\mathcal{Z})}[f(\mathcal{Z})], \tag{6}$$

where f is an arbitrary function. Assuming this expectation cannot be computed analytically, how can you approximate it?

Exercise 2 (25 minutes - 4.5 points)

We observe N independent and identically distributed (i.i.d) random variables $\mathcal{X} = \{x_i \in \mathbb{R}\}_{i=1}^N$ following a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+^*$ (where $\mathbb{R}_+^*$ denotes $]0;+\infty[$).

Question 1 **(0.5 point)** We first consider the mean μ and the variance σ^2 as deterministic parameters. Why can we factorize the likelihood as in equation (7)?

$$p(\mathcal{X};\mu,\sigma^2) = \prod_{i=1}^N p(x_i;\mu,\sigma^2), \qquad \text{where } p(x_i;\mu,\sigma^2) = \mathcal{N}(x_i;\mu,\sigma^2). \tag{7}$$

Question 2 **(1 point)** Using the probability density function (pdf) of the Gaussian distribution defined in equation (14) of the appendix, show that the maximum-likelihood estimate of the mean is given by:

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{i=1}^N x_i. \tag{8}$$

Question 3 (0.5 point) 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. $p(\mathcal{X}|\mu; \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2)$, where the conditioning bar '|' indicates that μ is now a random variable. The variance σ^2 is still considered as a deterministic parameter.

Why do we call the prior over μ a conjugate prior for the above likelihood?

Question 4 (0.5 point) Explain how we can show that the posterior distribution of μ is given by equation (9) (do not do the math, just explain the procedure).

$$p(\mu|\mathcal{X}; \sigma^2) = \mathcal{N}(\mu; \mu_\star, \sigma_\star^2), \qquad \text{where} \begin{cases} \mu_\star &= \dfrac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \dfrac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}} \\ \dfrac{1}{\sigma_\star^2} &= \dfrac{1}{\sigma_0^2} + \dfrac{N}{\sigma^2} \end{cases}, \tag{9}$$

where μ_{ML} is defined in (8).

Question 5 (1 point) 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 (1 point) Give the limit of $\mu_\star$ and $\sigma_\star^2$ when the number of observations N goes to infinity and interpret the result.

Exercise 3 (50 minutes - 8 points)

Let $\mathcal{X} = \{x_i \in \mathbb{R}\}_{i=1}^N$ denote a set of observed random variables, $\mathcal{Z} = \{z_i \in \mathbb{R}_+\}_{i=1}^N$ a set of latent random variables and $\theta = \{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+^*\}$ a set of unknown deterministic model parameters. We consider the following generative model:

$$p(\mathcal{X}|\mathcal{Z}; \theta) = \prod_{i=1}^{N} p(x_i|z_i; \theta) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, z_i\sigma^2), \tag{10}$$

$$p(\mathcal{Z}) = \prod_{i=1}^{N} p(z_i) = \prod_{i=1}^{N} \mathcal{IG}\left(z_i; \frac{\alpha}{2}, \frac{\alpha}{2}\right), \tag{11}$$

where $\alpha \in \mathbb{R}_+^*$ is a hyper-parameter assumed to be known (which is why we omit it in the notation of $p(\mathcal{Z})$).

Question 1 (0.5 point) Draw the Bayesian network corresponding to this model.

Question 2 (0.5 point) Show that $p(\mathcal{X}; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$.

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

Question 3 (1 point) Show that for every couple $(i, j) \in \{1, ..., N\} \times \{1, ..., N\}$ such that $i \neq j$ we have

$$p(z_i, z_j|\mathcal{X}; \theta) = p(z_i|x_i; \theta)p(z_j|x_j; \theta). \tag{12}$$

Question 4 (2 points) Show that $p(z_i|x_i; \theta) = \mathcal{IG}(z_i; a_i, b_i)$ and give the expression of a_i and b_i.

Question 5 (2 point) We want to derive an expectation-maximization (EM) algorithm to estimate the unknown model parameters θ. To do so, you first have to compute

$$Q(\theta, \tilde{\theta}) = \mathbb{E}_{p(\mathcal{Z}|\mathcal{X};\tilde{\theta})}[\ln p(\mathcal{X}, \mathcal{Z}; \theta)], \tag{13}$$

where $\tilde{\theta}$ denotes the current value of the model parameters.

In your development, you can omit to denote any constant with respect to θ. You can refer to the appendix for computing expectations with respect to an inverse-gamma distribution.

Question 6 (2 point) Compute the update of the model parameters by maximizing $Q(\theta, \tilde{\theta})$ with respect to θ (simply cancel the partial derivatives).

Appendix

Usual derivatives The derivative of

- $x \mapsto ax$ is $x \mapsto a$, with $x \in \mathbb{R}$;
- $x \mapsto x^2$ is $x \mapsto 2x$, with $x \in \mathbb{R}$;
- $x \mapsto \ln(x)$ is $x \mapsto \dfrac{1}{x}$, with $x \in \mathbb{R}_+^*$;
- $x \mapsto \dfrac{1}{x}$ is $x \mapsto -\dfrac{1}{x^2}$, with $x \in \mathbb{R}^*$.

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

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{14}$$

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

$$\mathcal{IG}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right), \tag{15}$$

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:

$$\mathbb{E}[x^{-1}] = \alpha/\beta, \tag{16}$$

$$\mathbb{E}[\ln(x)] = \ln(\beta) - \psi(\alpha), \tag{17}$$

$$\tag{18}$$

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