# W4240/W6240

# Data Mining/ Statistical Machine Learning

Frank Wood

January 17, 2012

# Introduction

- Data mining is the search for patterns in large collections of data
  - Designing models
  - Fitting models to data
  - Using models to perform inference/prediction
- Pattern recognition is concerned with *automatically* finding patterns in data / learning models
- Machine learning is pattern recognition with concern for computational tractability and full automation
- Data mining = Machine Learning = Applied Statistics
  - Scale
  - *Computation*

# High Level Course Goals

- Problem formulation
  - Starting from data and/or a question, you will learn how to create and design a model that will answer the question(s) of interest.
  - You will learn how to formally, mathematically codify a model.
  - You will learn how to fit models.
  - You will learn to think about computational/inferential trade-offs.
- Tool *Creation*
  - You will learn how to design and implement general purpose learning algorithms
  - You will implement inference algorithms for several models such as latent Dirichlet allocation, Bayesian logistic regression, Gaussian mixture models and more.
- Practice
  - Through your homework and final project you will be evaluated on how well you can put into practice the theory that will be taught in class.

# Style of Instruction

- Great textbook (Bishop, Pattern Recognition and Machine Learning, *required!*)!
  - Will follow second half of text closely.
- Class time will be spent on *theory* and "*proof*," explaining the tricky parts of the text by doing math on the board.
- Homework (extensive and hard) will be spent on practice.
- Team-based final project
- This is *not* a *"What tool do we use and how do we use it?"* course!

# Grading - What you should expect.

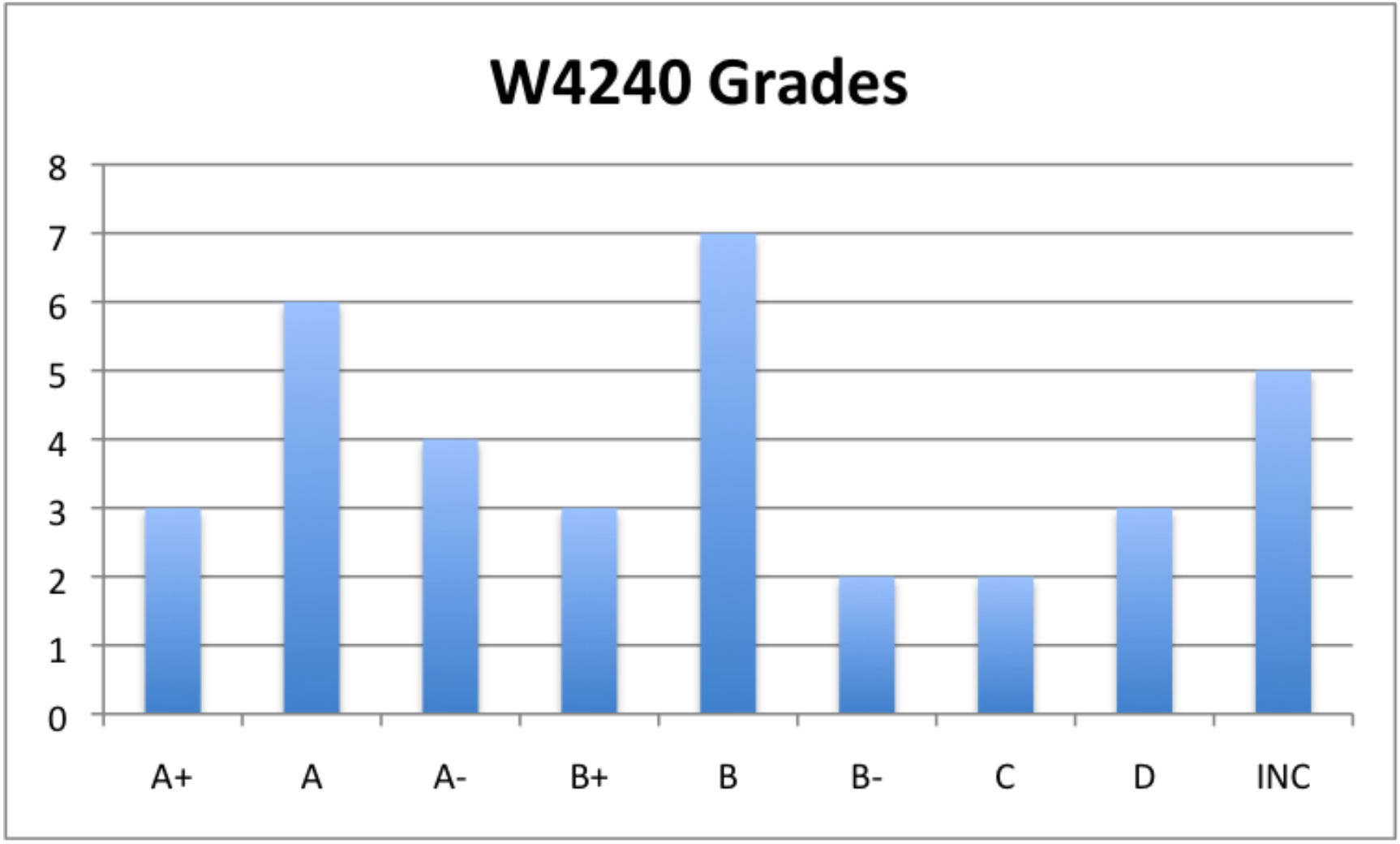


Figure: Fall 2010 Grade Distribution

# Links and Syllabus

- Course home page :
  http://www.stat.columbia.edu/∼fwood/w4240/
- *Bookmark* this page (not courseworks)
- Guest lectures may be sprinkled throughout the course.

# Prerequisites

- Linear algebra
- Multivariate calculus (matrix and vector calculus)
- Probability and statistics at a masters level
- Programming experience in some language like pascal, matlab, c++, java, c, fortran, scheme, etc.
- Some algorithmic complexity theory (basics) useful
- Information theory (entropy, KL-divergence)

# Review

Good idea to familiarize yourself with PRML [1] Chapter 1 and 2 and Appendices B,C,D, and E. In particular:

- Information theory
- Multivariate Gaussian distribution
- Discrete, multinomial, and Dirichlet distributions
- Lagrange multipliers
- Matlab

We will offer extra Matlab programming sections, a review of both the multivariate Gaussian and information theory.

# Homework (from anonymous student feedback)

> *The assignments were challenging and probably the best part of the course. It was with bated breath, and nervous anticipation - almost like a blind date with someone you knew would be pretty - that we waited for the homework ok, maybe that was 10% exaggerated. ....*
>
> *The assignments in this course were more challenging than any other course I have taken in the past.*
>
> *Really enjoyed the assignments in this class. Excellent job putting them together.*
>
> *Very difficult programming assignments, and I learned a lot doing them,*

The homework in this course is *programming*. You will *implement* various data mining / machine learning algorithms. If you have not programmed before the course is doable but difficult.

# Project

The final project is a semester-long, significant piece of team-based work that demonstrates your data mining / statistical machine learning knowledge on a problem domain of interest to you (and hopefully also of interest to a larger academic, governmental, or industry community). The deliverables include a 2-3 page proposal; a short, publication-quality paper; and a 10-20 minute presentation.

# Syllabus

- Review
- Graphical models
  - Belief propagation
- Expectation Maximization
- Variational Inference
- Sampling
- Misc.

Along the way you will implement estimation and inference procedures for the following models (minimally)

- Gaussian mixture model
- Bayesian linear regression
- Bayesian logistic regression
- Latent Dirichlet allocation
- ...

# Past Example Projects

See website for links to related talks.

# Bibliograpy I

[1] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006. ISBN 0387310738. doi: 10.1117/1.2819119. URL `http://www.library.wisc.edu/selectedtocs/bg0137.pdf`.