EE 660

Homework 6 (Week 8): Type 2 Project Proposal

Posted: Fri., 10/16/2020 Due: Fri., 10/23/2020, 5:00 PM PDT

**This proposal form is for Type 2 projects: Experimental exploration of machine learning issues.**

**Please fill in both the Project Proposal form (pp. 1-2) and the Dataset Information Form (p. 3)**. **This is required of everyone (each team submits one HW6 with all their names on it).** \*All fields except “other comments” are required. In each field, replace instructions (black text) with your descriptions. Preferred format is to enter your answers into the Word version of this form, then convert to pdf before submission. If you prefer to use another app instead of Word, then submit a typed version with each field labeled with its title (“Clear statement of the problem…”, etc.), and submit as a pdf file.

Please note that this proposal will not be graded like a regular homework. The primary purpose is to give you some feedback on your project topic and plans; the scoring on this homework will be primarily based on whether you put in a reasonable effort, whether the content makes good technical sense, and whether you responded to the fields as required.

|  |
| --- |
| \*Insert Project Title Here |
| \*Project team: Your name(s) and email address(es) |
|  |
| \*Clear statement of the problem, machine learning issue, or variable dependence you will investigate |
| This will guide your work throughout the project period, and help you define when you have completed your project work. (For example, is it to  study how some factors relate to a phenomenon e.g. overfitting? Or is it to find the probability that a model selection module does pick the best model, and what parameters will influence the probability? Other examples and ideas are listed in Type 2 Project Detail.) |
| \*Your plan for dataset, any preprocessing and feature extraction |
| Describe your plans, any preprocessing you expect to try, and similarly for feature extraction if any.  If you use synthetic data, describe the dataset overall and how it relates to your project plan (including how it relates to your problem statement above). (Detailed description of your dataset will be given on the dataset form.) |
| \*A plan of your experiment(s) (approach) |
| Describe what experiments you intend to perform. Give a preliminary plan on how you will proceed. This could include, for example, models/methods you will use to do experiments for explaining and interpreting a theory; how you will use the dataset (test sets, out-of-sample data, training sets, etc.); what parameters will be varied and what results you will show.  Your plan can be revised as you do the work, because you will learn more about the problem as you go. But having an initial plan for your approach is helpful. |
| \*What you hope to learn or demonstrate or refute |
| How will you interpret your results? What conclusions do you hope to draw? Will your results be compared with anything (e.g. theory, general principles, common practice, etc.)? |
| \*A description of any other work of yours that is related to your class project |
| If none, state “none”. Otherwise, briefly summarize your related work here, and state how your EE 660 class project work will be different. Examples of related work you need to describe here: research work on the same or similar topic; work for hire (at your place of employment) on the same or similar topic; work for other classes (projects, etc.) on the same or similar topic. |
| \*If yours is a team project, roughly describe how work will be divided |
| Describe how you are planning to divide up the work and how you will collaborate. (Contributions to all aspects of the project (theory, coding, planning, analysis, writing etc.) will be stated in the final report.) |
| Other Comments |
| Anything else you think should be included. For example, if you see potential problems that might arise, you can mention them here with ideas of how you might deal with them. Or, anything that is particular to your project idea that wasn’t requested in the form fields above. |

EE 660 Synthetic Dataset Information Form Fall 2020

*This form is for synthetic datasets only.*

*Include one form for each dataset you plan to use. (For each dataset’s form, you may continue onto an additional page if necessary.)*

**Note:**  for some of the aspects requested below, you might not yet have an exact answer. Please answer based on your preliminary plan. For quantities that will be variable over some range, it is understood that it may make sense to change the range covered after you’ve gotten some preliminary results.

\*Problem type: regression, classification/logistic regression

**\*Brief description of dataset**:

Include target function f(x), any parameters in f(x) that will be varied (give preliminary ranges); anticipated values of N.

**\*Will all data drawn be i.i.d?**  Yes, No.

**If no, \*please describe**.

**\*What will underlying pdf’s be? Will the feature dimensions be independent or dependent?**

Give the pdf’s: e.g., p(x) and p(y|x); or p(y) and p(x|y); or p(x) and p(n) with y = f(x) + n or y = sign {f(x) + n}.

Give (preliminary) values for parameters like mean and variance; if they will be varied, then leave as variables and give preliminary ranges.

If features are dependent or correlated, state how (e.g., give covariance matrix for a multi-variate Gaussian); use variable(s) with preliminary ranges if appropriate.

**\*Will all feature dimensions have same underlying pdf’s?** Yes, No.

**If no, \*how will the difference be?**

**\*Has Noise?** Yes, No.

**If yes, \*what kind of noise?** e.g., is noise additive? pdf of the noise? State “given above” if already stated in “underlying pdf’s” field above.

\*Label (output) type: numeric, binary categorical, or multi-class categorical

**\*If Label Type is Categorical, will the number of samples across classes be balanced or unbalanced ?**  Balanced, unbalanced.

**If unbalanced, \*indicate how unbalanced (ratio of class representation, or a range for it if variable),**

**\*If other aspects will be variable, how will they be varied?**

e.g. number of points – Over what preliminary range would N vary?

e.g. number of features – How will the number of features (dimensions) be varied?

e.g. variance of noise

**Any other comments on the dataset:**

If there are other aspects of the dataset relevant to your project that weren’t asked for above, please state them here (e.g., if you choose to make some (or all) features non-numeric).

EE 660 Real-world Dataset Information Form Fall 2020

*Only include this dataset form if you intend to use a real-world dataset.*

*Include one form for each dataset you plan to use. (For each dataset’s form, you may continue onto an additional page if necessary.)*

\*Dataset or competition title:

\*Link:

**\*Problem type**: regression, classification/logistic regression

**\*Brief description of dataset and problem domain**:

**\*Number of data points**:

**\*Number of features or input variables**:

\*Feature or input-variable types: numeric (how many variables and what type), categorical (how many variables), other (describe and give how many variables)

\*Label (output) type: numeric, binary categorical, or multi-class categorical

**\*If Label Type is Categorical, is the number of samples significantly unbalanced (maximal variation of more than a factor of 2)**? Yes, No.

**If yes, \*rate as:**

significant (maximal variation is factor of 2 to 10)

major (factor of 10 to 100)

extreme (more than factor of 100))

**\*Has Missing Data**? Yes (give idea of how prevalent, if known), or No.

If yes, try to explain how you will deal with it in the “A plan of preprocessing and feature extraction” section in the Project Proposal

**\*Is the problem/dataset a Kaggle competition (current or past)?** yes or no

**If yes, answer:**

**\*(i) Is the competition current (give the end date), or past**?

**\*(ii) How much information is available on the Kaggle website (e.g., in “kernels” and links therein)**? Briefly describe what type of information and code is available.

**Any other comments on the dataset:**