CS-E4820 Machine Learning: Advanced Probabilistic Methods
Pekka Marttinen, Paul Blomstedt, Homayun Afrabandpey, Reza Ashrafi, Betül Güvenç, Tianyu Cui, Pedram Daee, Marko Järvenpää, Santosh Hiremath (Spring 2019)
Exercise problems, round 7, due on Tuesday, 19th March 2018, at 23:55
Please return your solutions in MyCourses as a single PDF file.

Problem 1. *"Bayes factors."*

(a) Suppose we have two bags, each containing a large number of black and white marbles. To learn about the contents of the bags, we have done 5 draws from each bag. After each draw, the marble drawn has been returned to the bag. The draws from the first bag are as follows (B, W, W, B, B) and the draws from the second bag are (B, B, B, B, W), where B corresponds to a Black marble and W to a White marble. Consider two models

$$\begin{aligned} M_1 &: \text{the proportions of marbles are the same in the two bags} \\ M_2 &: \text{the proportions of marbles are different in the two bags.} \end{aligned}$$

Write out the two models explicitly. Assuming that *a priori* all proportions are equally probable, compute the Bayes factor in favor of M_1.

Hint: Beta distribution is the conjugate prior for the Binomial/Bernoulli likelihood, and a uniform proportion corresponds to the $Beta(1, 1)$ distribution.

(b) The same as (a), but now the first set of draws contains 300 black and 200 white draws, and second set of draws 250 black and 250 white draws.

(c) **NB! This is an optional question for which you can get one extra point to complement missing points in previous or future rounds.** The same as (a), but now in addition to black and white marbles, some marbles may be red (R). The observations are as follows: (B, B, W, B, R, R, W, R) and (R, B, B, B, B, R, R, B).

Hint: The Dirichlet distribution is the conjugate prior for the multinomial/categorical likelihood.

Problem 2. *"Model selection for GMM with BIC and cross-validation."*

In many machine learning applications model selection is crucial. In this exercise, you will practice two common approaches for model selection:

1. Bayesian Information Criterion (BIC) (as an approximation to 'Bayesian model selection').
   Hint: What is the total number of parameters needed to specify the component means and covariance matrices, and the mixture weights? You will need this number to compute BIC.

2. Cross-Validation (as a representative for a predictive model selection criterion).

You are given a data set (1000 samples of dimension 2) contained in the file `data.pickle`, which has been sampled from a Gaussian Mixture Model (GMM) using three classes (the true class labels are given for your convenience, but they should not be used in learning the model).

In the given code template (see below), the data will be divided into training and test sets.

(a) Use both criteria to select the number of components in the GMM using the training data. Plot the both the BIC and the validation log-likelihoods as a function of the number of components, as well as the data with the best model. Do both methods find a model with three components as the most likely?

(b) Use the selected models to evaluate the test set log-likelihood.

(c) Explain briefly the pros and cons of the two approaches and comment which approach you would consider better and why.

Use `ex7_21_template.py` and `ex7_22_template.py` and modify relevant parts into your solution. The module `GMMem.py` provides a set of tools for learning GMMs, which you can use for model fitting (`GMMem.GMMem`) and evaluating the test set log-likelihood (`GMMem.GMMloglik`), see further documentation in the file.

Problem 3. *"Bayesian linear regression model using Edward."*

The file `linear_regression_commented.py` shows an implementation of the Bayesian linear regression model using Edward. The corresponding tutorial is available at http://edwardlib.org/tutorials/supervised-regression.

(a) Run the model.

(b) Analyse the same simulated data set with multiple values for the regularizing hyperparameter (prior std), by changing the value manually over a grid, with log prior std ranging from -4 to 4. Plot the test prediction accuracy (MSE) as a function of the regularizing hyperparameter. Compare the optimal value with the value that can be learned by optimizing the hyperparameter as part of the inference (by defining it as a `tf.Variable`).

(c) Repeat (b) for data set sizes 10, 50, 200 (the size used in (b)). Comment on the results.