

CSE 6740 Lecture 7

How Do I Start Doing Machine Learning? (The Georgia Tech Machine Learning Project)

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Quiz Answers

1. Learning a 2-class generative classifier boils down to learning two separate density estimators. T.
2. The nearest-neighbor rule is a generative classifier. F.
3. The nearest-neighbor rule is a nonparametric classifier. T.

Today

1. How the project works
2. The topics

My way of making your project *real*. You will develop good machine learning code that other people can use.

Motivations for Project Structure

- **Learning by doing.** Implementing the math as code is the only way to really understand things.
- **Creating code for the field.** There are existing collections of machine learning code, but there is a need for a *comprehensive* collection of *efficient* code.

FASTlib and MLPACK

MLPACK is what the collection will be called. Its first public release will be at the end of the semester. Its codes are built using the *FASTlib* C++ library, which features:

- Support for standardized argument passing and management of experiment output
- Linear algebra methods (through LAPACK and Trilinos)
- Optimization and other numerical methods (through Numerical Recipes)
- Data structures for fast discrete algorithms
- Use of templates for elegance/comprehensibility
- Memory and CPU efficiency
- Standards for code reviews, unit testing, and consistent style

Learning FASTlib

- Tutorial code example
- About 15-20 ML methods to look at
- In the manual: FAQ and “cookbook”
- If all else fails: FASTlib development mailing list

Steps

- Choose the ML methods you'll implement (starts today)
 - two people per method, with slightly different twists
- Implement your method completely on your own
- Compare your implementation with your partner's, for correctness, and efficiency; review your partner's code for unit testing, style, documentation and correctness/performance on standard datasets
- Merge to create one function for the ML method, with options, and write up its compliance checklist – project score will be largely based on this merged code
- Review an anonymous team's merged code

Comparisons

As part of the code testing requirement, the methods will be compared statistically and computationally on a selection of provided datasets:

MSE:

Method	Synth Data 1	Astro Data
Linear regression	6.3	8.4
Nadaraya-Watson regression	1.1	2.3

CPU seconds:

Method	Synth Data 1	Astro Data
Linear regression	33	57
Nadaraya-Watson regression	109	483

Scoring

- To get an A, you need a total of 2 points minimum on the first part of the project and 3 points on the second part.
- Each ML method is worth 1-3 points depending on overall difficulty. You can do as many as you want.
- You can also implement in Matlab, which gets 1/3 the amount of points you would have gotten with a FASTlib implementation. Matlab implementors can only pair with each other.
- The person with the most points at the end of the class will be bestowed with the title of “King of Machine Learning” or “Queen of Machine Learning” for 2008.

The Methods

- For the first part, you'll choose from my list of standard ML methods. For the second, you can choose a recent ML paper (subject to my approval), worth 3 points except in unusual circumstances, or propose extensions to the existing methods in the collection, worth 1/3 point each.
- Which methods to pick? Based on your aptitudes and research interests.
- Only two people can do each method (per language), so we will have a priority system for choosing assignments.
- There are “sequences” – a method can't be done unless the some team has done or is doing the first method in the sequence – e.g. LASSO can't be done before linear regression.

The Methods

“Ensembles” sequence:

1. Boosting, bagging, and stacking (1)
2. Feature selection by forward and backward selection (1)
3. Bootstrap and RANSAC (1)
4. Covariate shift training (1)

The Methods

“Linear regression” sequence:

1. Linear regression and ridge regression (2)
2. LASSO (2)
3. Variational linear regression (2)
4. Gaussian process regression and classification (3)

The Methods

“Neural networks” sequence:

1. Perceptron and multi-layer neural network: regression and classification (2)
2. Mixture density networks (2)

The Methods

“PCA” sequence:

1. PCA (2)
2. CCA and factor analysis (2)
3. Principal curves (2)
4. Autoassociative networks (2)

The Methods

“Decision tree” sequence:

1. CART decision tree: classification and regression (3)

The Methods

“SVM” sequence:

1. Relevance vector machine (3)
2. Distance-weighted discrimination (3)