STA414: Lecture 9 Quiz (Gaussian Processes)

Instructor: Thibault Randrianarisoa

Questions

Question 1 In Gaussian Process regression with noisy observations $y_i = f(x_i) + \varepsilon_i$, the posterior predictive variance $\mathrm{Var}[f_*|X, y, x_*]$ depends on which of the following?

a) Both the observed outputs y and the input locations X, x_*.

b) Only the observed outputs y and the noise variance σ^2_{noise}.

c) Only the input locations X, x_*, and the noise variance σ^2_{noise}.

d) Only the test input x_* and the prior variance $k(x_*, x_*)$, regardless of the training inputs.

Question 2 Why can we perform inference in a Gaussian Process model despite the fact that standard probability density functions (with respect to Lebesgue measure) do not exist in infinite-dimensional spaces?

a) We approximate the GP by truncating the infinite-dimensional prior to a fixed finite number of basis functions chosen in advance.

b) We use the marginalization property: any finite collection of function values is jointly Gaussian, so we only ever work with finite-dimensional distributions.

c) We replace Bayes' rule with a variational objective that avoids computing densities altogether.

d) We regularize the GP prior so that its density with respect to Lebesgue measure remains well-defined in the limit.

Question 3 According to Mercer's theorem, a Gaussian Process $f \sim \mathcal{GP}(0, k)$ with a continuous positive semi-definite kernel on a compact domain is equivalent to which of the following?

a) A finite Bayesian linear regression model whose number of basis functions equals the number of training points.

b) An infinite model $f(x) = \sum_{j=1}^{\infty} w_j \sqrt{\lambda_j}\, \phi_j(x)$ with independent standard normal weights.

c) A kernel density estimator centered at each training point with bandwidth determined by the length scale.

d) A mixture of finitely many Gaussian distributions, each corresponding to one eigenfunction of the kernel.

Question 4 In the log marginal likelihood of a Gaussian Process, $\log p(y|X, \mathcal{H}) = -\frac{1}{2}y^\top[K + \sigma_n^2 I]^{-1}y - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log(2\pi)$, why does optimizing the length scale ℓ not simply drive it to the smallest possible value (which would fit the data most closely and could be 0)?

a) A very small length scale makes the covariance matrix K singular, so the marginal likelihood is undefined.

b) The log-determinant term $\frac{1}{2}\log|K + \sigma_n^2 I|$ acts as a complexity penalty that increases for overly flexible models, counterbalancing the data-fit term.

c) The marginal likelihood is convex in ℓ, so gradient descent always finds the global minimum which corresponds to a moderate length scale.

d) The noise variance σ_n^2 is always fixed at some value, which prevents the length scale from shrinking.

Question 5 Consider a GP regression model with prior $f \sim \mathcal{GP}(0, k)$ and observations $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$. The posterior mean at a test point can be written in two equivalent forms:

$$\mu(x_*) = \sum_{n=1}^{N} \beta_n \, y_n = \sum_{n=1}^{N} \alpha_n \, k(x_*, x_n)$$

where $\boldsymbol{\beta} = K(x_*, X)[K(X,X) + \sigma_n^2 I]^{-1}$ and $\boldsymbol{\alpha} = [K(X,X) + \sigma_n^2 I]^{-1} y$.

Suppose we use a Matérn-$\frac{1}{2}$ kernel $k(x, x') = \exp(-|x - x'|/\ell)$ with a very small length scale $\ell \to 0^+$, while the noise variance $\sigma_n^2 > 0$ is held fixed. Which of the following best describes the behavior of the posterior predictive distribution $f_* \mid X, y, x_*$ for $x_* \notin X$?

a) The posterior mean interpolates the data exactly and the posterior variance collapses to zero everywhere.

b) The posterior mean tends toward zero and the posterior variance tends toward $k(x_*, x_*) = 1$ everywhere, recovering the prior.

c) The posterior mean converges to the nearest neighbor's observation value, and the posterior variance converges to $1 - \exp(-2|x_* - x_{\text{nearest}}|/\ell)$.

d) The posterior mean converges to the sample average $\bar{y}$ and the posterior variance converges to σ_n^2/N.

Question 6 Recall from the lecture that the random linear model $f(x) = ax + b$ with independent priors $a \sim \mathcal{N}(0, \sigma_w^2)$ and $b \sim \mathcal{N}(0, \sigma_b^2)$ corresponds to a GP with covariance function $k(x, x') = \sigma_w^2 x x' + \sigma_b^2$. Suppose we use this kernel in GP regression with noise variance $\sigma_n^2 > 0$ and observe $N \geq 2$ training points at distinct, non-zero locations. Now consider the posterior predictive variance $V(x_*)$ as the test point moves far from all training data, $|x_*| \to \infty$. Assuming $V(x_*)$ is not constant in x_*, which of the following is correct?

a) $V(x_*)$ converges to a constant σ_b^2, because the intercept is the only component whose uncertainty survives far from the data.

b) $V(x_*)$ converges to zero, because the two parameters a and b are fully determined by $N \geq 2$ observations.

c) $V(x_*)$ grows without bound as $|x_*|^2$, because the residual uncertainty on the slope a is amplified by the squared distance from the origin.

d) $V(x_*)$ converges to the prior variance $k(x_*, x_*) = \sigma_w^2 x_*^2 + \sigma_b^2$ exactly, because the training data becomes irrelevant far away.

Solutions

1. Correct Answer: c) Only the input locations X, x_*, the covariance function k, and the noise variance σ^2_{noise}, but not the observed outputs y.
Reasoning: The posterior variance is $\Sigma_{*|y} = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma^2_{\text{noise}} I]^{-1} K(X, X_*)$, which depends only on the input locations and the kernel, not on the observed values y. This is a distinctive property of GP regression highlighted in the lecture.

2. Correct Answer: b) We use the marginalization property: any finite collection of function values is jointly Gaussian, so we only ever work with finite-dimensional distributions.
Reasoning: The lecture explicitly addresses the "dimensionality hurdle" that densities with respect to Lebesgue measure vanish as $d \to \infty$. The key insight is that we never need the full infinite-dimensional density: we restrict attention to the finite set of training and test points, for which the joint distribution is a well-defined multivariate Gaussian, and apply standard Gaussian conditioning formulas.

3. Correct Answer: b) An infinite linear-in-the-parameters model $f(x) = \sum_{j=1}^{\infty} w_j \sqrt{\lambda_j}\, \phi_j(x)$ with independent standard normal weights.
Reasoning: Mercer's theorem decomposes the kernel as $k(x, x') = \sum_{m=1}^{\infty} \lambda_m \phi_m(x) \phi_m(x')$, which shows that a GP is exactly a Bayesian linear regression with infinitely many basis functions $\psi_m(x) = \sqrt{\lambda_m}\, \phi_m(x)$ and i.i.d. $\mathcal{N}(0, 1)$ weights. Option (a) is incorrect because the representation generally requires infinitely many basis functions, not finitely many.

4. Correct Answer: b) The log-determinant term $\frac{1}{2} \log |K + \sigma_n^2 I|$ acts as a complexity penalty that increases for overly flexible models, counterbalancing the data-fit term.
Reasoning: The marginal likelihood balances a data-fit term ($-\frac{1}{2} y^\top [K + \sigma_n^2 I]^{-1} y$) against a complexity penalty ($-\frac{1}{2} \log |K + \sigma_n^2 I|$). A very small length scale produces a near-diagonal, highly flexible covariance matrix whose log-determinant is large, penalizing the model. The lecture illustrates this with the length-scale fitting example, showing that the marginal likelihood does not favour overfitting.

5. Correct Answer: b) The posterior mean tends toward zero and the posterior variance tends toward $k(x_*, x_*) = 1$ everywhere, recovering the prior.
Reasoning: As $\ell \to 0^+$, the kernel $k(x_i, x_j) \to 0$ for all $i \neq j$, while $k(x_i, x_i) = 1$. Thus $K(X, X) \to I$ and, crucially, the cross-covariance vector $\mathbf{k}_* = (k(x_*, x_n))_n \to \mathbf{0}$ for any test point $x_* \notin \{x_1, \dots, x_N\}$. The posterior mean $\mu(x_*) = \mathbf{k}_*^\top [K + \sigma_n^2 I]^{-1} y \to 0$ and the posterior variance tends to $k(x_*, x_*) - \mathbf{0}^\top (\cdots)\, \mathbf{0} = 1$, which is exactly the prior. Intuitively, a negligible length scale makes every training point uncorrelated with any other location, so the data carries no information and the posterior reverts to the prior. Option (a) confuses $\ell \to 0$ with $\sigma_n^2 \to 0$; option (c) incorrectly assumes the kernel retains local influence; option (d) describes a different regime entirely.

6. Correct Answer: c) $V(x_*)$ grows without bound as $|x_*|^2$, because the residual uncertainty on the slope a is amplified by the squared distance from the origin.
Reasoning: The posterior variance is $V(x_*) = k(x_*, x_*) - \mathbf{k}_*^\top [K + \sigma_n^2 I]^{-1} \mathbf{k}_*$, where $\mathbf{k}_* = \sigma_w^2 x_* \mathbf{x} + \sigma_b^2 \mathbf{1}$. Expanding, $V(x_*)$ is a quadratic in x_* (N.B: this is already enough to answer, a nonzero quadratic function always diverges) with leading coefficient $\sigma_w^2 (1 - \sigma_w^2\, \mathbf{x}^\top [K + \sigma_n^2 I]^{-1} \mathbf{x})$. This coefficient is strictly positive because $\sigma_n^2 > 0$ means the data never fully determines the slope a, so $V(x_*) \sim C\, x_*^2$ as $|x_*| \to \infty$. Option (a) ignores the slope uncertainty. Option (b) is wrong because $\sigma_n^2 > 0$ prevents the parameters from being exactly determined. Option (d) is wrong because the data does reduce the variance (the leading coefficient drops from σ_w^2 to $\sigma_w^2 (1 - \sigma_w^2\, \mathbf{x}^\top M^{-1} \mathbf{x}) < \sigma_w^2$), just not enough to prevent quadratic growth.