Machine Learning (CS4780)

Kilian Q. Weinberger and Chris de Sa
Computer Science and Engineering

Cornell University 2018
Part I

Administrative Stuff

This baby says: BORING!
Course Information

• Instructors:
  – Chris de Sa and Kilian Q. Weinberger

• Homepage:
  – https://courses.cis.cornell.edu/cs4780
  – Vocareum: http://vocareum.com
  – Piazza: https://piazza.com/class/jcb1ar25kjd5vq

• TAs:
  – Many (more and better than you think)

• Office Hours / Recitations:
  • TA Office Hours: Every day. (Details will be posted on Piazza.)
  • Prof. Office Hours: Tu/Thu 10am-11am

• Questions:
  • Post all questions on Piazza (you can make them private)
  • Do not email directly (except in an emergency)

• (Optional) Course Books
  • Machine Learning a Probabilistic Perspective (K.P. Murphy)
  • The Elements of Statistical Learning (Hastie, Tibshirani, Friedman)
Course Breakdown

- **50% Theory: Midterm + Final (1/4 each)**
  - Closed book
  - No cheat sheets!
  - No personal notes

- **50% Programming Assignments**
  - Up to 2 members in each team
  - 2 days extension *per team per project*
  - Autograder (unlimited resubmissions)
  - *Extra credit if you beat my own submission*
  - *Extra credit if you win contests*

- **(0% or 100%) Homeworks**
  - Up to 5 members in each team
  - Binary grades \{0,100\}
  - Preparation for exam
  - Must submit a *faithful attempt*
Placement Exam

- Due Monday!!!
- Label with your NETID on front!!!
- Get started early!!
- It is there for your own protection!
- This is how you get a Vocareum invite!
Study Groups (2-5 people)

- You **must** join a study group by next week.
- Find people on Piazza
- This course will cover non-trivial material, learning in a group makes it less hard and more fun!
MLunch

- Lunch with Kilian or Chris every Monday
- Up to 8 students (online sign up sheet)
- (will post link on Piazza)
- Details will follow ...
Pass placement exam -> get account
There will be (roughly) 8 projects
You have (roughly) 2 weeks for each project
Unlimited submits until deadline
Costs $30 :-(

http://vocareum.com
BIG VOTE!!

JULIA

PYTHON
Important notes:

- **language.txt must be filled with one word** (Only in the unlikely case of a tie):
  - julia
  - python
- Setup a (secret) **screen name** for your team for the leaderboard
  - (top right corner - click on your login)
- Only text with `<GRADED>` and `</GRADED>` will be graded
- !!!!You MUST form teams before you get started!!!!!
# Course Topics

<table>
<thead>
<tr>
<th>We will cover:</th>
<th>We will not cover:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Parametric / Non-parametric learning</td>
<td>- Graphical Models</td>
</tr>
<tr>
<td>- Empirical Risk Minimization</td>
<td>- Reinforcement Learning</td>
</tr>
<tr>
<td>- Bias/Variance Trade-off</td>
<td>- Unsupervised Learning</td>
</tr>
<tr>
<td>- Boosting</td>
<td>- Genetic Programming</td>
</tr>
<tr>
<td>- Support Vector Machines</td>
<td></td>
</tr>
<tr>
<td>- Deep Learning</td>
<td></td>
</tr>
<tr>
<td>- (Recommender systems)</td>
<td></td>
</tr>
</tbody>
</table>
Prerequisites
Prerequisites

- Three pillars of ML:
  - Statistics / Probability
  - Linear Algebra
  - Multivariate Calculus

- Should be confident in at least 1/3, ideally 2/3.

- TAs might be able to give recitations on some topics if needed. (But don’t rely on it.)
About this course

* Take this course if …
  * you are interested in Machine Learning
  * you are comfortable with a decent amount of mathematics
  * you are not scared of programming

* Don’t take this course if …
  * matrices scare you
  * you don’t remember how to take derivatives
  * you want an easy A

* You cannot take this course if you fail the placement exam.
  * In that case, take appropriate prerequisites and come back next year.
Student comments

- “[…] Requires a good knowledge in math and derivatives.”
- “A TON of work, but mostly worth it for a very valuable skill.”
- “great course, but prepare to work your butt off.”
- “The topics were pretty complicated and difficult to understand quickly. I would have preferred a slightly slower pace.”
- “It's mostly a math class”
Academic Integrity

- Zero Tolerance policy
  - All occurrences will be reported
- We **actively** look for academic conduct violations
- The autograder checks for plagiarism **daily**
- Past outcomes
  1. Student cries in my office
  2. Student may ruin transcript or (much) worse
  3. (If on visa, student can lose permission to stay in U.S.)
Examples:

- Most common: Students steal from same source
- Students post to RentACoder.com or other page
- Students post solutions on the web
- Students use solutions from last year’s course
Part II: What is MACHINE LEARNING?
Traditional Computer Science

Traditional CS:

Data → Program → Computer → Output
Machine Learning

Traditional CS:

Data → Program → Computer → Output

Machine Learning:

Data → Output → Computer → Program
Machine Learning

Machine Learning:
- Data → Computer → Program → Output

Traditional CS:
- Data → Computer → Output
Machine Learning

Training:

Data → Computer

Program → Data → Computer

Testing:

Computer → Output

Output
What is Machine Learning?

* **Formally:** (Mitchell 1997): A computer program $A$ is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$.

* **Informally:** *Algorithms* that *improve* on some *task* with *experience*. 
A (very brief) History of ML
Samuel’s Checker Player (1952)

- Basically Shannon’s Minimax Algorithm (traditional AI)
- Included simple learning algorithm to improve board evaluation.
- Player improved over time!!
Perceptron 1957
(Frank Rosenblatt @ Cornell)

- Provable convergence properties
- Eventually leads to multilayer perceptron = Artificial Neural Networks = Deep Learning
AI Winter

- (1969) Minsky & Papert “killed” AI
- Burst huge expectation bubble
- Funding for AI research collapsed for decades
Rebirth as Machine Learning

- Machine Learning:
  - Originally: Mostly a name game to get funding.
  - Profound difference:
    - ML: Bottom up, AI: Top down
    - ML: More practical smaller goals
    - Based on Statistics and Optimization, not Logic
TD-Gammon (1994)

* Gerry Tesauro (IBM) teaches a neural network to play Backgammon. The net plays 100K+ games *against itself* and beats world champion [Neurocomputation 1994]

* Algorithm teaches *itself* how to play so well!!!
Deep Blue (1997)

- IBM’s Deep Blue wins against Kasparov in chess. Crucial winning move is made due to Machine Learning (G. Tesauro).
Example: Websearch
Example: Spam Filter

From: Leap 2 Upgrades <upgrade.noreply@kagi.com>
Subject: [Bulk] Leap 2 Upgrades: Kilian Weinberger, save over 50% when you upgrade to Leap 2
Date: September 21, 2010 6:33:55 PM CDT
To: Kilian Weinberger <kilianw@seas.upenn.edu>

Limited Time Special Upgrade Offer

Upgrade to Yep 2 or Leap 2 for only $19, or... buy the 4-product bundle for only $29.
Home assistant
Very soon: Autonomous Cars
When will it stop?

• The human brain is one big learning machine

• We know that we can still do a lot better!

• However, it is hard. Very few people can design new ML algorithms.

• But many people can use them!
What types of ML are there?

As far as this course is concerned:

* **supervised learning:** Given labeled examples, find the right prediction of an unlabeled example. (e.g. *Given annotated images learn to detect faces.*)

* **unsupervised learning:** Given data try to discover similar patterns, structure, sub-spaces (e.g. *automatically cluster news articles by topic*)

* **reinforcement learning:** Try to learn from delayed feedback (e.g. *robot learns to walk, fly, play chess*)
## Average salary by skill

### Computer centric skill value

<table>
<thead>
<tr>
<th>Skill</th>
<th>Average salary value of skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process Management</td>
<td>$170,371</td>
</tr>
<tr>
<td>Automation Tools</td>
<td>150,853</td>
</tr>
<tr>
<td>Load Runner</td>
<td>140,656</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>125,258</td>
</tr>
<tr>
<td>Big Data</td>
<td>119,938</td>
</tr>
<tr>
<td>NoSQL</td>
<td>119,471</td>
</tr>
<tr>
<td>Software as a Service (SaaS)</td>
<td>118,616</td>
</tr>
<tr>
<td>MongoDB</td>
<td>114,503</td>
</tr>
<tr>
<td>Systems Development Life Cycle (SDLC)</td>
<td>110,699</td>
</tr>
<tr>
<td>Concept Development</td>
<td>110,549</td>
</tr>
<tr>
<td>JBoss</td>
<td>109,661</td>
</tr>
<tr>
<td>IT Management</td>
<td>109,077</td>
</tr>
<tr>
<td>Data Modeling</td>
<td>108,970</td>
</tr>
<tr>
<td>Objective C</td>
<td>108,225</td>
</tr>
<tr>
<td>Data Architecture</td>
<td>107,875</td>
</tr>
<tr>
<td>Struts</td>
<td>107,443</td>
</tr>
<tr>
<td>Business Solutions</td>
<td>106,756</td>
</tr>
<tr>
<td>Apache Hadoop</td>
<td>105,986</td>
</tr>
<tr>
<td>JDBC</td>
<td>104,070</td>
</tr>
<tr>
<td>Java Message Service (JMS)</td>
<td>103,576</td>
</tr>
<tr>
<td>Scrum</td>
<td>103,461</td>
</tr>
</tbody>
</table>
## Average salary by skill

<table>
<thead>
<tr>
<th>IT Management</th>
<th>109,077</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Modeling</td>
<td>108,970</td>
</tr>
<tr>
<td>Objective C</td>
<td>108,225</td>
</tr>
<tr>
<td>Data Architecture</td>
<td>107,875</td>
</tr>
<tr>
<td>Struts</td>
<td>107,443</td>
</tr>
<tr>
<td>Business Solutions</td>
<td>106,756</td>
</tr>
<tr>
<td>Apache Hadoop</td>
<td>105,986</td>
</tr>
<tr>
<td>JDBC</td>
<td>104,070</td>
</tr>
<tr>
<td>Java Message Service (JMS)</td>
<td>103,576</td>
</tr>
<tr>
<td>Scrum</td>
<td>103,461</td>
</tr>
<tr>
<td>UNIX Administration</td>
<td>103,124</td>
</tr>
<tr>
<td>ERwin</td>
<td>102,499</td>
</tr>
<tr>
<td>Unified Modeling Language (UML)</td>
<td>102,498</td>
</tr>
<tr>
<td>Sybase</td>
<td>102,117</td>
</tr>
<tr>
<td>Integrated Development Environment (IDE)</td>
<td>102,085</td>
</tr>
<tr>
<td>Full Life Cycle</td>
<td>101,460</td>
</tr>
<tr>
<td>Verilog</td>
<td>101,201</td>
</tr>
<tr>
<td>Python</td>
<td>100,717</td>
</tr>
<tr>
<td>Eclipse</td>
<td>100,648</td>
</tr>
</tbody>
</table>

Data: Brookings/Burning Glass
A few quotes ... 

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, form. Director, DARPA)
- “Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Director Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)
- “Machine learning is today’s discontinuity” (Jerry Yang, CEO, Yahoo)
- “AI vastly more risky than North Korea” (Elon Musk, CEO Tesla, Spaxe X, …)
Data Privacy

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

[Wang and Kosinski 2017]
Weaponry

“Slaughterbots” https://www.youtube.com/watch?v=9CO6M2HsoIA
Impactful Decision Making

Compas Recidivism Algorithm
Now let’s get crack’n
To-do action items:

- Join Piazza!
- Finish Placement Exam!
- Find study group
- Autograder Setup:
  - Find Teammate
- Start Project 0