**Fontys RobotLab research**

The RobotLab program ‘Big Chemistry’ has received over 90 million euros from the National Growth Fund to position the Netherlands as a global leader in chemical robotics combined with artificial intelligence. By building an autonomous ‘RobotLab’, large numbers of experiments can be carried out, yielding large datasets on properties of molecular systems. The aim is to train new algorithms to predict the properties of molecular systems, e.g. solubility, phase separation, critical micelle concentration, smell, toxicity, and reaction rates.

Fontys contributes to this program through practice oriented research. Central to the research plan is the construction of a robot lab, combining chemical research, high technology (robotics) and artificial intelligence (big data + self-learning systems). Fixed and mobile robots that largely automate the lab will be develop to automate the lab, if not completely, to a large degree. The data collected during chemical expert experiments will help to perform new expert experiments faster and better (recommender system).

**Experiment planning through Bayesian optimization**

In a lot of applications and fundamental research, chemists aim to optimize one or more physical quantities that are a result of a chemical reaction/process. For example, the yield of a certain substance is to be maximised in the least amount of time, consuming the least amount of chemicals. The design of experiments in this context is to be constructed as the way to find the (unique) combination of parameters that may be varied in a chemical experiment (e.g. pH, concentration and nature of a buffer or co-colvent, temperature, etc.) that optimizes some quantity of interest.

Mathematically, this problem may be thought of as a problem of optimizating an expensive to evaluate function (i.e. doing the experiment) in a (typically) high-dimensional space. Bayesian optimization (BO) is emerging as a standard approach to the design of experiments in this context. In BO the quantity of interest, which is regarded as a function of the (large number of) chemical parameters, is modeled as a Gaussian process, giving information about the expected values of, as well as a measure of the uncertainty about the quantity for which the experiments are to be optimized. The first step in BO is to perform a Gaussian process regression to all available experimental data. Once this is complete, a so-called acquisition function, which can be constructed from the Gaussian process, is to be optimized as this function codifies for which combination of chemical parameters it is either possible or likely (exploration vs. exploitation) that the optimum will be reached.

**Treating categorical variables**

In high-dimensional spaces, Gaussian process regression is very effective as long as the dimensions that are to be explored are of a continuous nature. In practice, many of the parameters/dimensions are not of a continuous, but rather a categorical (or binary) nature. It is possible to employ one-hot encoding technique to ‘deal’ with these categorical variables, but this is quickly becoming an ineffective approach if a large number of parameters is of a categorical nature. In the (recent) literature, it is proposed to embed a large number of categorical variables to a low-dimensional continuous latent space. Two seemingly successful embeddings are (1) a (Variational Auto) Encoder-Decoder network or (2) using (random) dictionary embeddings.

**Assignment**

The assignment is as follows:

* Study which approaches (VAE/random dictionary embedding/yet another approach) from the literature are effective in treating categorical variables.
* Implement one or more of these approaches in an (existing) Bayesian optimization code.
* Research and compare the performance of these approaches for some benchmark cases.
* Report concisely on your findings and explain the success/failure of compared approaches.