From Oleg Sysoev

Project 1. PSICA trees were recently proposed for the treatment selection: https://arxiv.org/abs/1811.09065. Your aim is

- Improve performance of the corresponding R package **psica** by using profiling tools in R and by re-implementing decision tree building step in C++.
- Comparing the running time of the current package and the updated version for different data sizes

Project 2. A special kind of longitudinal data is when only inputs are longitudinal but the response is observed in the very end of the process (for example inputs can be physiological measurements of the fetus at several time points and the response is “mortality”). The aim is

- Make a detailed survey of the longitudinal regression methods and find out which ones can handle longitudinal X variables (i.e. $X_1, .. X_t$) but a single response variable Y_t
- Implement one such method (which has no R implementation yet) as an R package
- Compare performance of this method to a Random forest model in which Y_t is response and $X_1, .. X_t$ are predictors for the given medical data set.

(NO VACANT PLACES LEFT) Project 3. A special kind of longitudinal data is when only inputs are longitudinal, but the response is observed in the very end of the process (for example inputs can be physiological measurements of the fetus during several time points and the response is whether the delivery was successful). Your aim is:

- Implement a recurrent neural network predicting Y_t from $X_1, \dots, X_t$ by using keras/Tensorflow
- Compare this network to a usual multilayer perceptron model $Y = f(X_1, \dots X_t)$ by applying it to the given data.

(NO VACANT PLACES LEFT) Project 4. Conformal prediction is a new fast way of computing prediction intervals, see for ex https://ieeexplore.ieee.org/abstract/document/6729517 . Your aim is:

- Implement computing prediction intervals for Decision Trees in R by using a) conformal prediction b) bootstrap and
  - Compare the quality of intervals depending on size of data
  - Compare the CPU time needed depending on size of data

Project 5. Normalization of single cell data is an important step for different genetic analyses, see for example https://www.nature.com/articles/nmeth.4292 The aim is:

- Make a detailed survey of normalization methods available (both for single-cell data and other RNA data)

- Implement one of these methods (which has no implementations in R yet) as an R package and compare the quality of hierarchical clustering when normalization is done and when it is not done

Project 6. Finding important genes that differentiate between different cell types (differential expression analysis) is a key ingredient of single-cell data analyses, see for example https://www.nature.com/articles/nmeth.2967. Your aim is:

- Make a detailed survey of differential expression analysis methods available for single-cell (count) data (both for single-cell data and other microRNA data)
- Implement one of these methods (which has no implementations in R yet) as an R package and compare the quality of hierarchical clustering when differentially expressed genes are used only compared to when all genes are used in clustering.

Project 7&8 (two students are needed)

SPAV algorithm was recently proposed for univariate monotonic regression, see https://link.springer.com/article/10.1007/s10115-018-1201-2

Student 1:

- Implement a GAM algorithm in which some of input variables are spline components and some other input variables are monotonic components that use SPAV algorithm, implement it as an R package.
- Compare predictive power of this algorithm and the usual GAM model that uses splines only (use some datasets where predictors can assumed to be monotonically related to response)

Student 2:

- Make a detailed survey of monotonic regression methods that generate smooth predictions and that can handle many predictors
- Implement one of these methods (which has no implementations in R yet) as an R package and compare this algorithm to the GAM algorithm created by student 1 (use some datasets where predictors can assumed to be monotonically related to response).

Project 9. Causal decision trees were recently proposed to find tree-like causal representations, see https://ieeexplore.ieee.org/abstract/document/7600471. Your aim is

1. Implement this approach as an R package
2. Compare the tree structure produced by this method and by PSICA trees https://arxiv.org/abs/1811.09065 (CRAN package ”psica”) for the given data set

Project 10. In longitudinal clinical trials (patients observed over a time period), some patients may decide to quit the trial. Your task is:

- Make a detailed survey of the methods for the longitudinal data with dropout
- Implement one of these methods (that has no R implementation yet) as an R package
- Compare a conventional method for imputation, for example Random Forest Imputation, with your selected method for a given data set.

Project 11. A number of decision tree methods is proposed for choosing individualized (optimal) treatment regimes, for example https://academic.oup.com/biomet/article/102/3/501/2365724 . Your task is:

- Make a detailed survey of the methods for the optimal treatment regimes
- Implement one of these methods (that has no R implementation yet) as an R package
- Compare predictions done by PSICA trees (package psica) with the implemented method for given data.

Project 12. It is known it telecommunications that if several signals are measured in a given point then the weakest signals might be not heard (i.e. dominated by strongest signals). This can be modeled by censored prediction models like this one: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4280. Your task is to

- Implement https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4280 as an R package
- Find out how the prediction accuracy and the computational time changes with the growing number of features (for given artificial datasets).

From Krzysztof Bartoszek

Predicting birds' flight range
Currently, flight range is estimated using C. J. Pennycuick's Flight program [1]. There are two problems with the program. Firstly, it is available only for Windows and secondly, more importantly, it allows for estimation of a single bird's range with manual imputation of its characteristics. This makes for extremely tedious work when one has measured hundreds or thousands of birds. Therefore, the first aim of the project is to build an R-package, publicly available on CRAN, that implements provided algorithms that estimate the flight range. The package should have the possibility to read-in a file (e.g. .csv) with measurements from multiple birds and automatically calculate the range for all of them.

[1] C.J. Pennycuick. Modelling the Flying Bird. Elsevier (Academic Press), 2008.

From Fredrik Lindsten

Project 1. Evaluating a bias-compensation method for Sequential Monte Carlo

Sequential Monte Carlo methods are a powerful class of computational algorithms for Bayesian inference. One of their key merits is that they provide an estimate of the log-marginal likelihood, which however is biased. In this project we will evaluate empirically if the asymptotic distribution of the log-marginal likelihood estimate can be used to compensate for this bias, and whether or not this improves the estimate compared to conventional methods.

Project 2. Comparing Hamiltonian Monte Carlo and Elliptical Slice Sampling for constrained Gaussian distributions

High-dimensional Gaussian distributions are frequently used in probabilistic machine learning models, such as Gaussian process regression, probit models, Bayesian LASSO, etc. When subject to linear constraints, simulating from such distributions exactly is not possible, and we need to resort to approximate methods. It has been recognized that the Hamiltonian Monte Carlo (HMC) algorithm is well suited for this task, since it is able to exploit the Gaussianity and the linearity of the constraints, to construct an MCMC sampler without rejection. However, the same type of structure exploitation is possible also for the Elliptical Slice Sampler (ESS), another highly efficient MCMC procedure. In this project we will investigate empirically if ESS results in a more efficient method than HMC when applied to high-dimensional Gaussian distributions under linear constraints.

From Patrick Lambrix

Project 1. Performance of players in ice hockey is often measured using traditional stats such as goals, assists, points and +/-. In one of our research projects we use reinforcement learning to define an advanced performance measure related to player impact in connection to goals scored.

In the current method the reward for a state where a goal is scored is 1. This means, for instance, that a goal made when a team is leading by 5 goals is equally important as a goal made when the game is tied. We would like to extend the current method by learning a reward function for the goal states where context is taken into account (e.g., what is the goal difference at the time the goal is scored?, when is the goal scored?). This will allow us to weigh in the importance of particular goals towards winning games in a better way.

From Anders Eklund

(NO VACANT PLACES LEFT) **Project 1.** CycleGAN [1] has successfully been used for unsupervised image to image translation for many different applications, see [2] for an overview of generative adversarial networks (GANs) in medical imaging. In this project the idea is to use CycleGAN to translate images from magnetic resonance imaging (MRI) scanner A to MRI scanner B, since deep learning algorithms trained on images from scanner A do normally not perform as well when tested on images from scanner B. The project will involve running existing Python / Keras / Tensorflow code on different openly available MRI datasets, and to evaluate how well CycleGAN can do the translation in 2D and in 3D.

[1] Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, arXiv:1703.10593

[2] Yi, X., Walia, E., & Babyn, P. (2018). Generative adversarial network in medical imaging: A review. arXiv:1809.07294