CMSC 422: Machine Learning

Hal Daumé III

01 Sep 2015

Course Background

What is this course about?

- Finding (and exploiting) patterns in data
- Replacing “human writing code” with “human supplying data”
  - $\Rightarrow$ System figures out what the person wants based on examples
  - $\Rightarrow$ Need to abstract from “training” examples to “test” examples
  - $\Rightarrow$ Most central issue in ML: *generalization*

Course Background

What is this course about?

- Finding (and exploiting) patterns in data
- Replacing “human writing code” with “human supplying data”
  - ⇒ System figures out what the person wants based on examples
  - ⇒ Need to abstract from “training” examples to “test” examples
  - ⇒ Most central issue in ML: *generalization*

Why is machine learning so cool?

- Broad applicability
  - Finance, robotics, vision, machine translation, medicine, etc.
- Close connection between theory and practice
- Open field, lots of room for new work
- `http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9026623`

Course Goals

By the end of the semester, you should be able to:

- Look at a problem and identify if ML is an appropriate solution
- If so, identify what types of algorithms might be applicable
- Apply those algorithms
- Conquer the world

In order to get there, you will need to:

- Do a lot of math (calculus, linear algebra, probability)
- Do a fair amount of programming
- Work hard (this is a 3-credit class)

Student comments on this course

- The field of machine learning seems too wide for any one-semester long course to completely cover everything

Student comments on this course

- The field of machine learning seems too wide for any one-semester long course to completely cover everything
- This course shouldn't be CS 422, it should be MATH 422!

Student comments on this course

- The field of machine learning seems too wide for any one-semester long course to completely cover everything
- This course shouldn't be CS 422, it should be MATH 422!
- All in all, great class, and clear interest in the needs of the students

Student comments on this course

- The field of machine learning seems too wide for any one-semester long course to completely cover everything
- This course shouldn't be CS 422, it should be MATH 422!
- All in all, great class, and clear interest in the needs of the students
- [The professor] needs to learn how to write proper exams

Student comments on this course

- The field of machine learning seems too wide for any one-semester long course to completely cover everything
- This course shouldn't be CS 422, it should be MATH 422!
- All in all, great class, and clear interest in the needs of the students
- [The professor] needs to learn how to write proper exams
- . . . and so on . . .

Student comments on this course

- The field of machine learning seems too wide for any one-semester long course to completely cover everything
- This course shouldn't be CS 422, it should be MATH 422!
- All in all, great class, and clear interest in the needs of the students
- [The professor] needs to learn how to write proper exams
- . . . and so on . . .

I try to take your comments seriously!
(but some things won't change. . .)

Topics Covered

- Supervised learning: learning with a teacher
- Unsupervised learning: learning without a teacher
- Complex settings: learning in a complicated world

Topics Covered

- Supervised learning: learning with a teacher
- Unsupervised learning: learning without a teacher
- Complex settings: learning in a complicated world
  - Time-series models
  - Structured prediction
  - Semi-supervised learning
  - Large-scale learning

Topics Covered

- Supervised learning: learning with a teacher
- Unsupervised learning: learning without a teacher
- Complex settings: learning in a complicated world
  - Time-series models
  - Structured prediction
  - Semi-supervised learning
  - Large-scale learning

- Not a zoo tour!
- Not an introduction to tools!

Topics Covered

- Supervised learning: learning with a teacher
- Unsupervised learning: learning without a teacher
- Complex settings: learning in a complicated world
  - Time-series models
  - Structured prediction
  - Semi-supervised learning
  - Large-scale learning

- Not a zoo tour!
- Not an introduction to tools!

- You will learn how these techniques work and how to implement them.

Syllabus

On Piazza, linked from

`http://hal3.name/`

Reading: I expect you to do it!

(but most are $\leq$ 12 pages, all are $\leq$ 20)

Online book draft (missing some figures) linked off the web page.

(Extra credit for new/unique bugs submitted or *fixed* on GitHub!)

Reading: I expect you to do it!

(but most are $\leq$ 12 pages, all are $\leq$ 20)

Online book draft (missing some figures) linked off the web page.

(Extra credit for new/unique bugs submitted or *fixed* on GitHub!)

Class time is for:

- Discussing questions from the reading
  - There are questions in the margins: be prepared to answer them
- Discussing homework assignments / labs
  - On each homework you must answer a “what didn’t you understand” question
- Me providing an insider’s view

Things you need to do now!

Complete Homework 00

- Due 08 Sep (that's next Tuesday!, by 8:00a)
- Submit using handin

Complete the first reading

- See syllabus
- Due by class Thursday (I mean it!)

Sign up to get mails

- Subscribe to the Piazza group.
- But be sure to actually read it!

Read the Piazza course description information!

Now, on to some real content. . .

(but first, questions?)

Now, on to some real content. . .

(but first, questions?)

- Training/testing

Classification

- How would you write a program to distinguish a picture of me from a picture of someone else?
- How would you write a program to determine whether a sentence is grammatical or not?
- How would you write a program to distinguish cancerous cells from normal cells?

Classification

- How would you write a program to distinguish a picture of me from a picture of someone else?
  - ⇒ Provide examples pictures of me and pictures of other people and let a *classifier* learn to distinguish the two.
- How would you write a program to determine whether a sentence is grammatical or not?
- How would you write a program to distinguish cancerous cells from normal cells?

Classification

- How would you write a program to distinguish a picture of me from a picture of someone else?
  - ⇒ Provide examples pictures of me and pictures of other people and let a *classifier* learn to distinguish the two.
- How would you write a program to determine whether a sentence is grammatical or not?
  - ⇒ Provide examples of grammatical and *ungrammatical* sentences and let a *classifier* learn to distinguish the two.
- How would you write a program to distinguish cancerous cells from normal cells?

Classification

- How would you write a program to distinguish a picture of me from a picture of someone else?
  - ⇒ Provide examples pictures of me and pictures of other people and let a *classifier* learn to distinguish the two.
- How would you write a program to determine whether a sentence is grammatical or not?
  - ⇒ Provide examples of grammatical and *ungrammatical* sentences and let a *classifier* learn to distinguish the two.
- How would you write a program to distinguish cancerous cells from normal cells?
  - ⇒ Provide examples of cancerous and *normal* cells and let a *classifier* learn to distinguish the two.

Data (“weather” prediction)

Example dataset:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Data ("weather" prediction)

Example dataset:

Class	**Outlook**	**Temperature**	**Windy?**
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

1. Class label
2. Features
3. Feature values

Data (“weather” prediction)

Example dataset:

Class	**Outlook**	**Temperature**	**Windy?**
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

1. Class label (aka “label”, denoted y)
2. Features
3. Feature values

Data ("weather" prediction)

Example dataset:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

1. Class label (aka "label", denoted y)
2. Features (aka "attributes")
3. Feature values

Data ("weather" prediction)

Example dataset:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

1. Class label (aka "label", denoted y)
2. Features (aka "attributes")
3. Feature values (aka "attribute values", denoted x)

Data ("weather" prediction)

Example dataset:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

1. Class label (aka "label", denoted y)
2. Features (aka "attributes")
3. Feature values (aka "attribute values", denoted x)

   $\Rightarrow$ Features can be binary, nomial or continuous

Data ("weather" prediction)

Example dataset:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Three principle components:

1. Class label (aka "label", denoted y)
2. Features (aka "attributes")
3. Feature values (aka "attribute values", denoted x)
   $\Rightarrow$ Features can be binary, nomial or continuous

A *labeled* dataset is a collection of (x, y) pairs

Data ("weather" prediction)

Example dataset:

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Task:

Class	Outlook	Temperature	Windy?
???	Sunny	Low	No

Predict the class for this "test" example"

Data ("weather" prediction)

Example dataset:

Class	**Outlook**	**Temperature**	**Windy?**
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

Task:

Class	**Outlook**	**Temperature**	**Windy?**
???	Sunny	Low	No

Predict the class for this "test" example"

Requires us to **generalize** from the training data

Ingredients for classification

Whole idea: Inject *your* knowledge into a learning system

Sources of knowledge:

1. Feature representation
2. Training data: labeled examples
3. Model

Ingredients for classification

Whole idea: Inject *your* knowledge into a learning system

Sources of knowledge:

1. Feature representation
   - Not typically a focus of machine learning
   - Typically seen as “problem specific”
   - *However*, it’s hard to learn on bad representations
2. Training data: labeled examples
3. Model

Ingredients for classification

Whole idea: Inject *your* knowledge into a learning system

Sources of knowledge:

1. Feature representation
   - Not typically a focus of machine learning
   - Typically seen as “problem specific”
   - *However*, it’s hard to learn on bad representations
2. Training data: labeled examples
   - Often expensive to label lots of data
   - Sometimes data is available for “free”
3. Model

Ingredients for classification

Whole idea: Inject *your* knowledge into a learning system

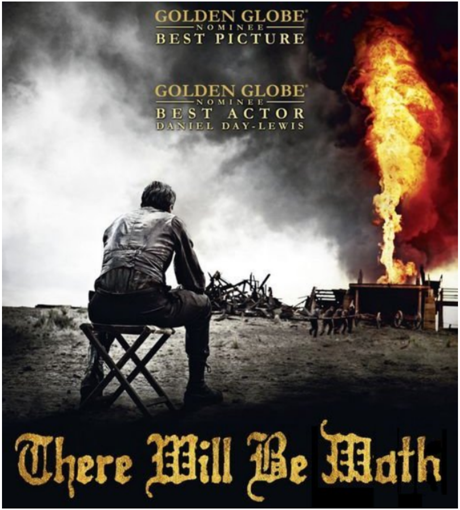
Sources of knowledge:

1. Feature representation
   - Not typically a focus of machine learning
   - Typically seen as "problem specific"
   - *However*, it's hard to learn on bad representations
2. Training data: labeled examples
   - Often expensive to label lots of data
   - Sometimes data is available for "free"
3. Model
   - No single learning algorithm is always good ("no free lunch")
   - Different learning algorithms work with different ways of representing the learned classifier
   - When the data has nothing to say, which model is better
   - Typically requires some control over generalization

More on generalization later. . .

Why do we care about math?!

- **Calculus and linear algebra:**
- **Probability:**
- **Statistics:**

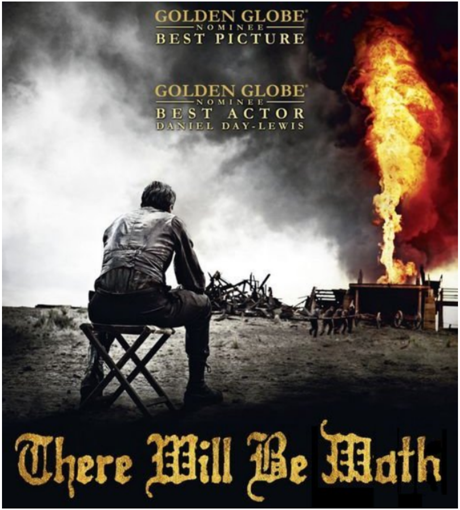


Why do we care about math?!

- **Calculus and linear algebra:**
  1. Techniques for finding maxima/minima of functions

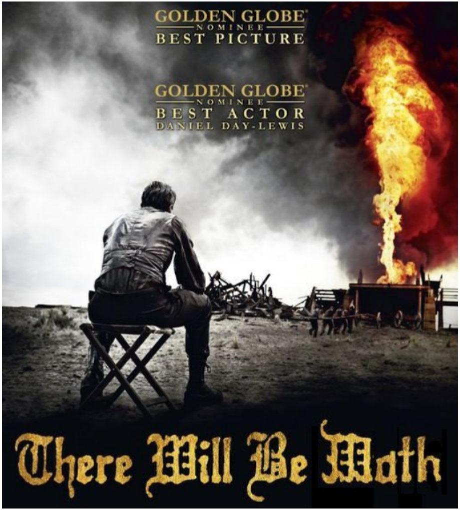
- **Probability:**

- **Statistics:**



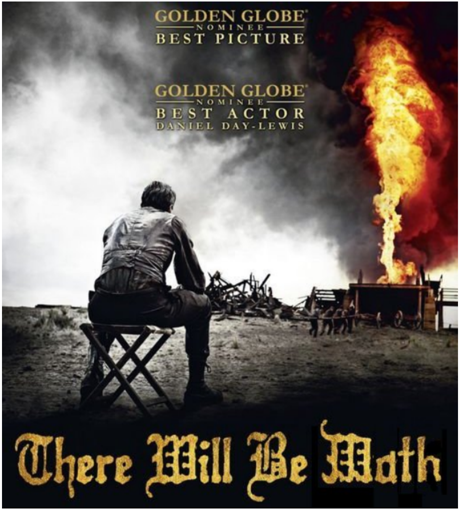
Why do we care about math?!

- **Calculus and linear algebra:**
  1. Techniques for finding maxima/minima of functions
  2. Convenient language for high dimensional data analysis
- **Probability:**
- **Statistics:**



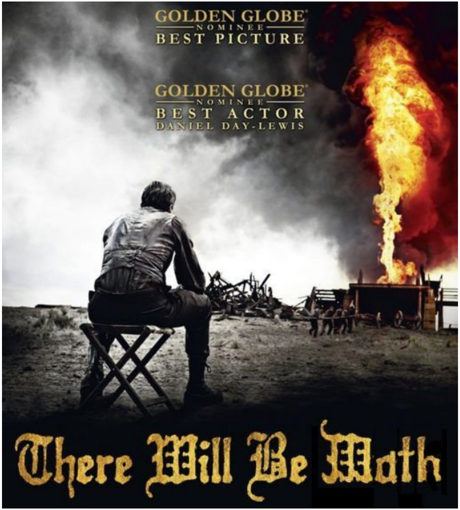
Why do we care about math?!

- **Calculus and linear algebra:**
  1. Techniques for finding maxima/minima of functions
  2. Convenient language for high dimensional data analysis
- **Probability:**
  1. The study of the outcome of repeated experiments
- **Statistics:**



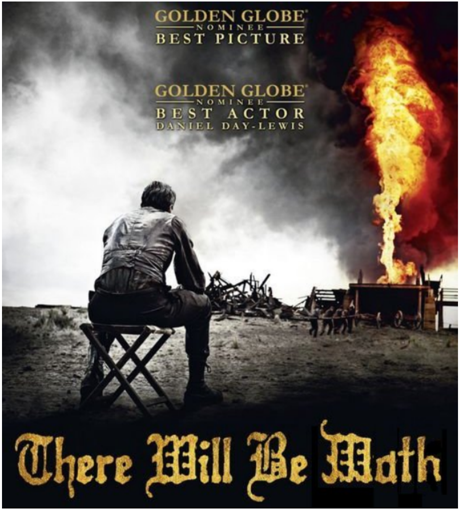
Why do we care about math?!

- **Calculus and linear algebra:**
  1. Techniques for finding maxima/minima of functions
  2. Convenient language for high dimensional data analysis
- **Probability:**
  1. The study of the outcome of repeated experiments
  2. The study of the plausibility of some event
- **Statistics:**



Why do we care about math?!

- **Calculus and linear algebra:**
  1. Techniques for finding maxima/minima of functions
  2. Convenient language for high dimensional data analysis
- **Probability:**
  1. The study of the outcome of repeated experiments
  2. The study of the plausibility of some event
- **Statistics:**
  1. The analysis and interpretation of data



Why do we care about probability and statistics?

Recall, statistics is the analysis and interpretation of data.

In machine learning, we attempt to generalize from one “training” data set to general “rules” that can be applied to “test” data.

Why do we care about probability and statistics?

Recall, statistics is the analysis and interpretation of data.

In machine learning, we attempt to generalize from one “training” data set to general “rules” that can be applied to “test” data.

How is machine learning *different* from statistics?

Why do we care about probability and statistics?

Recall, statistics is the analysis and interpretation of data.

In machine learning, we attempt to generalize from one “training” data set to general “rules” that can be applied to “test” data.

How is machine learning *different* from statistics?

1. Stats cares about the model, we care about predictions

Why do we care about probability and statistics?

Recall, statistics is the analysis and interpretation of data.

In machine learning, we attempt to generalize from one “training” data set to general “rules” that can be applied to “test” data.

How is machine learning *different* from statistics?

1. Stats cares about the model, we care about predictions
2. Stats cares about model fit, we care about generalization

Why do we care about probability and statistics?

Recall, statistics is the analysis and interpretation of data.

In machine learning, we attempt to generalize from one “training” data set to general “rules” that can be applied to “test” data.

How is machine learning *different* from statistics?

1. Stats cares about the model, we care about predictions
2. Stats cares about model fit, we care about generalization
3. Stats tries to explain the world, we try to predict the future

Why do we care about probability and statistics?

Recall, statistics is the analysis and interpretation of data.

In machine learning, we attempt to generalize from one "training" data set to general "rules" that can be applied to "test" data.

How is machine learning *different* from statistics?

1. Stats cares about the model, we care about predictions
2. Stats cares about model fit, we care about generalization
3. Stats tries to explain the world, we try to predict the future

It all started with a lady drinking tea. . .

History of ML?

- Initial attempts at object recognition [Rosenblatt, 1958]
- Learning to play checker [Samuel, 1959, 1963]
- Rosenblatt can't learn XOR [Minsky & Pappert, 1969]
- Symbolic learning, spectroscopy [Winston, 1975; Buchanan 1971]
- Backpropagation for neural nets [Werbos, 1974; Rummelhart, 1986]
- PAC model of learning theory [Valiant, 1984]
- Optimization enters machine learning [Bennett & Mangasarian, 1993]
- Kernel methods for non-linearity [Cortes & Vapnik, 1995]
- Machine learning behind day-to-day tasks [2005ish]
- Machine learning takes over the world [2010ish]
- Neural networks and the biggest come-back ever [2013ish]