Probabilistic Inference — Test, 2022-01-31

Duration: 50 minutes

1 Mathematical identities

- Subscripts of the covariance matrix of vector-valued random variables determine the ordering of the axes of the matrix. So for $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^E$, we have $\boldsymbol{\Sigma}_{\mathbf{xy}} \in \mathbb{R}^{D\times E}$ with

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{xy}} &= \mathrm{Cov}[\mathbf{x},\mathbf{y}] = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}[(\mathbf{x}-\mathbf{m}_{\mathbf{x}})(\mathbf{y}-\mathbf{m}_{\mathbf{y}})^\intercal] \\ &= \mathbb{E}[\mathbf{x}\mathbf{y}^\intercal] - \mathbf{m}_{\mathbf{x}}\mathbf{m}_{\mathbf{y}}^\intercal\,, \qquad (1) \\ \implies [\boldsymbol{\Sigma}_{\mathbf{xy}}]_{ij} &= \mathrm{Cov}[x_i, y_j]\,. \qquad (2) \end{aligned}$$

- Covariance matrices are symmetric by definition.

- Covariance matrices are always positive semidefinite (PSD), i.e. $\mathbf{a}^\intercal\boldsymbol{\Sigma}\mathbf{a} \geq 0, \forall \mathbf{a}$. This comes from the fact that for a random variable $\mathbf{x}$ with covariance $\boldsymbol{\Sigma}$, we can define a scalar random variable $\mathbf{a}^\intercal\mathbf{x}$ for a constant $\mathbf{a}$. Its variance must be $\mathbf{a}^\intercal\boldsymbol{\Sigma}\mathbf{a}$, and variances are always positive.

- The family of Gaussian distributions is **closed under linear transformations**. I.e. transforming the outcome of a Gaussian random vector $\mathbf{x}$ by a matrix A ($A\mathbf{x}$) will also be Gaussian distributed (see above for its variance).

  This is the **single most important** property of Gaussians that leads to many of its other properties.

- Gaussians are closed under **marginalisation** (take A to be a row vector with a element being 1), i.e. for a Gaussian $p(\mathbf{x},\mathbf{y})$ we have

$$p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{y} = \int \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{m}_{\mathbf{x}}\\ \mathbf{m}_{\mathbf{y}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{bmatrix}\right)\mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{x}; \mathbf{m}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{xx}})\,. \qquad (3)$$

- Gaussian probability density function (pdf) with input $\mathbf{x} \in \mathbb{R}^D$, which in my notes I designate by $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\big(-(\mathbf{x}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big)\,. \qquad (4)$$

- For a joint Gaussian density

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}; \begin{bmatrix}\mathbf{m}_{\mathbf{x}}\\ \mathbf{m}_{\mathbf{y}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{bmatrix}\right), \qquad (5)$$

  we have the conditional density

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\big(\mathbf{x}; \quad \mathbf{m}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y}-\mathbf{m}_{\mathbf{y}}), \quad \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\big)\,. \qquad (6)$$

2 Multiple choice questions

2.1 Finite basis function models

Consider a finite basis function model $f(x) = \boldsymbol{\phi}(x)^\intercal \mathbf{w}$ basis functions $\phi_i(x) = \exp(-(x - c_i)^2)$ with a prior $p(\mathbf{x}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{I})$, if for all i, $0 \leq c_i \leq 10$.

Question 1 If we observe data in the region $0 \leq x \leq 10$ through e.g. a Gaussian likelihood, the posterior variance of $f(\cdot)$ at $x = 20$ will be

A Very large. B 1 C Very close to zero. D 0

Question 2 If we observe data in the region $20 \leq x \leq 30$ through e.g. a Gaussian likelihood, the posterior variance of $f(\cdot)$ at $x = 50$ will be

A 0 B 1 C Very close to zero. D Very large.

Question 3 What is the prior variance on $f(x)$ for $x > 20$?

A 0 B Very close to zero. C Very large. D 1

2.2 Gaussian processes

If not otherwise stated, assume a GP model with

- a zero-mean GP prior,
- squared exponential prior covariance function: $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp((\mathbf{x} - \mathbf{x}')^\intercal (\mathbf{x} - \mathbf{x}')/(2\ell^2))$ with $\sigma_f = \ell = 1$, and
- the likelihood $p(\mathbf{y} \mid f(X), X) = \mathcal{N}\big(\mathbf{y}; f(X), \sigma^2\big)$,
- no more than 100 observations.

Question 4 The posterior for $f(X^*)$ of a model with a GP prior and a Gaussian likelihood is independent over all outputs.

A False. B True.

Question 5 The posterior of the GP model is also a Gaussian process.

A True. B False.

Question 6 The likelihood $p(\mathbf{y} \mid f(X), X) = \mathcal{N}\big(\mathbf{y}; f(X), \sigma^2\mathbf{I}\big)$ is independent over all observations.

A True. B False.

Question 7
The posterior over function values $p(f(X^*) \mid X, \mathbf{y})$ for a model with Gaussian likelihood $p(\mathbf{y} \mid f(X), X) = \mathcal{N}\big(\mathbf{y}; f(X), \sigma^2\mathbf{I}\big)$ is

$$\mathcal{N}\Big(f(X^*); \mathbf{K}_{X^*X}[\mathbf{K}_{XX} + \mathbf{A}]^{-1}\mathbf{y}, \mathbf{K}_{X^*X^*} + \mathbf{B} - \mathbf{K}_{X^*X}[\mathbf{K}_{XX} + \mathbf{A}]^{-1}\mathbf{K}_{XX^*}\Big)\,, \tag{7}$$

where $\mathbf{K}_{X_1X_2} = k(X_1, X_2)$, i.e. the prior kernel evaluated at points $X_1 \in \mathbb{R}^{N_1\times D}$ and $X_2 \in \mathbb{R}^{N_2\times D}$, giving an $N_1 \times N_2$ matrix. The correct $\mathbf{A}$ and $\mathbf{B}$ are

A $\mathbf{A} = \sigma^2\mathbf{I}$, $\mathbf{B} = \mathbf{0}$ B $\mathbf{A} = \mathbf{0}$, $\mathbf{B} = \mathbf{0}$ C $\mathbf{A} = \mathbf{0}$, $\mathbf{B} = \sigma^2\mathbf{I}$ D $\mathbf{A} = \sigma^2\mathbf{I}$, $\mathbf{B} = \sigma^2\mathbf{I}$

Question 8
The posterior over observations $\mathbf{y}^*$ at locations X^* $p(\mathbf{y}^* \mid X, \mathbf{y}, X^*)$ for a model with Gaussian likelihood $p(\mathbf{y} \mid f(X), X) = \mathcal{N}\big(\mathbf{y}; f(X), \sigma^2\mathbf{I}\big)$ is

$$\mathcal{N}\Big(\mathbf{y}^*; \mathbf{K}_{X^*X}[\mathbf{K}_{XX} + \mathbf{A}]^{-1}\mathbf{y}, \mathbf{K}_{X^*X^*} + \mathbf{B} - \mathbf{K}_{X^*X}[\mathbf{K}_{XX} + \mathbf{A}]^{-1}\mathbf{K}_{XX^*}\Big)\,, \tag{8}$$

where $\mathbf{K}_{X_1X_2} = k(X_1, X_2)$, i.e. the prior kernel evaluated at points $X_1 \in \mathbb{R}^{N_1\times D}$ and $X_2 \in \mathbb{R}^{N_2\times D}$, giving an $N_1 \times N_2$ matrix. The correct $\mathbf{A}$ and $\mathbf{B}$ are

A $\mathbf{A} = \sigma^2\mathbf{I}$, $\mathbf{B} = \sigma^2\mathbf{I}$ B $\mathbf{A} = \sigma^2\mathbf{I}$, $\mathbf{B} = \mathbf{0}$ C $\mathbf{A} = \mathbf{0}$, $\mathbf{B} = \sigma^2\mathbf{I}$ D $\mathbf{A} = \mathbf{0}$, $\mathbf{B} = \mathbf{0}$

Question 9 If we observe data in the region $20 \leq x \leq 30$ through e.g. a Gaussian likelihood, the posterior variance of $f(\cdot)$ at $x = 50$ will be

A Very close to 1. B Very close to zero. C 0 D Very large.

Question 10 A Gaussian process is completely defined by its mean function and covariance function.

A False. B True.

Question 11 A Gaussian process with a squared exponential covariance function behaves as a basis function model with `<blank>` basis functions. Substitute for `<blank>`:

A 1 B a very large but finite number of C 0 D infinite

2.3 Model selection & low-rank kernels

Question 12 In a Bayesian inference problem, the prior $p(\boldsymbol{\theta})$, likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$, and marginal likelihood $p(\mathbf{y})$ are related through Bayes' rule:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}\,. \tag{9}$$

To perform maximum a-posteriori (MAP) inference, we need to be able to evaluate

A the likelihood and prior B the posterior and likelihood C the marginal likelihood and prior D the posterior

Question 13 The marginal likelihood of a GP model with Gaussian likelihood is independent over all observations.

[A] False. [B] True.

Question 14 Integrating over the posterior uncertainty in the hyperparameters $p(\boldsymbol{\theta} \mid X, \mathbf{y})$ can be done exactly to give an analytic expression of the result.

[A] True. [B] False.

Question 15 For a linear kernel $k(\mathbf{x}, \mathbf{x}) = 1 + \mathbf{x}^\intercal \mathbf{x}$, which for an arbitrary input matrix $X \in \mathbb{R}^{N \times D}$ gives a kernel matrix of $\mathbf{K} = XX^\intercal + \mathbf{1}_{N \times N} = [X, \mathbf{1}_N][X, \mathbf{1}_N]^\intercal$, what is the best computational complexity that GP regression be performed in?

[A] $O(N^2 D)$ [B] $O(ND)$ [C] $O(N^3)$ [D] $O(ND^2)$

2.4 Bayesian optimisation

Question 16 Bayesian optimisation is most useful when the true function we are trying to optimise is very cheap to evaluate.

[A] True [B] False

Question 17 When designing an acquisition function for minimising a black box function, we need to balance exploration and exploitation. This is done by

[A] Exploring regions where the mean function is low by choosing locations where the mean function is minimised

[B] Choosing locations with a trade-off between low mean function and high uncertainty