CSE 515T (Fall 2019) Project

A portion of your grade this semester will be based on a significant project investigating Bayesian methods in depth. You will have two possible paths to satisfy this project requirement.

The main goal of the project is to give you hands-on experience applying Bayesian methods to a real-world dataset. The use of real-world data can have many interesting (and potentially frustrating!) aspects that are difficult to convey without getting your hands dirty. The scope of the project is intended to be more than a homework problem, but less than a full-fledged research paper.

Note that I don't necessarily expect your idea to "work." Part of research is trying out ideas, regardless of whether they're ultimately successful. If your idea does "work," I expect you to think about why it was succesful. If not, I expect you to think about why not!

Option 1: Novel project with existing data

Some of you may be working on existing projects where the techniques of this class could be brought to bear; for example if you are a Ph.D. student wishing to analyze existing data. In previous semesters, several students have used the project as a springboard for investigations that eventually lead to publications.

If this is the case you may propose a project applying Bayesian techniques to your problem. There is a great deal of freedom in defining this project to suit your needs. However, a critical requirement to this option is that *you be an expert in the data you are proposing.* In particular, the following datasets are *insufficient:*

- Kaggle competitions, etc.
- reference datasets from machine learning, UCI datasets, etc.

However, in previous semesters I have seen people use data from surprising sources that lead to interesting projects! Examples from previous semesters include:

- Split times from cross country races, with the goal of predicting a runner's performance from their previous races.
- Several projects on predicting sports outcomes from various features, predicting tournament outcomes, etc.
- A Bayesian ray tracer based on Bayesian quadrature.

As you can see, you can be passionate about and an expert in data that is not academic/scientific!

If you wish to pursue this route, your project will comprise four parts:

- A one-page proposal outlining the problem you want to solve, its importance and motivation, your proposed solution, and the data will use for your investigation. This will be due **Monday, 4 November 2019.** We will review your proposal and give you feedback. At this stage we may also *reject* your proposal if we feel the problem is not appropriate.
- A two-page status report reporting on your current progress and detailing any issues/questions you may have, due **Friday, 22 November 2019.** We will provide feedback and guidance.
- A four-page final report detailing the final approach and the outcomes of the investigation, due **Monday, 9 December 2019.**
- A ten-minute presentation to the class during the **last week of class.**

This project option can optionally have teams of 2–4 people; however at least 50% of the team members must rise to the level of "experts" on the data. Please comment on this in your proposal.

Option 2: Supervised project

If you do not wish to pursue the first project option, you can instead complete a guided project incorporating many notions from this class following the outline below. There is still considerable room for creativity in this option.

You can work in teams of 2–3 on this project if you desire.

The project will be an investigation into applying Bayesian optimization to automatic and efficient hyperparameter tuning of machine learning models.

We will use two sources of data for this investigation. The first is a classical synthetic benchmark function used in optimization, the *Branin* (sometimes *Branin–Hoo*) function. See here for a description, formula, and some implementations:

```
https://www.sfu.ca/~ssurjano/branin.html
```

This is a function of a two-dimensional input and can be useful for visualizing the behavior of the methods you will implement. The goal is to minimize the function over the specified box-bounded domain.

The second is real data from a hyperparameter tuning task on two different models: SVM and online LDA. These datasets were used in the paper "Practical Bayesian Optimization of Machine Learning Algorithms" by Snoek, et al. The paper is available at

```
http://tiny.cc/bopt
```

You should read it!

Given some data, the performance of each methods was evaluated on a three-dimensional grid of hyperparameter values, defining an objective function $f(\theta)$, where θ is a three-dimensional hyperparameter vector. Our goal is to maximize the performance of the model as a function of the hyperparameters. As the cost of training and evaluating the performance of these models is so great, we will use the precomputed grids instead. The data is available here:

```
https://github.com/mwhoffman/benchfunk/tree/master/benchfunk/functions/data
```

For each problem, the goal is to minimize the value in the fourth column given the values of the hyperparameters in the first three. You can ignore the final column.

Note that all of these are minimization problems.

The project will comprise a series of smaller components that you must complete. You will then compile your findings into a report. I expect every bulleted item to be addressed in the report. The report will be due **Wednesday, 4 December 2019,** the last day of class.

The remaining details of the project will be posted shortly.