1(a) Write down the log posterior $\log p(f_x | y, X)$ for the representers.

$f_x = [f(x_1), f(x_2), \ldots, f(x_N)]$

① prior over latent f : $p(f) = GP(f; m, k)$

② $y_i \in \{-1, 1\}$ input output pairs : $= \{(x_i, y_i)\}_{i=1}^N$

③ $p(y | f_x) = \prod_{i=1}^N \frac{1}{1 + \exp(-y_i f(x_i))}$ $\sigma(y_i f(x_i)) = \frac{1}{1 + \exp(-y_i f(x_i))}$

$$\log p(f_x | y) = \log p(y | f_x) + \log p(f_x) - \log (p(y)$$

$$= \sum_{i=1}^n \log \sigma(y_i f(x_i)) - \frac{1}{2} (f_x - m_x)^T K_{xx}^{-1} (f_x - m_x) + \text{constant terms.}$$

$$\approx - \sum_{i=1}^n \log (1 + \exp(-y_i f(x_i)) - \frac{1}{2} (f_x - m_x)^T K_{xx}^{-1} (f_x - m_x)$$

(b) Compute the gradient of this log posterior

$$\nabla_{f_x} \log\left(p(f_x \mid y, X)\right)$$

Answer: ① prior over latent f: $p(f) = GP(f; m, k)$

② $y_i \in \{-1, 1\}$ input output pairs $:= \{(x_i, y_i)\}_{i=1}^{N}$

③ $p(y \mid f_x) = \prod_{i=1}^{N} \frac{1}{1 + \exp(-y_i f(x_i))}$

$$\log p(f_x \mid y) = \log p(y \mid f_x) + \log p(f_x) - \log(p(y))$$

$$= \sum_{i=1}^{n} \log \sigma(y_i f(x_i)) - \frac{1}{2} (f_x - m_x)^T K_{xx}^{-1} (f_x - m_x) + \text{constant terms.}$$

$$\nabla \log\left(p(f_x \mid y)\right) = \sum_{i=1}^{n} \nabla_{f_x} \log \sigma(y_i f_{x_i}) - K_{xx}^{-1} (f_x - m_x)$$

$$= \sum_{i=1}^{n} -\nabla_{f_x} \log\left(1 + \exp(-y_i f_{x_i})\right) - K_{xx}^{-1} (f_x - m_x)$$

$$-\frac{\partial}{\partial f_{x_j}} \log\left(1 + \exp(-y_i f_{x_i})\right) = \delta_{ij} \left(\frac{y_i + 1}{2} - \sigma(f_{x_i})\right)$$

here, as $y_i \in \{-1, 1\}$, $\sigma(f_{x_i}) = \frac{1}{1 + \exp(-f_{x_i})}$

$\frac{y_i + 1}{2} \in \{0, 1\}$

Consider $-\frac{\partial}{\partial f_{x_i}} \log\left(1+\exp\left(-y_i f_{x_i}\right)\right) = \frac{-1}{1+\exp\left(-y_i f_{x_i}\right)} \cdot -y_i \exp\left(-y_i f_{x_i}\right)$

$$= \frac{y_i \exp\left(-y_i f_{x_i}\right)}{1+\exp\left(-y_i f_{x_i}\right)}$$

① $y_i = -1 \Rightarrow \frac{-\exp\left(f_{x_i}\right)}{1+\exp\left(f_{x_i}\right)} = \frac{-1}{1+\exp\left(-f_{x_i}\right)} = -\sigma\left(f_{x_i}\right)$

② $y_i = 1 \Rightarrow \frac{\exp\left(-f_{x_i}\right)}{1+\exp\left(-f_{x_i}\right)} = 1 - \frac{1}{1+\exp\left(f_{x_i}\right)} = 1 - \sigma\left(f_{x_i}\right)$

$$\Rightarrow \frac{y_i+1}{2} - \sigma\left(f_{x_i}\right)$$

$$\nabla \log\left(p\left(f_x \mid y\right)\right) = \begin{bmatrix} \frac{y_1+1}{2} - \sigma(f(x_1)) \\ \frac{y_2+1}{2} - \sigma\left(f(x_2)\right) \\ \vdots \\ \frac{y_n+1}{2} - \sigma\left(f(x_n)\right) \end{bmatrix} - K_{xx}^{-1}\left(f_x - m_x\right)$$

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

$$\mathbb{E}_q[f(\cdot)] = m_\cdot + K_{\cdot x} \nabla \log p(y|\hat{f}_x)$$

that is using the gradient of the log-likelihood.

Answer:

① $\hat{f}_x$ is the mode of posterior.

② $q(f_x) = \mathcal{N}(f_x; \hat{f}_x, \Sigma)$ is the approximate Gaussian posterior

③ Map estimate of latent function f can be written as

$$\mathbb{E}_q(f(\cdot)) = m(\cdot) + K_{\cdot x} \nabla \log p(y|\hat{f}_x)$$

④ We know that at $f_x = \hat{f}_x$ the mode.

$$\nabla_{f_x} \log p(f_x|y, X) = 0$$

$$\nabla_{f_x} \log p(y|f_x) + \nabla_{f_x} \log p(f_x) = 0$$

$$\nabla_{f_x} \log p(y|f_x) + \nabla_{f_x} \left[\frac{1}{2}(f_x - m_x)^T K_{xx}^{-1} (f_x - m_x)\right] = 0$$

$$\nabla_{f_x} \log p(y|f_x) - K_{xx}^{-1}(f_x - m_x) = 0$$

$$\nabla_{f_x} \log(p(y|f_x)) = K_{xx}^{-1}(f_x - m_x)$$

thus $f_x = \hat{f}_x$ we have:

$$\nabla_{f_x} \log(p(y|\hat{f})) = K_{xx}^{-1}(\hat{f} - m_x)$$

⑤ As we know $\hat{f}$ the approximate Gaussian posterior at training points

$$q(f_x) = \mathcal{N}(f_x; \hat{f}, \Sigma)$$

⑥ The posterior predictions at f_* (Lecture-14, Slide-17)

$$q(f_* | y) = \int p(f_* | \vec{f}_x)\, q(\vec{f}_x)\, d\vec{f}_x$$

$$= \mathcal{N}(f_*; m_* + K_{*x} K_{xx}^{-1} (\hat{f} - m_x), \Sigma_{posterior})$$

Thus $E_q(f(\cdot)) = m(\cdot) + K_{\cdot x} K_{xx}^{-1} (\hat{f} - m_x)$

as expected value is precisely mean of gaussian.

From ④ we have at $f_x = \hat{f}$

$$K_{xx}^{-1} (\hat{f} - m_x) = \nabla_{f_x} \log(p(y | \hat{f}))$$

Thus $E_q(f(\cdot)) = m(\cdot) + K_{\cdot x} \nabla_{f_x} \log(p(y | \hat{f}))$

1(d) By writing down the explicat gradient of $\log p(y|\hat{f}_x)$, 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 $\mathbb{E}_q[f(\cdot)]$.

Ans: $$\nabla_{f_x} \log\left(p(y|\hat{f}_x)\right) = \sum_{i=1}^{n} \nabla_{f_x} \log\left(\sigma(y_i f(x_i))\right)$$

$$= \sum_{i=1}^{n} \nabla_{f_x} \left[-\log\left(1+\exp(-y_i f(x_i))\right)\right]$$

$$\frac{\partial}{\partial f_i} -\log\left(1+\exp(-y_i f(x_i))\right) = \frac{+y_i \exp(-y_i f(x_i))}{1+\exp(-y_i f(x_i))}$$

$$y_i = -1 \Rightarrow \frac{(-1)\exp(f(x_i))}{1+\exp(f(x_i))} = \frac{-1}{1+\exp(-f(x_i))} = -\sigma(f(x_i))$$

$$y_i = 1 \Rightarrow \frac{(1)\exp(-f(x_i))}{1+\exp(-f(x_i))} = 1 - \frac{1}{1+\exp(-f(x_i))} = 1-\sigma(f(x_i))$$

using mapping $y_i = -1 \;;\; \frac{y_i+1}{2} = 0$

$y_i = 1 \;;\; \frac{y_i+1}{2} = 1$

$$\frac{\partial}{\partial f_i} -\log\left(1+\exp(-y_i f(x_i))\right) = \frac{y_i+1}{2} - \sigma(f(x_i))$$

$$\nabla_i \log\left(p(y|\hat{f}_x)\right) = \frac{y_i+1}{2} - \sigma(f(x_i))$$

Now consider $|\hat{f}_x| >> 1$

$$\sigma(|\hat{f}_x|) = \frac{1}{1+\exp(-|\hat{f}_x|)} = 1 \quad \text{as} \quad \lim_{a \to \infty} \exp(-a) = 0$$

Thus when $y_i = 1$ and $|\hat{f}(x_i)| >> 1$ and $\hat{f}(x_i) > 0$

$$\nabla_i \log(p(y|\hat{f}_x)) = \frac{1+1}{2} - 1 = 0$$

Thus x_i will not contribute to $E_q[f(\cdot)]$

Similarly when $y_i = -1$ and $|\hat{f}(x_i)| >> 1$ ad $f(x_i) < 0$

$$\nabla_i \log(p(y|\hat{f}_x)) = \frac{-1+1}{2} - \frac{1}{1+\exp(-f(x_i))}$$

$$= 0 - \frac{1}{1+\exp(-f(x_i))} = 0$$

$\lim$ as $-f(x_i) \to \infty$

When the signs of y_i & $f(x_i)$ agree that is $y_i f(x_i) > 0$ and $|\hat{f}(x_i)| >> 1$ $\sigma(y_i f(x_i)) \to 1$ these are highly likely training points as given by logistic likelihood at the same time the gradient of their log-logistic likelihood $\nabla_i \log(p(y|\hat{f}_x) \cong 0$

thus those training point will not contribute to update equation. as the gradient of log likelihood

$$E_q[f(\cdot)] = m_\cdot + K_{\cdot x} \nabla \log p(y|\hat{f}_a)$$ those component is zero.

1 (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 using the hinge loss

$$\ell(y_i; f(x_i)) = \max(0, 1 - y_i f(x_i)) \quad \text{instead of the logistic loss.}$$

Because this loss has zero gradient whenever $y_i f(x_i) > 1$

the 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:

$$r(y_i; f(x_i)) = \exp(-\ell(y_i; f(x_i))) \quad \text{as a likelihood.}$$

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

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.

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

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

Answer: We consider the normalization of suggested likelihood.

$$\sum_{y_i \in \{-1,1\}} c\, r(y_i; f(x_i))$$

where $r(y_i; f(x_i)) = \exp\left[-\max(0, 1 - y_i f(x_i))\right]$

we separate out +ve and -ve classifications.

s.t.

$$\sum_{\text{where } y_i = -1} c \exp\left[-\max(0, 1 + f(x_i))\right] + \sum_{y_i = 1} c \exp\left[-\max(0, 1 - f(x_i))\right]$$

Without loss of generality consider train points in +ve and negative classes with same $f(x_i) = f'$ value f'. Then either $f' \geq 1$ or $f' \leq 1$

Let $f' > 1$

without loss of generality consider training points in the +ve and -ve classes with $f(x_i) = f'$

then for two such points we have.

$$S = \underbrace{c \exp\left[-\max(0, 1 + f')\right]}_{y_i = -1} + \underbrace{c \exp\left[-\max(0, 1 - f')\right]}_{y_i = 1}$$

without loss of generality let $f' > 1$

then $\max(0, 1 + f') = 1 + f'$

$\max(0, 1 - f') = 0$

$$S = c\left[\exp\left[-(1+f')\right] + P^{\circ}\right]$$

$$= c\left[1 + e^{-(1+f')}\right]$$

The only normalizer for S is when

$$c = \frac{1}{1 + e^{-(1+f')}}$$ but this is not independent of $f(x_i)$

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