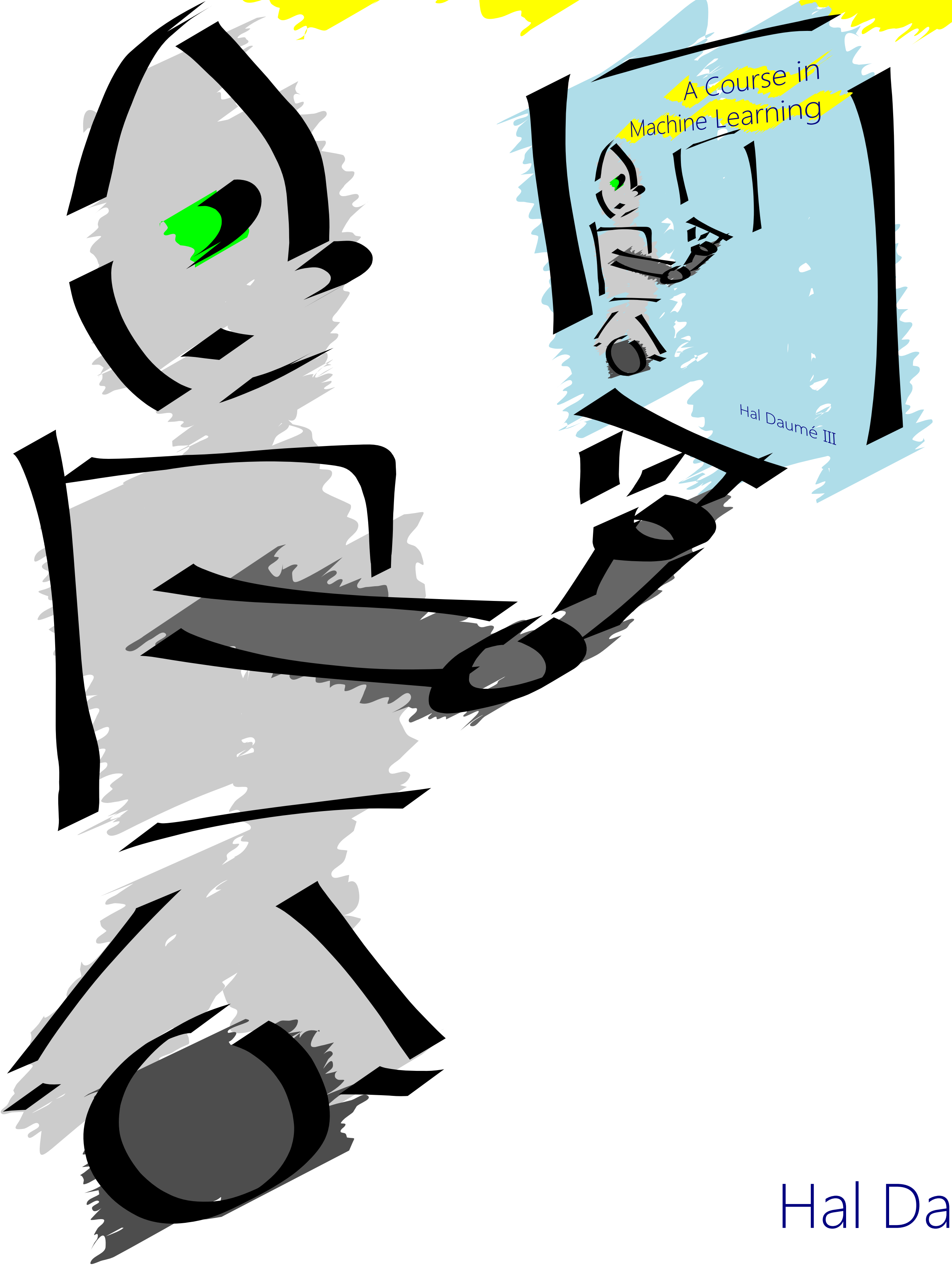
A Course in Machine Learning



Hal Daumé III

http://ciml.info

version 0.8 , August 2012

15 | Semi-Supervised Learning

--

Learning Objectives:

- Explain the cluster assumption for semi-supervised discriminative learning, and why it is necessary.
- Dervive an EM algorithm for generative semi-supervised text categorization.
- Compare and contrast the query by uncertainty and query by committee heuristics for active learning.

You may find yourself in a setting where you have access to some labeled data and some unlabeled data. You would like to use the labeled data to learn a classifier, but it seems wasteful to throw out all that unlabeled data. The key question is: what can you do with that unlabeled data to aid learning? And what assumptions do we have to make in order for this to be helpful?

One idea is to try to use the unlabeled data to learn a better decision boundary. In a discriminative setting, you can accomplish this by trying to find decision boundaries that don't pass too closely to unlabeled data. In a generative setting, you can simply treat some of the labels as observed and some as hidden. This is **semi-supervised learning**. An alternative idea is to spend a small amount of money to get labels for some subset of the unlabeled data. However, you would like to get the most out of your money, so you would only like to pay for labels that are useful. This is **active learning**.

Dependencies:

15.1 EM for Semi-Supervised Learning

naive bayes model

15.2 Graph-based Semi-Supervised Learning

key assumption
graphs and manifolds
label prop

15.3 Loss-based Semi-Supervised Learning

density assumption
loss function
non-convex

15.4 Active Learning

motivation

qbc
qbu

15.5 Dangers of Semi-Supervised Learing

unlab overwhelms lab
biased data from active

15.6 Exercises

Exercise 15.1. TODO. . .