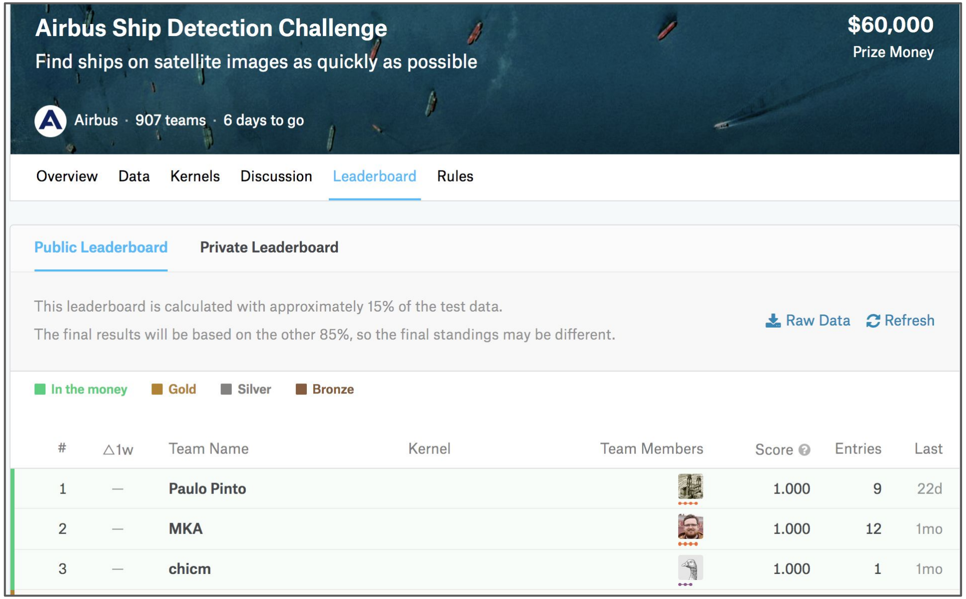
Week 6

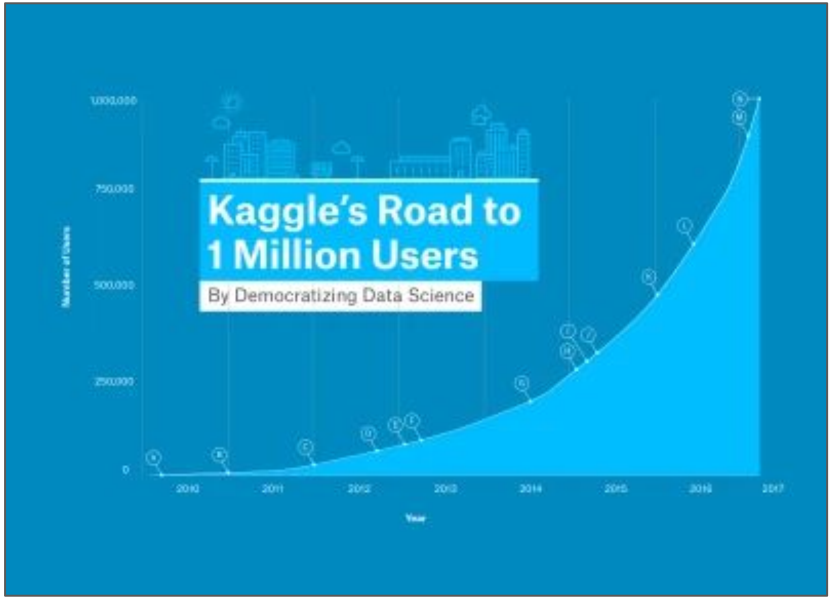
Agenda

1. Introduce Final Project
2. Gradient Descent and Regularization review
3. Introduction to neural network libraries
4. Start deep learning notebook

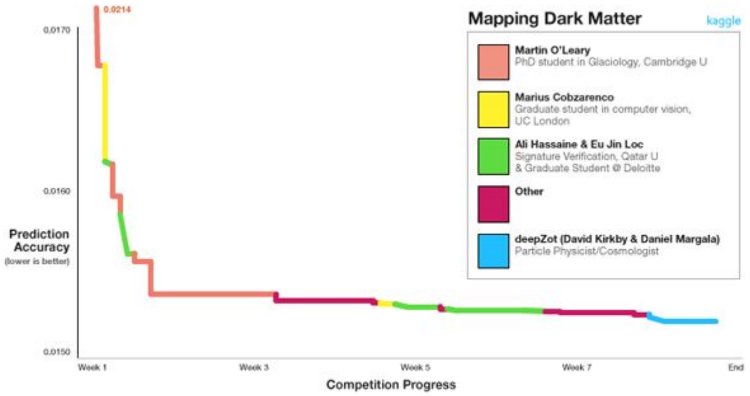
For next week: Read LeCun paper

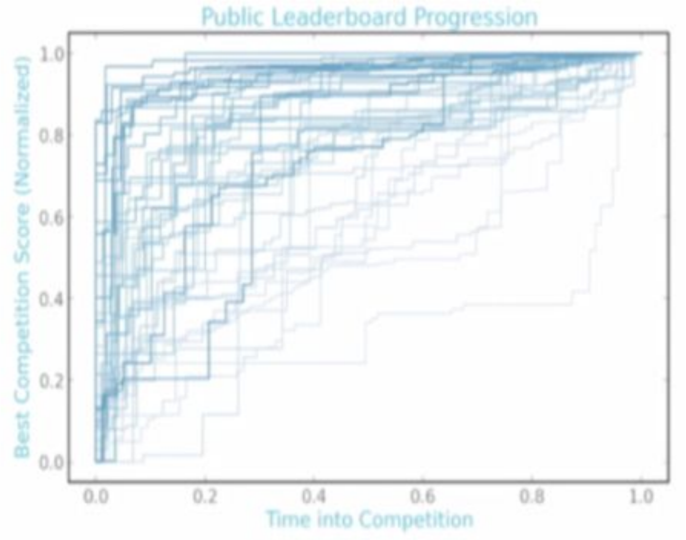
Kaggle

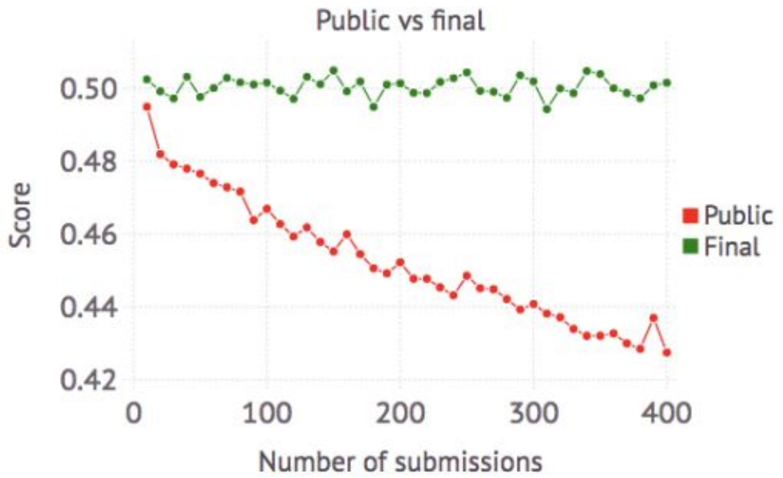

Airbus Ship Detection Challenge
Find ships on satellite images as quickly as possible
$60,000
Prize Money
Airbus · 907 teams · 6 days to go
Overview Data Kernels Discussion Leaderboard Rules
Public Leaderboard Private Leaderboard
This leaderboard is calculated with approximately 15% of the test data.
The final results will be based on the other 85%, so the final standings may be different.
Raw Data Refresh
In the money Gold Silver Bronze
△1w Team Name Kernel Team Members Score Entries Last
1 — Paulo Pinto 1.000 9 22d
2 — MKA 1.000 12 1mo
3 — chicm 1.000 1 1mo


Kaggle's Road to 1 Million Users
By Democratizing Data Science

Kaggle


Mapping Dark Matter
kaggle
0.0214
0.0170
0.0160
0.0150
Prediction Accuracy
(lower is better)
Martin O'Leary
PhD student in Glaciology, Cambridge U
Marius Cobzarenco
Graduate student in computer vision, UC London
Ali Hassaine & Eu Jin Loc
Signature Verification, Qatar U & Graduate Student @ Deloitte
Other
deepZot (David Kirkby & Daniel Margala)
Particle Physicist/Cosmologist
Week 1
Week 3
Week 5
Week 7
End
Competition Progress


Public Leaderboard Progression
Best Competition Score (Normalized)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Time into Competition


Public vs final
Score
0.50
0.48
0.46
0.44
0.42
0
100
200
300
400
Number of submissions
Public
Final

Final Project

FINAL PROJECTS

- XX Baseline Due
- XX Final Notebook Due
- XX Presentations
- Groups of 2-4
- You pick your groups.. Feel free to use this signup sheet to help find group members:
- https://docs.google.com/document/d/16supxu4kE1hTHrFeDzj1qvNKw5HE2WhmnDL6OeLGbqY/edit?usp=sharing

RANDOM ACTS OF PIZZA

- https://www.kaggle.com/c/random-acts-of-pizza
- People post pizza requests on Reddit
- Build 2-class classifier
- Classify whether post will get pizza
- Practice mining features from text

HOME PRICE PREDICTION

- https://www.kaggle.com/c/house-prices-advanced-regression-techniqu
- Predict the sale price of a property (like the Zestimate)
- Regression
- Feature engineering

FOREST COVER PREDICTION

- https://www.kaggle.com/c/forest-cover-type-prediction
- Classify canopy type in forest (e.g. Spruce)
- Multi-class classification: 8 classes
- Practice trying different algorithms

FACIAL KEYPOINTS DETECTION

- https://www.kaggle.com/c/facial-keypoints-detection
- Determine x,y of keypoints in image (e.g. left eye corner)
- 30 regression outputs (x,y of 15 labels)
- Practice Convolutional Neural Networks

Gradient Descent Review

Warm-up

Student Exercise

- Find:
  - Parameters
  - Cost function
  - Objective
- Model "hypothesis"

$$Y_i = \alpha + \beta X_i + y X_i^2$$

- Cost function twist:
  - Use "absolute error" cost function.

Gradient Descent

- Pseudocode:
  - Choose an initial vector of parameters α, β.
  - Choose learning rate R.
  - Repeat until an approximate minimum is obtained (randomly shuffl examples in training set).
  - For each example i:

$$\alpha <- \alpha - R \frac{\partial}{\partial \alpha} J(\alpha, \beta)$$

$$\beta <- \beta - R \frac{\partial}{\partial \beta} J(\alpha, \beta)$$

GD Performance

- What is the batch gradient descent algorithm?
- What is the stochastic gradient descent algorithm (SGD)?
- What is mini-batch?

Additional Q's

- What does convergence mean?
- What is the benefit of having a convex cost function?
- Why might feature scaling be important?
- What is alpha? How might we set alpha?

Sigmoid Function

- Logistic (sigmoid) function: $g(z) = \frac{e^z}{e^z+1} = \frac{1}{1+e^{-z}}$
- In logistic regression: $z = \alpha + \beta X + \ldots$
- Transforms: $[-\infty, +\infty] \rightarrow [0, 1]$
- Constrains output of our model between 0 and 1

Batch, Mini-Batch, or Stochastic?

- Choose an initial vector of parameters w and learning rate η.
- Repeat until an approximate minimum is obtained:
  - Randomly shuffle examples in the training set.
  - For $i = 1, 2, ..., n$, do:
    - $w := w - \eta \nabla Q_i(w)$.

Gradient Descent

Examining the Cost Function

- Logistic regression cost function:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \log \hat{Y}_i + (1 - Y_i)\log(1 - \hat{Y}_i)$$

- Can rewrite single part as two different components:

$$\mathrm{Cost}(\hat{Y}_i, Y_i) = \begin{cases} -\log(\hat{Y}_i) & \text{if } Y_i = 1 \\ -\log(1 - \hat{Y}_i) & \text{if } Y_i = 0 \end{cases}$$

Logistic Regression: Gradient Descent

- Benefit: leads to getting predicted cost values closer to actual values
  - Cost function:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \log \hat{Y}_i + (1 - Y_i)\log(1 - \hat{Y}_i)$$

- Use the update rule:

$$\theta <- \theta - R\frac{\partial}{\partial\theta}J(\theta)$$

- Benefit: derivative is very simple:

$$\frac{\partial}{\partial\theta}J(\theta) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)X_i$$

1. Why do we use a surrogate cost function?

Gradient Descent

Gradient Descent: Learning Rate

- $\beta \leftarrow \beta - R\frac{\partial\beta}{\beta}J\ (\alpha,\beta)$
- What does R do?
- Small R:
  - Slow and incremental gradient descent
- Large R:
  - Parameter values update quickly.
  - Could overshoot the minimum.
  - May fail to converge.

1. Can you imagine what an adaptive learning rate might look like?

Gradient Descent

Feature Scaling With Gradient Descent

- Critical in gradient descent, where we seek the direction of greatest decrease in cost function.
- Can be viewed as contour plot.
- With gradient descent, the direction of steepest descent can depend on the units in which variables are measured.
  - Inefficient when features or axes are on different scales

Local or Global Minima?

- Gradient descent is guaranteed to converge to a local minimum as long as R is small enough and the function is differentiable.
- How do we know we've hit a global versus local minimum?
- In regression it doesn't matter!

Regularization

Working With the Penalized Cost Function

- Penalized cost function:

Penalty

$$J(\alpha, \beta) = \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \theta_0 + \theta_1 X_i + \dots + \theta_k X_i^k)^2 + \boxed{\boxed{\lambda} \sum_{j=1}^{k} \theta_j^2}$$

Regularization parameter

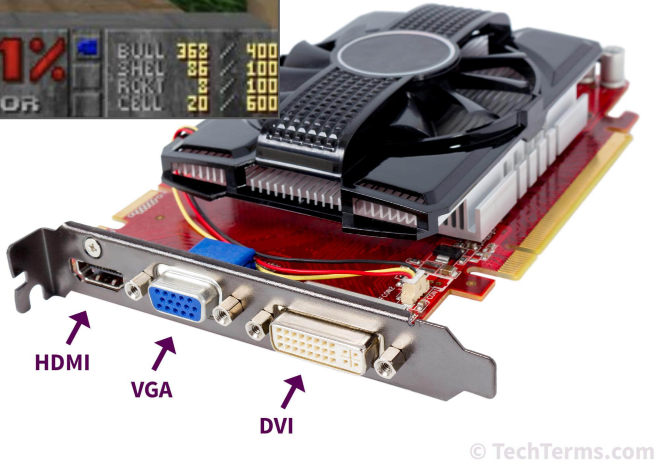
- The larger the θ parameter is, the higher the cost will be.

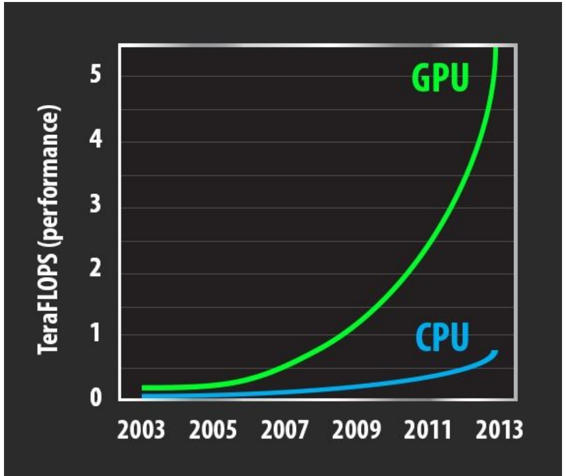
1. Is that L1 or L2 regularization?
2. What does L1 regularization accomplish?
3. What does L2 regularization accomplish?

Introduction to Neural Network Libraries

GPUs

1. GPUs grew out of gaming in the 90s
2. GPUs do vector/matrix/tensor operations faster than CPUs.





datascience@berkeley

Neural Network Frameworks

Support for deep learning frameworks

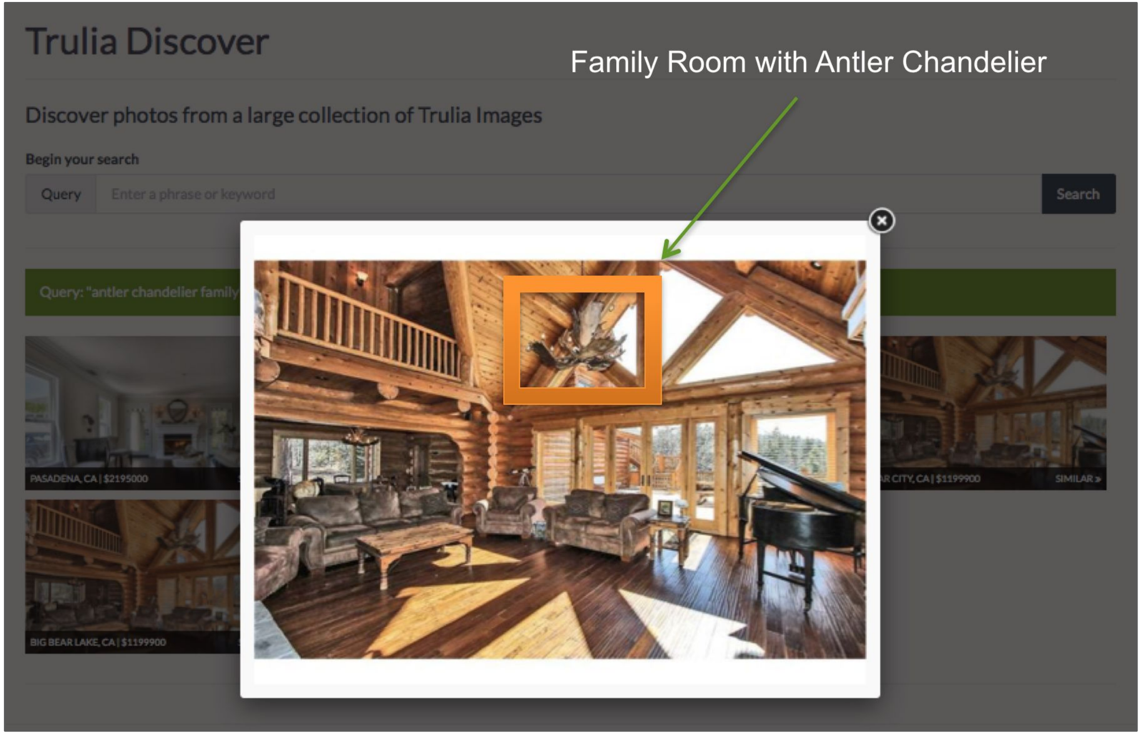
The AWS Deep Learning AMIs support all the popular deep learning frameworks allowing you to define models and then train them at scale. Built for Amazon Linux and Ubuntu, the AMIs come pre-configured with Apache MXNet and Gluon, TensorFlow, Microsoft Cognitive Toolkit, Caffe, Caffe2, Theano, Torch, PyTorch, and Keras, enabling you to quickly deploy and run any of these frameworks at scale.

Accelerate your model training

To expedite your development and model training, the AWS Deep Learning AMIs include the latest NVIDIA GPU-acceleration through pre-configured CUDA and cuDNN drivers, as well as the Intel Math Kernel Library (MKL), in addition to installing popular Python packages and the Anaconda Platform.

1. Large number of NN libraries have emerged in recent years
2. All support GPUs

Neural Network Frameworks



1. Exceptional at working with large scale text and image datasets

Final Thoughts?