MACHINE LEARNING WITH BIOMEDICAL DATA
BioML CLASS PROJECT
Fall 2017

Instructor: Mert R. Sabuncu, ECE/BME, Cornell

RULES and GUIDELINES

1. All deadlines are strict. No extensions will be granted. If no report is received by corresponding deadline, the team will receive a *zero* grade for that component of the project.

   Important Dates (Fall 2017)

   Project Proposal (5% of total class grade) due: 5pm, September 19
   Project Progress Report (5% of total class grade) due: 5pm, October 26
   Project Final Report (10% of total class grade) due: 5pm, November 30
   Project Presentations (10% of total class grade): In class, November 28 and 30

2. You will form teams of 3 (a 2-member team will be allowed if you are going with Option B – see below). You will need to form your team by end of week 3 and inform me by emailing msabuncu@cornell.edu (with SUBJECT LINE: "BioML Class Project Team") by **5pm on September 8, 2017.** Your email should cc all team members, and include a team name and all team member names, with a designated contact person who will upload the reports on blackboard. No form of between-team collaboration is allowed.
3. Each team has two options:
   a. **Option A** (Tadpole): I will provide you with *.csv files that will contain all the data you need[1]. The machine learning objectives will be made very clear. The data are

[1] IMPORTANT NOTE about Tadpole: These data are derived from a publicly-available, de-identified/anonymized human subjects dataset, and as such, you should be extremely careful about safety and security. You should NEVER attempt to: 1) identify the individuals in the data, 2) merge the provided files with other datasets, and 3) use the provided files for purposes other than the class project. You should always keep the files on a secure, password-protected computer, with controlled access. No person who is not a member of your class project team should be able to read these files. Importantly, you should permanently DELETE all data files after the completion of the class project and no later than December 15, 2017.

obtained from an ongoing data science challenge (https://tadpole.grand-challenge.org/) and promising teams will be encouraged to make a submission in mid November. For further details, refer to TadpoleReadMe.pdf.

   b. **Option B** (Alternative biomedical dataset): You can choose any publicly available or private biomedical dataset and identify your own machine learning problem[2]. If you decide to go with this option, you will need to schedule a 30-min meeting with Instructor (email: msabuncu@cornell.edu with subject line that starts with BioML) during Week 3 (week of **September 4, 2017**) to discuss your plans.

4. There are three reports due throughout the semester and an in-class presentation during the final week of classes. All reports will be submitted via **blackboard** (ONLY team contact person should upload) by **5 pm on deadline day.**
5. Proposal Report (pdf file with name: "BioML-TeamX-Proposal.pdf") is due on **September 19.** This report should be no longer than two pages and contain at least following information: team member names, chosen dataset and machine learning problem, a basic literature survey (5-10 most relevant published papers) in the area of choice, planned machine learning strategy (including the metric that will be optimized) and a brief description of the baseline approach(es) (simple prediction algorithm that one would try as a first attempt).
6. Progress Report (file: "BioML-TeamX-Progress.pdf") is due on **October 26.** This report should be no longer than four pages and contain at least following information: Details on the coding environment, implemented pre-processing steps including data clean-up and missing value filling, description of the *implemented* baseline algorithm, results obtained with the baseline algorithm, and preliminary description of the proposed algorithm and preliminary results. If you are participating in the Tadpole challenge (option A above), you will need to report training and testing performance.
7. Final Report (pdf file with name: "BioML-TeamX-Final.pdf") is due on **November 28.** This report should be no longer than eight pages and contain at least following information: description of data, ML problem, literature survey, pre-processing steps, baseline

[2] Data safety, security and compliance to other relevant rules will be your responsibility.

method and its results, proposed/explored/implemented ML methods and corresponding results. A rationale of the explored analytic steps, and *visualization and interpretation* (e.g., variable importance metrics) of the ML models should also be included. The final report should also contain a paragraph describing each team member’s contribution to the project. In addition to the pdf file, the final report email should include a zip folder with all Matlab or Python code that is necessary to reproduce final results. Note that the code should be well documented, describing the individual steps and contain instructions on how to execute on an arbitrary machine.

8. In-class Presentation – during last two lectures (November 28 and 30), each team will give a 10-minute slide (powerpoint, beamer, etc) presentation on their project. These presentations will be graded by your peers, so make sure to make the material accessible to your audience. Figures visualizing the data and machine learning models to help gain insights about the data and the prediction are strongly encouraged.

9. In total, the class project will have a 30% weight in the calculation of your final grade. The breakdown is as follows: 5% proposal, 5% progress, 10% final report, and 10% presentation.

10. The suggested coding environment is Matlab or Python. There is a growing list of Python libraries (E.g., http://scikit-learn.org/) that you can employ. You will need to use your own computing resources to implement and run analyses.