**Department of Computer Science**

**Machine Learning, 2**
**Prof. Dr. Marius Kloft and Dr. Antoine Ledent**

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

# Machine Learning 2
# Problem Sheet 3

## Problem 1

An **Exponential Family** of distributions on $\mathbb{R}^d$ is a family of distributions on $\mathbb{R}^d$ parameterized by a parameter $\theta \in \mathbb{R}^n$ whose probability density functions can be written in the form

$$\forall x \in \mathbb{R}^d, \quad \forall \theta \in \mathbb{R}^n)$$
$$\text{pdf}(x) = h(x)g(\theta)\exp\left(\langle \eta(\theta), T(x)\rangle\right),$$

where $\langle ., . \rangle$ indicates inner productm, for some fixed functions

$$T : \mathbb{R}^d \to \mathbb{R}^m$$
$$\eta : \mathbb{R}^n \to \mathbb{R}^m$$
$$g : \mathbb{R}^n \to \mathbb{R}, \quad \text{and}$$
$$h : \mathbb{R}^d \to \mathbb{R}.$$

A priori, $d, n, m$ are any fixed natural numbers.

*The reason for this definition is mathematical convenience: we will see later in the course that exponential families have some very desirable mathematical properties that simplify Bayesian calculations, and that many families of distributions that are used frequently are actually exponential families.*

(a) Let $X$ be a $U[0, a]$ random variable. Let $x_1, \ldots, x_N$ be $N$ independent random samples of $X$, what is the maximum likelihood estimator of $a$?

(b) Prove that the Poisson distribution is in the exponential family.

(c) Prove that the (multivariate) Normal distribution, with unknown variance and unknown mean, is in the exponential family.

## Problem 2

The Beta function is the defined by

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta,$$

that is the only thing you may assume about it. Proving the other properties required for the exercise is part of the problem. This problem guides you through a basic understanding of the Beta function and its use.

**First, a bit of theory...** Before reading further, you should convince yourself of the following obvious, but very important fact, which has been indirectly used previously in applications of Bayes's theorem etc.: If a parameter $\gamma$ is sampled from a certain distribution (probability measure) $\mu$, and then used to simulate, independently, a random variable with distribution

(probability measure) $\nu_\gamma$ (we assume that all the measures $\nu_\gamma$ for all the possible values of $\gamma$ live on the same $\sigma$-algebra $\mathcal{F}$), then for any event $E \in \mathcal{F}$, we have

$$\mathbb{P}(E) = \int_\gamma \nu_\gamma(E) d\mu(\gamma), \tag{3}$$

where $\mathbb{P}$ is the probability measure (on the product of $\mathcal{G}$ and $\mathcal{F}$, and where $\mathcal{G}$ is the $\sigma$-algebra associated to $\mu$) corresponding to independently sampling $\gamma$ from $\mu$ and then a random variable with probability measure $\nu_\gamma$.

**Remarks:** From a strict mathematical point of view, this corresponds to a *definition* rather than a proposition, but you should convince yourself that it conforms with basic probabilitic intuition. In Machine Learning, $\mu$ is the prior distribution on $\gamma$.

**Conjugate priors** are a very important tool in Bayesian statistics:

**Definition:** We say that a family of distributions $f_\mu$ (here $\mu$ is a parameter) is a conjugate (prior) distribution for $g_\theta$ (here $\theta$ is a parameter) if, when we take a $f_\mu$ prior on $\theta$, the posterior probability distribution of $\mu$ after observing some values of the data $x$ (which is assumed to be generated following $g_\theta$ ) follows the distribution $f_{\mu'}$ for some $\mu'$ depending on $\mu$ and $x$.

Conjugate distributions are important because choosing a conjugate distribution as a prior often allows for easy calculation of the posterior via a simple formula for $\mu'$, which gives an *analytical* solution that doesn't require simulation or computation of intractable integrals. If it is a reasonable assumption, it is always a good idea to choose a conjugate prior.

(a) Suppose that the families of prior distributions with pdf $f_{1,\alpha}$ (resp. $f_{2,\alpha}$), where $\alpha$ is a (possibly multidimensional) parameter, are conjugate with respect to a family of distributions with pdfs $f_\theta$. Show that the family of convex combinations of elements of the families $f_1$ and $f_2$ is conjugate with respect to $f$. For simplicity of notation, you may restrict yourself to showing that the posterior corresponding to the prior $pf_{1,\alpha} + (1-p)f_{2,\alpha}$ (for $0 < p < 1$) is in the family. (Please give a concise answer with an equation.)

(b) Show that the beta distribution is a conjugate distribution for the Bernouilli distribution.

(c) Show that the beta distribution is in the exponential family.

(d) We observe $a$ successes and $b$ failures of a Bernoulli random variable of unknown $\theta$. What is the Maximum likelihood estimate of the probability that the next trial is a success. Suppose we have a $\mathrm{Beta}(\alpha, \beta)$ prior over the parameter $\theta$, what is the probability that the next trial is a success? Comment on the case of a uniform prior.

**Problem 3**

In the lecture we saw different estimators for the regression problem. In this exercise, we implement the different types of estimators used in the lecture.

(a) Generate a 200 5-dimensional samples $X \in \mathbb{R}^{200\times 5}$ from a normal distribution of mean 0 and identity covariance matrix.

(b) Generate the corresponding one dimensional output $Y \in \mathbb{R}^{200}$ in the following way: 1) Multiply $X$ by the vector $w = (0.5, 2, 1, 20, 0.2)^T$ from the right. And 2) Add independent normal noise to each elements with mean 0 and variance 1.

(c) Compute the MLE of $w$ based on the data generated.

(d) Use Stan to compute the MAP estimate using Normal priors and Normal likelihood.

(e) Compare the Mean Squared Error(MSE) using the two methods