CS 584 – Data Mining

Project Guidelines[1]

For the class project, follow these procedures:

- Form a team of two students.
- Decide on a problem that your team would like to work on. For example, find a sufficiently large dataset and do something interesting with the data. Provide background and motivation in your proposal (and subsequently, project report). Explain why the problem is interesting.
- *All proposals need to be approved.* Your grade for the project will depend on several factors, including challenge level of the problem, novelty, demonstration of knowledge in course materials, evidence of research effort and depth of thinking, experimental design and evaluation, analysis of results, the report itself and the presentation.
  a. To help narrow down the topic that you'd like to work on, you can think of the project in terms of two focuses: application and/or methodology. For application, the focus is on solving a real-world problem (see Kaggle competitions, for example), but there should be some challenges that you try to address. For methodology, the focus is on the development of novel techniques that improve upon existing methods. Of course, your project may very well consist of both components.
  b. If you are unsure whether your project is challenging enough, compare it with the homework assignments. **It should be more challenging than the homework**. If you can apply some existing algorithm (e.g. decision tree) quickly without much effort and still get reasonable results, then the problem is not challenging enough. You can, however, use that as a baseline and try to improve the performance in some non-trivial ways.
  c. Your HW3 will be on clustering, and HW4 will be on recommendation systems.
- See a list of possible datasets in the bottom of this document. Feel free to find/create your own dataset. The dataset should contain a minimum of 10,000 instances. However, meeting the minimum would not guarantee an A in the project. See the grading criteria above.
- Your team will do a 2-minute "project pitch." The purpose of the project pitch is for you to see what other teams are planning to do, and for you to get some feedback.
- Write a 1-page project proposal, due a week after the project pitch. Your project proposal should be structured into the following sections (it should concisely answer the following questions):
  - **What is the problem your team is solving?** Give a brief but precise description or definition of the problem.

[1] Adapted from CS341 at Stanford, Project in Mining Massive Datasets

  - **What are the challenges of the problem?** Discuss the challenges you foresee, e.g. imbalanced data, large feature set, large data size, noisy data, etc.
  - **What data will you use?** Briefly describe the data, the sizes (number of records and features, file size) and where will you get the data.
  - **What's the state of art?** Do some research on related work (approaches other people have used or developed).
  - **How will you solve the problem?** Describe your approach: what method, algorithm, or technique do you plan to develop or use? *Be as specific as you can!*
  - **How will you evaluate your method?** Describe how you will measure performance or success of your method. Against what baseline methods will you compare your algorithm or how do you plan to obtain ground-truth labeled data so that you can then measure accuracy, precision, recall or some other metric that will tell me how well is your method really performing.
- Write a project report that is well-formatted, using the ACM template (https://www.acm.org/publications/proceedings-template). The report should be at least 5 pages long. Describe the problem, related work, your approach, experimental results, and your analysis (e.g. why do you think the method performed well/not well).

  The report should have the following sections:

  - **Abstract**: Summary of the report.
  - **Introduction:** Talk about motivation of the problem; provide a description or definition of the problem or hypothesis you set to evaluate.
  - **Related Work:** Describe what other people have done on the same or similar problem. What are the advantages/disadvantages of those techniques? Why is your proposed solution better, or why did you choose your proposed solution over the other existing techniques?
  - **Solution:** How did you solve the problem? Describe the technical approach. Tell us what method/algorithm you use, develop or extend and how you implement it.
  - **Experiments:**
    - **Data:** Briefly describe the data and its size (number of records, file size)
    - **Experimental setup:** Describe how did you setup your experiments, how the training/testing data was prepared, what performance metrics are you considering, what baseline methods for comparison are you using.
    - **Experimental results and analysis:** Describe your experimental results. Structure your experiments around particular aspects of your method. Some ideas: (1) a table showing results of your method using different types of features; (2) a table showing the results of your method using different parameter settings; (3) table comparing the performance of your method to the baselines; (4) a graph plotting the size of the training dataset vs. the time it takes to train the model; (5) Investigation of the learned model (what are the important features, etc.).

  - **Brief analysis & conclusion**
  - **At the end of the paper, also describe the contribution of each team member.**
- The project will be 35% of your overall grade. Here are the percentage breakdowns:
  - Proposal: 5%
  - Presentation: 5%
  - Code & Report: 25%

Resources (software/datasets/ideas):

- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/[2]
- http://www-users.cs.umn.edu/~kumar/dmbook/resources.htm
- http://www.stanford.edu/class/cs341/data.html
- Kaggle Competitions[3]: http://www.kaggle.com/competitions
- KDnuggets: https://www.kdnuggets.com/datasets/index.html
- 2015 ECML/PKDD Discovery Challenge: http://www.ecmlpkdd2015.org/discovery-challenges
- 2015 PAKDD Contest: https://knowledgepit.fedcsis.org/contest/view.php?id=107
- Google for more of such contests

Important Dates:

3/17 – Project Pitch (in class)

3/24 – proposal due (11:59pm)

5/12 – presentations

5/12 – project due date

[2] The archive contains datasets that vary in size, complexity, dimensionality, etc. Some may be too easy for the project.

[3] Kaggle competitions vary in different degrees of difficulty. Obviously, if you pick a project that's not very challenging, then you won't get very high score.