Probabilistic Machine Learning

Exercise Sheet #8

Exponential Families, and Bayesian Neural Networks

1. **Exam-Type Question — Exponential Families** Reminder: Consider a random variable X taking values $x \in \mathbb{X} \subset \mathbb{R}^n$. A probability distribution for X with pdf of the form

$$p(x \mid w) = h(x) \cdot \exp\left(\phi(x)^\intercal w - \log Z(w)\right), \qquad \text{where} \qquad Z(w) := \int_{\mathbb{X}} \exp\left(\phi(x)^\intercal w\right)\, dh(x) \tag{1}$$

is called an **exponential family** of probability distributions.

   (a) What are the commonly used names for the quantities $\phi(x)$, w, $h(x)$ and $\log Z(w)$, respectively?

   (b) Give an expresssion for the expected value $\mathbb{E}_{p(x|w)}[\phi(x)]$ that can be evaluated without computing an integral.

2. **Theory Question — Conjugate Prior Inference** Assume we are given n independent draws $\boldsymbol{x}_i \in \mathbb{R}^d$, $i = 1, \dots, n$ from a *multivariate* Gaussian distribution.

$$p(\boldsymbol{x}_i \mid \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\intercal \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})\right) \tag{2}$$

   Assume that we know the mean $\boldsymbol{\mu} \in \mathbb{R}^d$, but we do not know the symmetric positive definite covariance matrix Σ. We know from the lecture that Gaussian distributions are an exponential family, and that all exponential families have conjugate priors.

   Show that the *Wishart distribution*, the exponential family with pdf

$$\begin{aligned} \mathcal{W}_d(\Sigma^{-1}; V, \nu) :&= p(\Sigma^{-1} \mid V, \nu) \quad \text{with} \nu > d - 1 \in \mathbb{R}, \text{ and s.p.d. } V \in \mathbb{R}^{d \times d} \\ &= \frac{1}{2^{\nu d/2}|V|^{\nu/2}\Gamma_d(\nu/2)} |\Sigma^{-1}|^{(\nu-d-1)/2} \exp\left(-\operatorname{tr}(\Sigma^{-1}V^{-1})/2\right) \end{aligned}$$

   is the conjugate prior for the *inverse* covariance (aka. precision) matrix Σ^{-1}. Here Γ_d is the multivariate Gamma function,[1] and $\operatorname{tr}(Z) = \sum_i [Z]_{ii}$ is the trace operation. What is the posterior distribution $p(\Sigma^{-1} \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ over Σ^{-1}?

3. **Practical Question**

   Bayesian Deep Learning: Now that you have implemented GP classification, consider the more general case of training a Deep Neural Network. The loss function of a DNN is not a quadratic, but now that we have already approximated the likelihood with a quadratic, we might as well do so for the prior, too. Train a DNN (architecture provided in the notebook) with SGD (or Adam), then construct a Laplace approximation to the weights, using the Hessian of the entire loss function (implementing this Hessian manually would

[1] The multivariate Gamma function is defined as

$$\Gamma_d(a) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma(a + (1-j)/2),$$

but this is not needed for a solution of this exercise.

be tedious. Thankfully, there are software packages available that provide this object for you). Sample a set of weights from this approximate posterior and, for a set of test points, plot the corresponding predictive probabilities for the two classes.

On ILIAS you can find a `jupyter` notebook that describes the exercise in more detail.