Probabilistic Machine Learning

# Exercise Sheet #8

due on Monday, 26 June 2023, 10am sharp

1. **Theory Question:** Maximum a-posteriori estimation for GP classification (logistic regression). The goal of this exercise is to elucidate some connections between GP classification and support vector machines (SVMs). Consider again the GP classification model discussed in the lecture, with input-output pairs $\{(x_i, y_i)\}_{i=1}^N$, where $y_i \in \{-1, 1\}$, assumed to be generated by the logistic likelihood

$$p(\boldsymbol{y} \mid f_X) = \prod_{i=1}^{N} \sigma(y_i f(x_i)) \qquad \text{with} \qquad \sigma(a) = \frac{1}{1+\exp(-a)}.$$

As prior over the latent function $f$, we consider a GP, $p(f) = \mathcal{GP}(f; m, k)$.

(a) Write down the **log posterior** $\log p(f_X \mid \boldsymbol{y}, X)$ for the "representers"

$$f_X = [f(x_1), \ldots, f(x_N)]$$

arising from this model. You can drop terms that are constant w.r.t. $f_X$.

(b) Compute the **gradient** of this log posterior, $\nabla_{f_X} \log p(f_X \mid \boldsymbol{y}, X)$.

(c) Assume that we have found the mode $\hat{f}_X$ of this posterior (i.e. the MAP estimate), and used it to construct an approximate Gaussian posterior $q(f_X) = \mathcal{N}(f_X; \hat{f}_X, \Sigma)$ on $f_X$, where $\Sigma$ is some covariance matrix, which does not feature in the following. Show that **the MAP estimate** for the latent *function* $f$ can then be written compactly as

$$\mathbb{E}_q[f(\bullet)] = m_\bullet + k_{\bullet X} \nabla \log p(\boldsymbol{y} \mid \hat{f}_X)$$

That is, using the gradient of the log *likelihood*.

(d) By writing down the explicit gradient of $\log p(\boldsymbol{y} \mid \hat{f}_X)$, make an argument that those training points $x_i$ at which $|\hat{f}(x_i)| \gg 1$ – those "far from the decision boundary" – do **almost not contribute** to this estimate $\mathbb{E}_q[f(\bullet)]$.

(e) This observation is reminiscent of the notion of *support vectors* in SVMs, where only a few training points contribute to the solution. In SVMs, this is achieved by using the hinge loss

$$\ell(y_i; f(x_i)) = \max(0, 1 - y_i f(x_i))$$

instead of the logistic loss. Because this loss has zero gradient whenever $y_i f(x_i) > 1$, associated training pairs do not contribute at all to the optimization problem, which can be leveraged for efficient numerical optimization. It would thus be tempting to use

$$r(y_i; f(x_i)) = \exp(-\ell(y_i; f(x_i)))$$

as a likelihood in our GP classification model. Unfortunately, this is not a valid likelihood. To make this clear, show that there is no constant $c$ such that

$$\sum_{y_i \in \{-1,1\}} c r(y_i; f(x_i)) = 1 \qquad \text{for all } f(x_i) \in \mathbb{R}$$

(thus $r(y_i; f(x_i))$ is not a family of probability distributions, and can't be a likelihood).

2. **Practical Question:** can be found in `Ex08.ipynb`