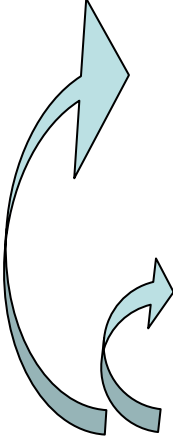




Some Tips on Project Proposal

April 15, 2010

Course Project

- 
1. Start with an interesting task and find real-world data
 2. Perform **research** to find out appropriate data mining / machine learning algorithms
 3. Implement several different algorithms
 4. Evaluate the performance of the algorithms on data
(if unsuccessful, return to step 3, 2, or 1)
 5. Write up results

Please be proactive!
Please be creative!

Components of Proposal

- Introduce task(s)
- Describe data set(s)
- Propose potential algorithm(s)
- Discuss potential evaluation strategies
- Review related works
- Propose plan for completion

Introduce task(s)

- What real-world problem are you tackling?
- Why is solving this problem important? In the proposal you would need to provide a **motivation** for performing this project.
- Specifically, what machine learning task(s) are you proposing?
 - Classification
 - Regression
 - Clustering

Describe data set(s)

- Every project would need to be centered around data set(s)
- If you have the data set(s) already, please describe details
 - How many training examples? (the bigger the data set, the better...)
 - How many features? What type of features? Is there missing data?
 - What is unique about this data? You can perform quick EDA...
- If you don't have a data set yet, you **must** provide a concrete plan for obtaining the data in the proposal
 - e.g. if you wish to scrape data from the web, mention the URLs
 - by the first progress report (Week 5) you must have a data set.

Propose Potential Algorithms

- We haven't covered many data mining / machine learning algorithms yet, so do please just do your best on this section.
- This is where reading various related papers will help...
- Propose more than one algorithm:
 - “Baseline” algorithms: kNN, logistic regression, linear regression
 - More complicated algorithms: SVMs, neural networks, etc.

A quick listing of algorithms

- **Classification:**
 - Nearest neighbors, logistic regression (binary), multinomial logit (extension of logistic regression), neural networks, naïve Bayes, support vector machines (SVMs), decision trees, etc.
- **Regression:**
 - Linear regression, generalized linear regression, regression trees, etc.
- **Clustering:**
 - K-means, mixture of Gaussians, agglomerative/divisive clustering, etc.
- **Dimensionality reduction:**
 - Principal component analysis (PCA), singular value decomposition (SVD), topic models (for text), etc.

Some algorithms only work with certain types of data (e.g. interval data)

A quick listing of algorithmic techniques

- **Ensemble methods:**
 - Bagging (take average of many different classifiers/regressors)
 - Boosting (adaptively weight each data case, e.g. AdaBoost)
 - Random forests (combining multiple decision trees)
- **Regularization (for regression):**
 - L2 regularization (“ridge regression”, “tikhonov regularization”)
 - L1 regularization (“LASSO”) -- this is harder case
- **Manipulating features:**
 - Centering, standardizing, converting categorical to binary
 - Feature selection techniques, adding nonlinear features (e.g. $x_1 * x_2$)
- **Optimization:**
 - Gradient descent
 - Newton’s method
 - Conjugate gradient
 - Grid search
 - Stochastic search
 - Beam search

There are many papers, tutorials, slides, and Wikipedia pages available on these topics

Proposed evaluation

- For classification/regression:
 - Learn model on training set, and calculate accuracy/error on test set
 - Can be a function of many different quantities (e.g. amount of training data, number of features, complexity of model, characteristics of data set)
 - Checks to see if you are overfitting/underfitting
 - “Cross-validation”
 - More specialized metrics for various tasks, e.g. root mean squared error (RMSE), expected reciprocal rank (ERR)
- For clustering:
 - Visualize clusters and see if it “looks” correct
 - If probabilistic model, evaluate likelihood on test data

Related work

- Read 3 “papers” related to your project
 - Can be related to the domain (e.g. sports statistics + machine learning)
 - Can be related to the algorithms that you potentially would use (e.g. ranking algorithms for yahoo challenge)
 - Can be “tutorial” or “survey” papers
- How to find these papers?
 - Google scholar, ACM digital library, Citeseer, etc.
 - If you find 1 good paper, do a breadth-first search on the references
 - Skim papers (read abstract) and if it is not related, move to the next one
- Would need to summarize these papers (e.g. what were their results and how can this paper be potentially useful to your project) and cite these papers in your bibliography

Plan for completion

- Suggest a rough timeline for completing the project:
 - What do you plan to accomplish each week?
 - Will you have weekly team meetings?
 - How do you plan to divide work among team?
 - Will you use third-party tools or code the algorithms yourself?
 - What “bottlenecks” do you foresee?

You are allowed to use third-party tools such as SVM-light, Weka, etc.

Format of Proposal

- Use NIPS conference format
 - <http://nips.cc/PaperInformation/StyleFiles>
 - Should be 4 pages
 - You can use the Word (.rtf) template
- 5% extra credit for using LaTeX system
 - Very easy to learn:
 - <http://heather.cs.ucdavis.edu/~matloff/latex.html>
 - Requires unix/linux environment to “compile” (for Windows, use Cygwin)

Quick Presentation!

- On Tuesday, April 20, one member of your team will be **required** to give a 2 minute “lightning” presentation of your project proposal
- Please prepare one Powerpoint slide (.ppt) to augment your presentation
- Please be ready with a USB key