Hacking Machine Learning Algorithms

Kyle Polich

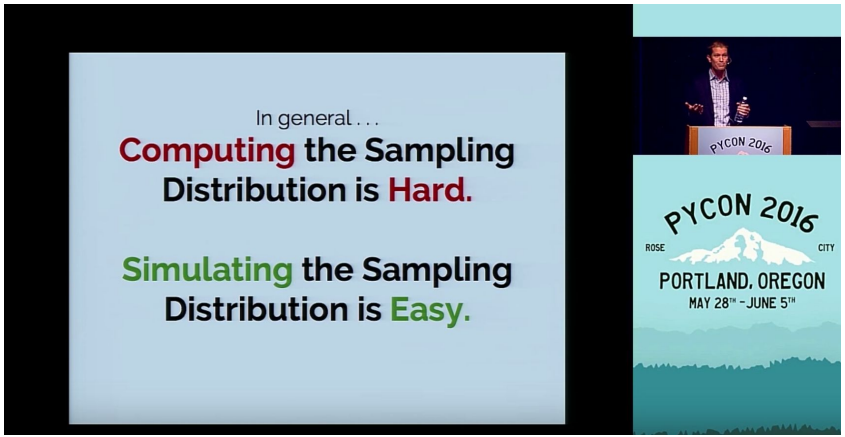
Outline

- What is machine learning?
- Why would it be interesting to hack?
- What do I mean by “hacking” ML?

What do others mean by “hack”?

...working in
areas related
to machine
learning...

What do others mean by “hack”?

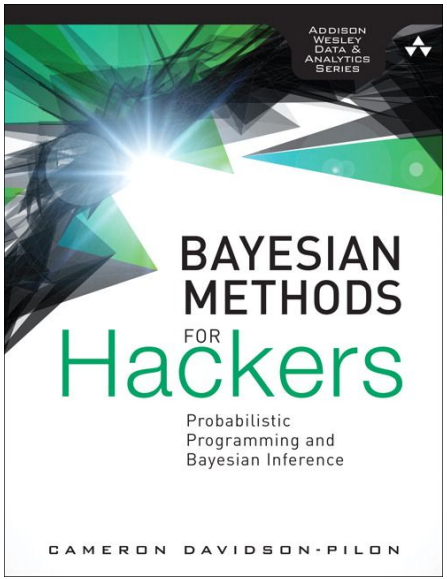


“A person whose natural approach to problem-solving involves writing code.”

http://bit.ly/29PmJHY

- Jake VanderPlas

What do others mean by “hack”?

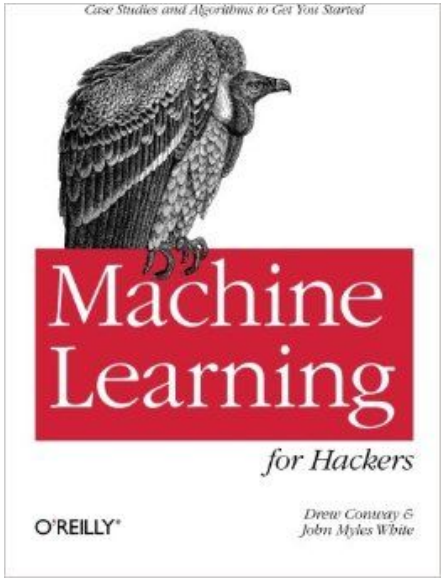


http://bit.ly/1ON3qld

“... documentation is lacking in certain areas, especially those that bridge the gap between beginner and hacker. One of this book's main goals is to solve that problem.”

- Cameron Davidson-Pilon

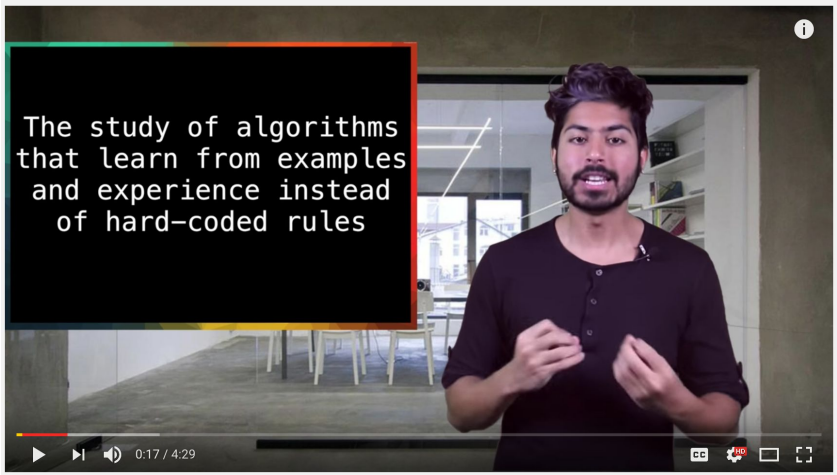
What do others mean by “hack”?



“We believe a hacker is someone who likes to solve problems and experiment with new technologies”

- Drew Conway, John White

What do others mean by “hack”?



“This video will get you up and running with your first ML app in just 7 lines of Python”

- Siraj Raval

http://bit.ly/29VBSva

What does Kyle Polich mean by “hack”?

- Information should (typically) be free
- Promoting decentralization
- Sharing
- Openness
- World Improvement
- **Figuring out how exactly things work, possibly to exploit them**

What is Machine Learning?

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.

- Classification / Labeling
- Clustering
- Decision problems

1 Minute ML Crash Course

Start with the dataset and objective variable

Using dictionary password?	Age	Email domain	Education	OS	Admin user
No	18	@gmail.com	GED	Osx	No
No	35	@2600.com	B.S.	FreeBSD	Yes
Yes	25	@spacex.com	PhD	Ubuntu	No
No	24	@yahoo.com	M.S.	TAILS	No
Yes	51	@aol.com	B.S.	Windows	Yes
Yes	45	@polich.com	B.S.	Osx	No
No	33	@gmail.com	GED	Red hat	No
No	51	@rr.com	Ms.	Osx	No

1 Minute ML Crash Course

Clean and transform the data

Using dictionary password?	Age	Email domain	Education	OS	Admin user
No	245	null	null	Android	No
Yes	-1	@3fk93f20f.org	PhD	Android 1.2.343.343.6.33.3	Yes

1 Minute ML Crash Course

Remove outliers and erroneous data

Using dictionary password?	Age	Email domain	Education	OS	Admin user
No	65536	HTTP/1.1	200	Osx	No
No	-1	@@@	301	FreeBSD	Yes
Yes	999999	a@b.co	404	Ubuntu	No
No	0	billg@bing.com	500	TAILS	No

1 Minute ML Crash Course

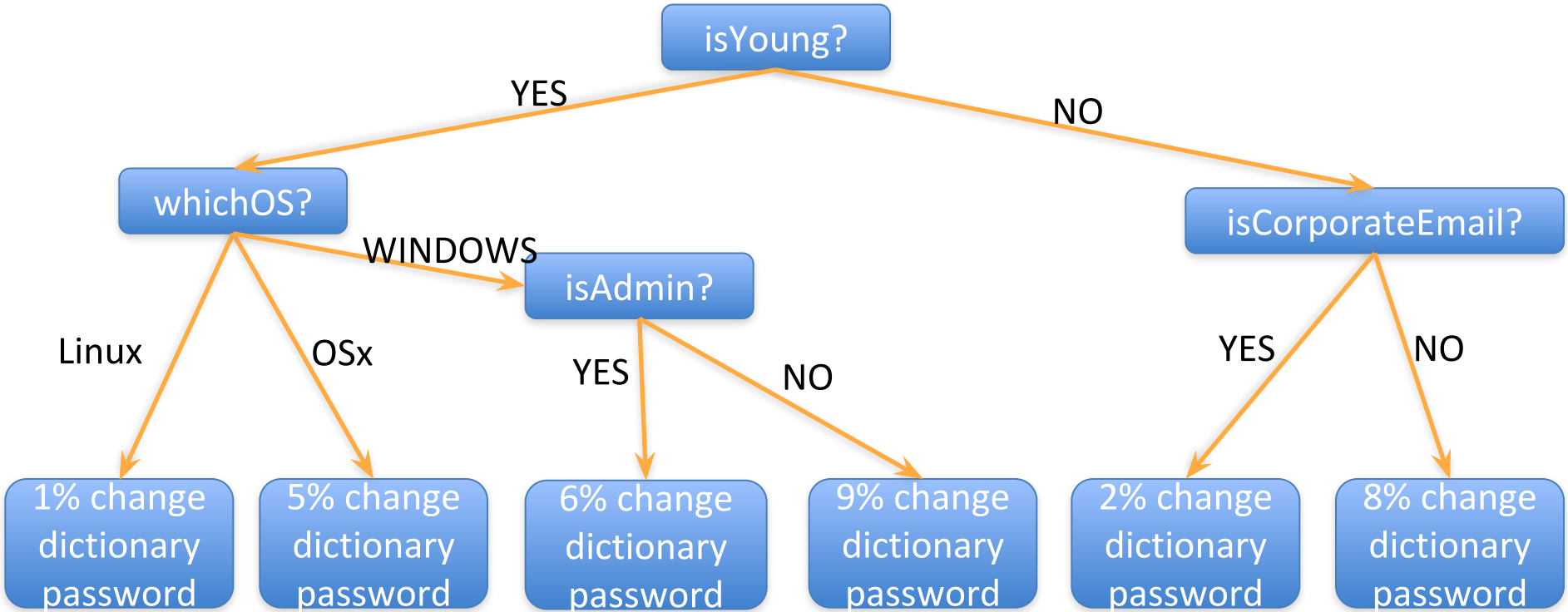
Run an algorithm…

…like logistic regression

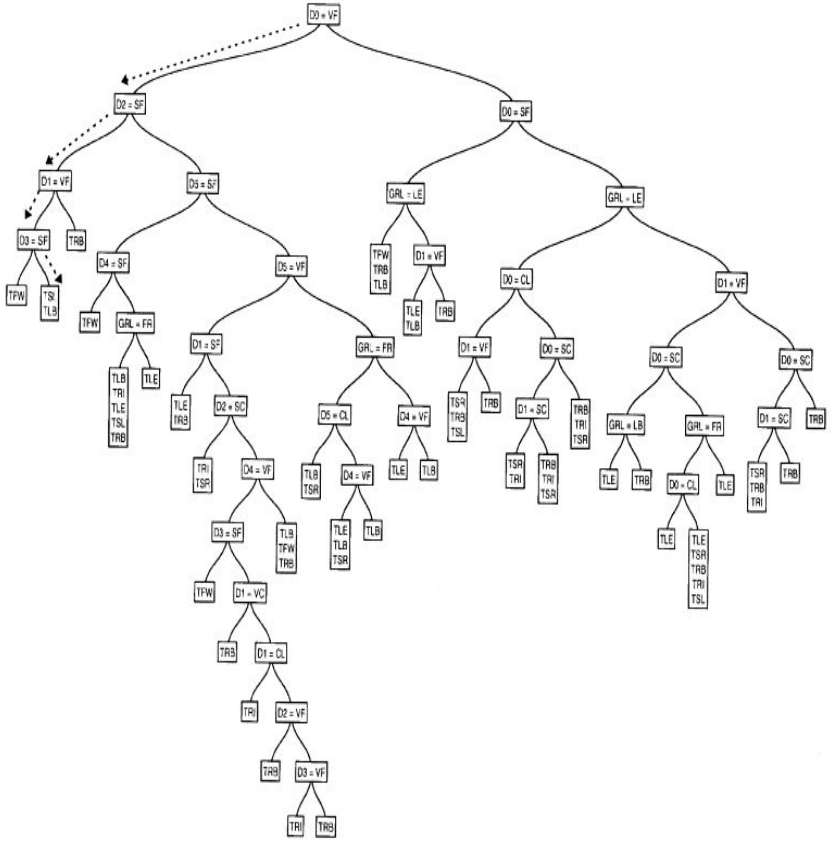
$$Pr(insecure) = \beta_0 + \beta_1 \cdot isYoung + \beta_2 \cdot linuxUser + \beta_3 \cdot isAdmin$$

1 Minute ML Crash Course

…like a decision tree



Model Interpretability????



1 Minute ML Crash Course

Cross validation

Training

Using dictionary password?	Age	Email domain	Education	OS	Admin user
No	18	@gmail.com	GED	Osx	No
No	35	@2600.com	B.S.	FreeBSD	Yes
Yes	25	@spacex.com	PhD	Ubuntu	No
No	24	@yahoo.com	M.S.	TAILS	No
Yes	68	@aol.com	B.S.	Windows	Yes
Yes	45	@polich.com	B.S.	Osx	No
No	33	@gmail.com	GED	Red hat	No
No	51	@rr.com	Ms.	Osx	No

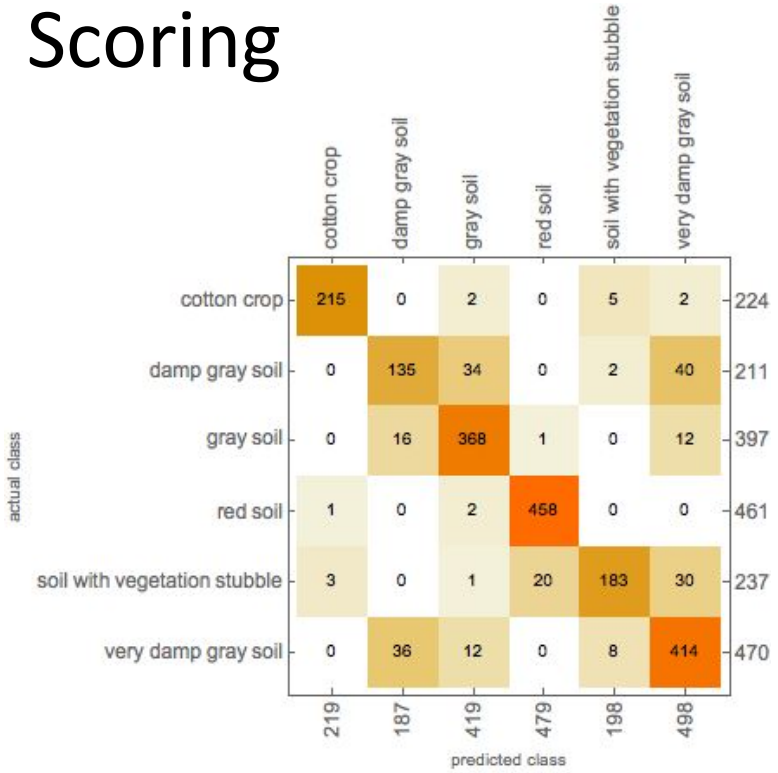
Testing

Using dictionary password?	Age	Email domain	Education	OS	Admin user
No	245	null	null	Android	No
Yes	-1	@3fk93f20f.org	PhD	Android 1.2.343.343.6.33.3	Yes

1 Minute ML Crash Course

Scoring

Actual Class	Predicted Class: Yes	Predicted Class: No
Yes	TP	FN
No	FP	TN

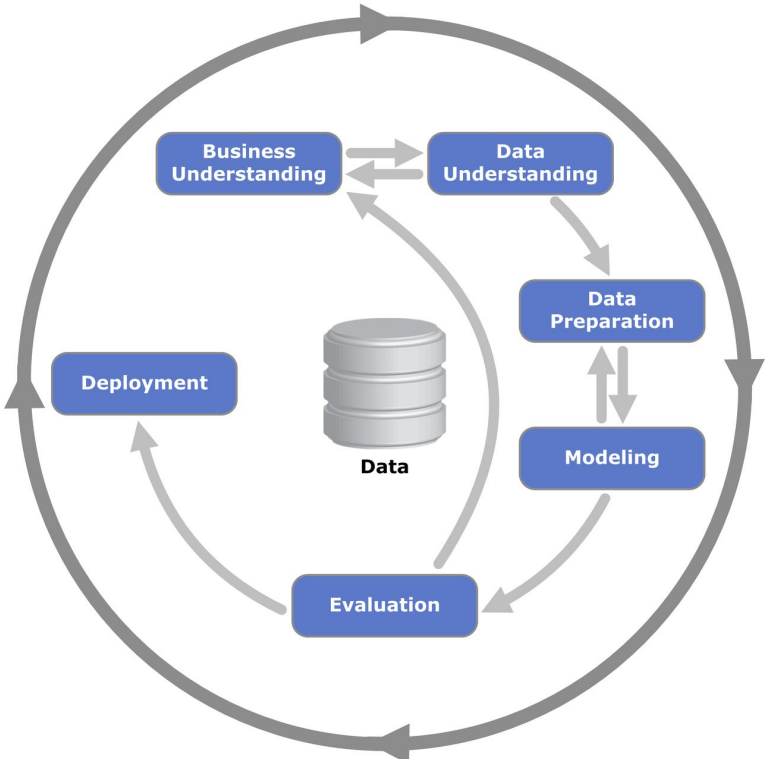


1 Minute ML Crash Course

Make predictions for fun, personal enrichment, improving the world, or profit!

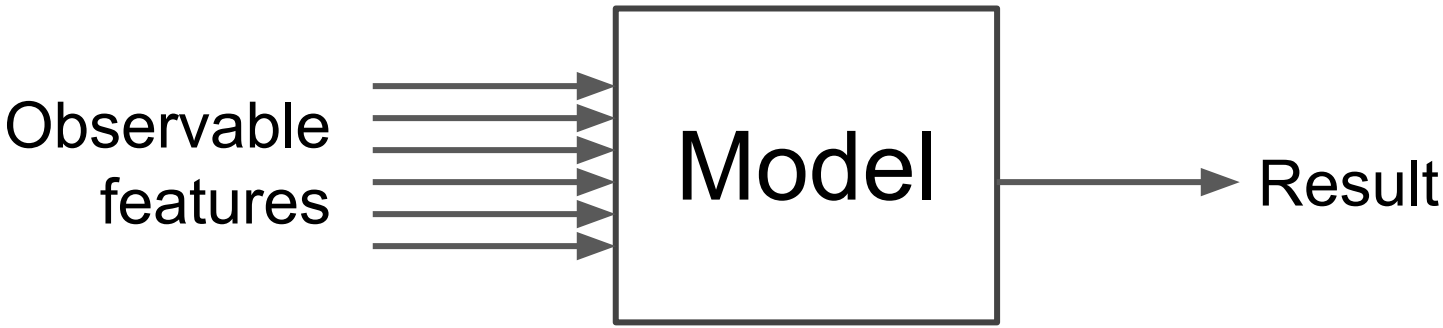
What is the process of ML?

- Feature engineering
- Algo selection
- Validation
- Tuning
- Deployment



What is ML?

A system that takes some input and produces some output by a mechanism that was algorithmically generated.



Why hack ML?

- Learn
- Advance the field
- Personal gain
- Watchdog

Predictive Policing

- HRDAG found DOJ Bureau of Justice Statistics substantially understated the number of homicides by police.
- Idea: optimization of department location to minimize emergency arrival time.

Unintended Racist Algorithm

Company: Real Estate site

Project: Recommend best matches of properties for sale given previous searches

Obvious considerations: Distance to max price, percentage of houses viewed with pools, location by IP address

Unintended Racist Algorithm

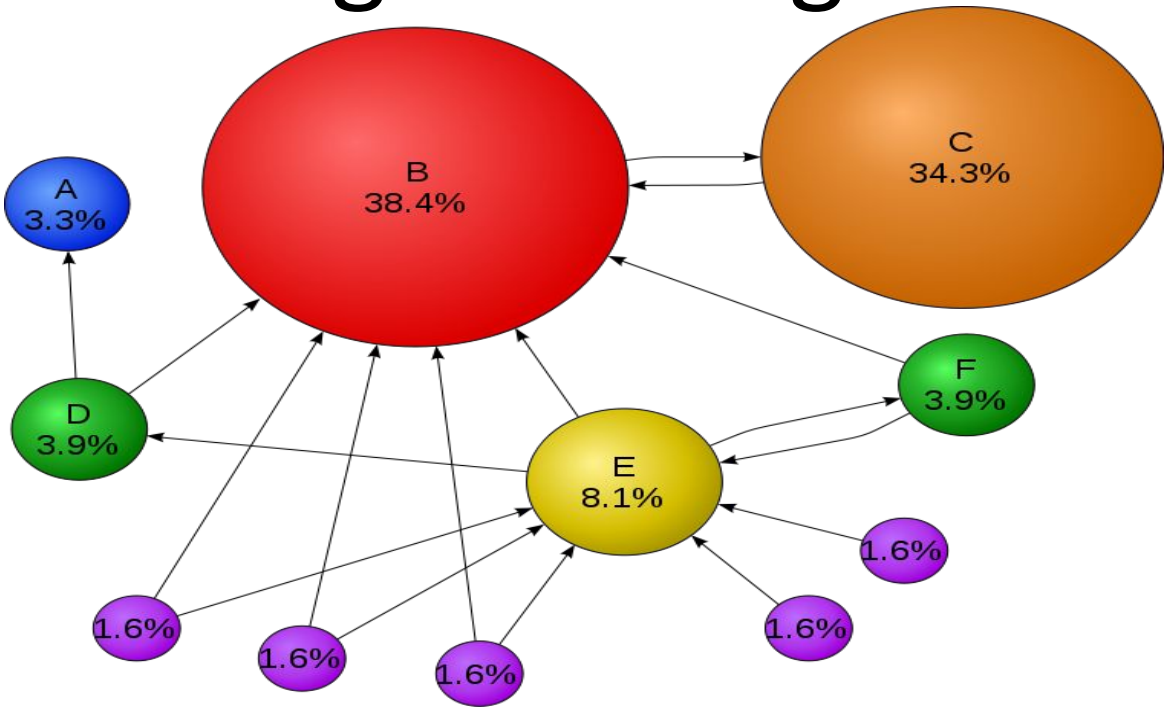
Predict move destination
by current IP address?

Destination ZIP	Current ZIP	# baths (current)	# beds (current)
11221	11365	1	1
11779	10011	1	2
11234	11226	1	1

Tay

“Bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we’ve got.”
- Tay

The PageRank algorithm



Incoming links from authoritative pages convey authority to a web page.

The Making of a Fly

by Peter A. Lawrence

< Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

List Price: $70.00
Used: from $35.54
New: from $1,730,045.91

Have one to sell? Sell yours here

All | New (2 from $1,730,045.91) | Used (15 from $35.54)

Show ⊙ New ○ Prime offers only (0) Sorted by Price + Shipping

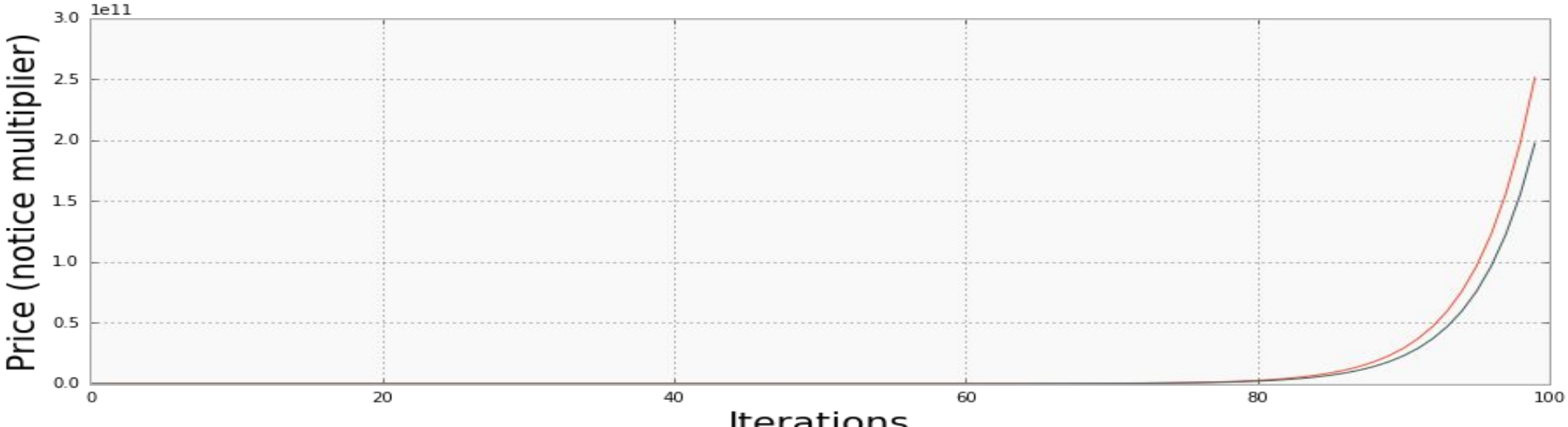
New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$1,730,045.91 + $3.99 shipping	New	Seller: profnath Seller Rating: 93% positive over the past 12 months. (8,193 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$2,198,177.95 + $3.99 shipping	New	Seller: bordeebook Seller Rating: 93% positive over the past 12 months. (125,891 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher	Add to Cart or Sign in to turn on 1-Click ordering.

Runaway pricing

```
def algorithmA(competitor_prices):
    max_price = max(competitor_prices)
    return 1.270589 * max_price

def algorithmB(competitor_prices):
    max_price = max(competitor_prices)
    return 0.9983 * max_price
```



The Hathaway Effect

Claim: Social media news and/or search traffic about Anne Hathaway has affected trading of Berkshire Hathaway stock

- Rumor at premiere of Les Miserables

The Hathaway Effect

Claim: Social media news and/or search traffic about Anne Hathaway has affected trading of Berkshire Hathaway stock

- Rumor at premiere of Les Miserables
- Berkshire Hathaway announced stock buyback

Spam filters

Filtering approach	Hacking approach
Blacklist of words (e.g. viagra, free shipping, breast)	Misspell words, adjust content
IP blacklist	Distributed delivery and botnets
Machine learning	Bayesian poisoning
Optical character recognition (OCCR)	Content hidden in images
Behavioral features (open rates, flagging, replying)	Callback hacking
Social network analysis	?

Where are ML at work?

Spam filters, Advertising, Fraud detection, Assess creditworthiness, Speech recognition, Route police efforts, Recommendation engines, Facial detection, Detect malware, Medical diagnosis, Self driving cars, Develop pharmaceuticals, and many more.

My call to action

- Please hack machine learning models!
- Watchdog
- Joy of finding things out
- Learn a useful skill by inspection
- Improve ML applications as a whole

What does it mean to hack it?

- Apply wrong label
- Put person in a different cluster
- Bias output of regression prediction
- Make inefficient/sub-optimal decision
- Weaken the usefulness of an observable variable

Similarities to other hacking

- Take machine apart vs. manipulate inputs and outputs
- Brute force
- May not have direct access to a system
- **Looking for oversights or mistakes made by model creators**

Similarities to other hacking

- How did the creator build it and what design choices did they make?
- How can I get the system to reveal it’s own weaknesses?

Diff from other hacking

- Inspect model parameters, not source code
- ML systems are can use online learning / adaptive
- The inner workings are more obfuscated

Auditing Algorithms

1. Code audit (algorithmic transparency)
2. Noninvasive User Audit (logs, user survey)
3. Scraping Audit (direct observation)
4. Sock Puppet Audit (direct interaction)
5. Crowdsource Audit (collaborative audit)

"Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms"

- Sandvig, Hamilton, Karahalios, and Langbort

Discussion of Techniques

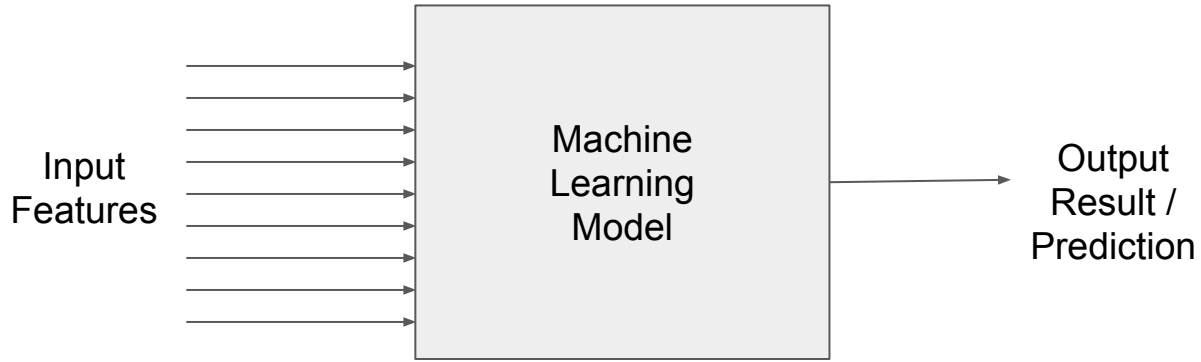
Ability to interact with model directly and many times

vs.

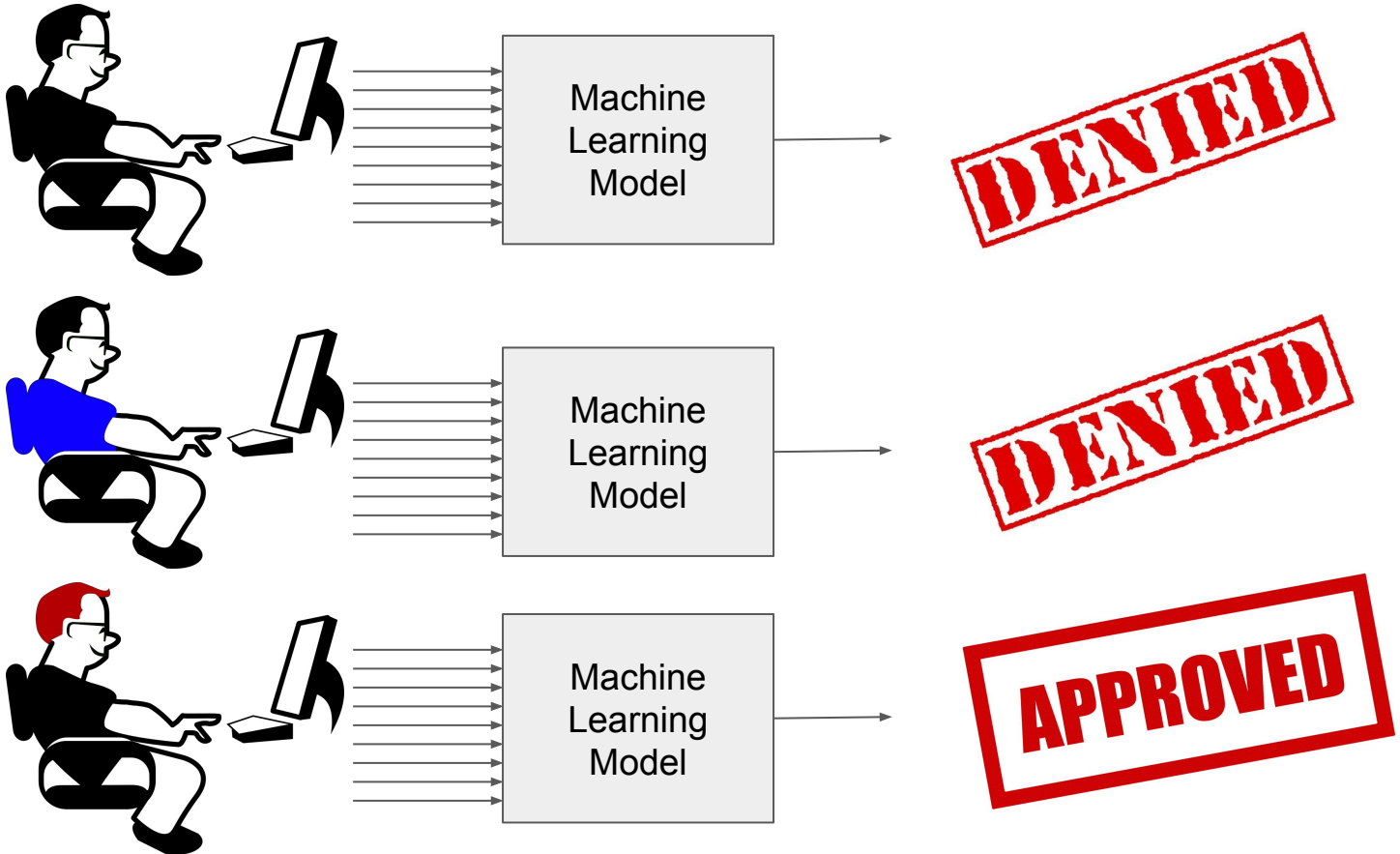
Limited access to interacting with model.

Code Audit

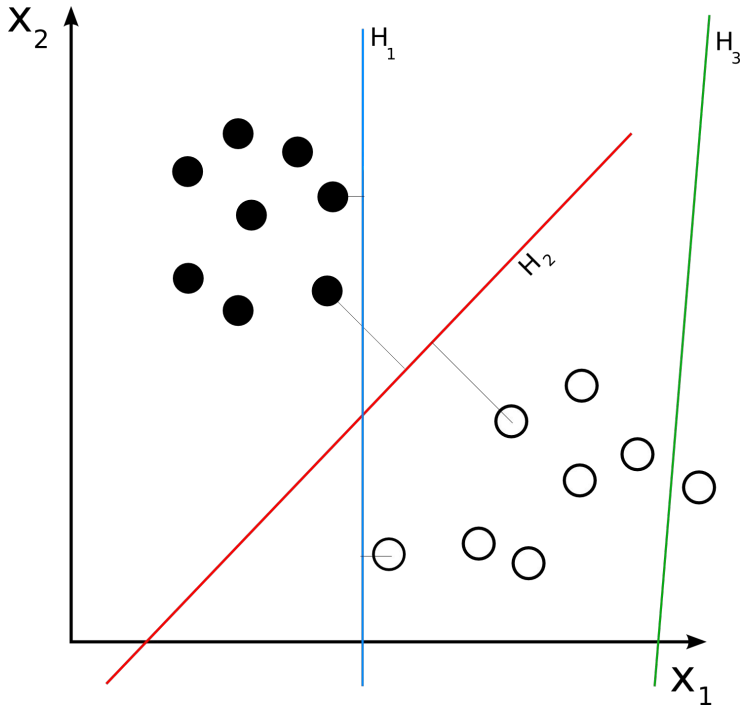
- Source code vs. model code
- Can’t pour over model code effectively
- Assume the black box



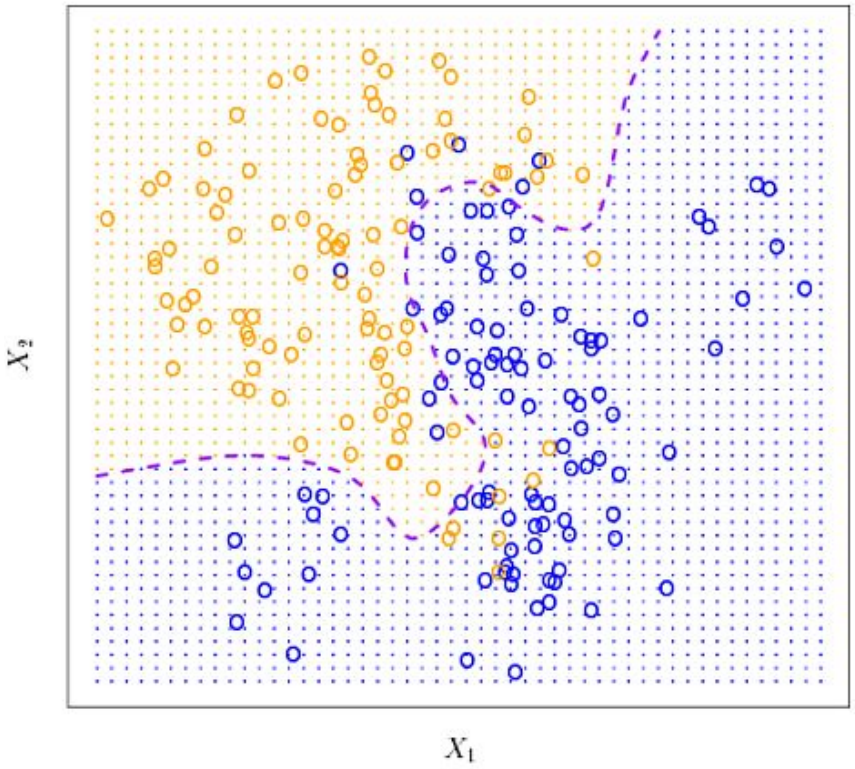
Feature Space Exploration



Decision Boundaries



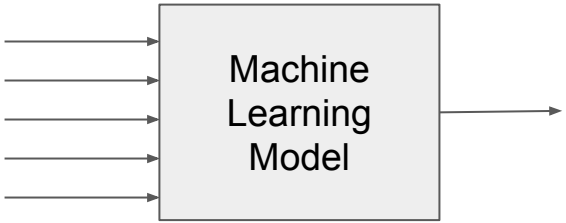
Decision Boundaries

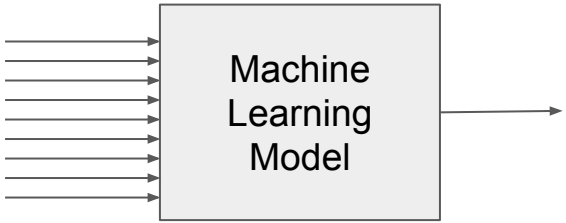


Decision Boundaries

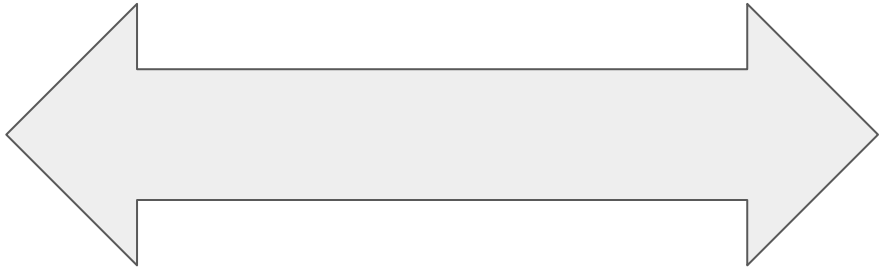
- Feature vector manipulation
- The curse of dimensionality
- Priors and domain knowledge
- Gradient descent

Feature space





Hand picked feature sets

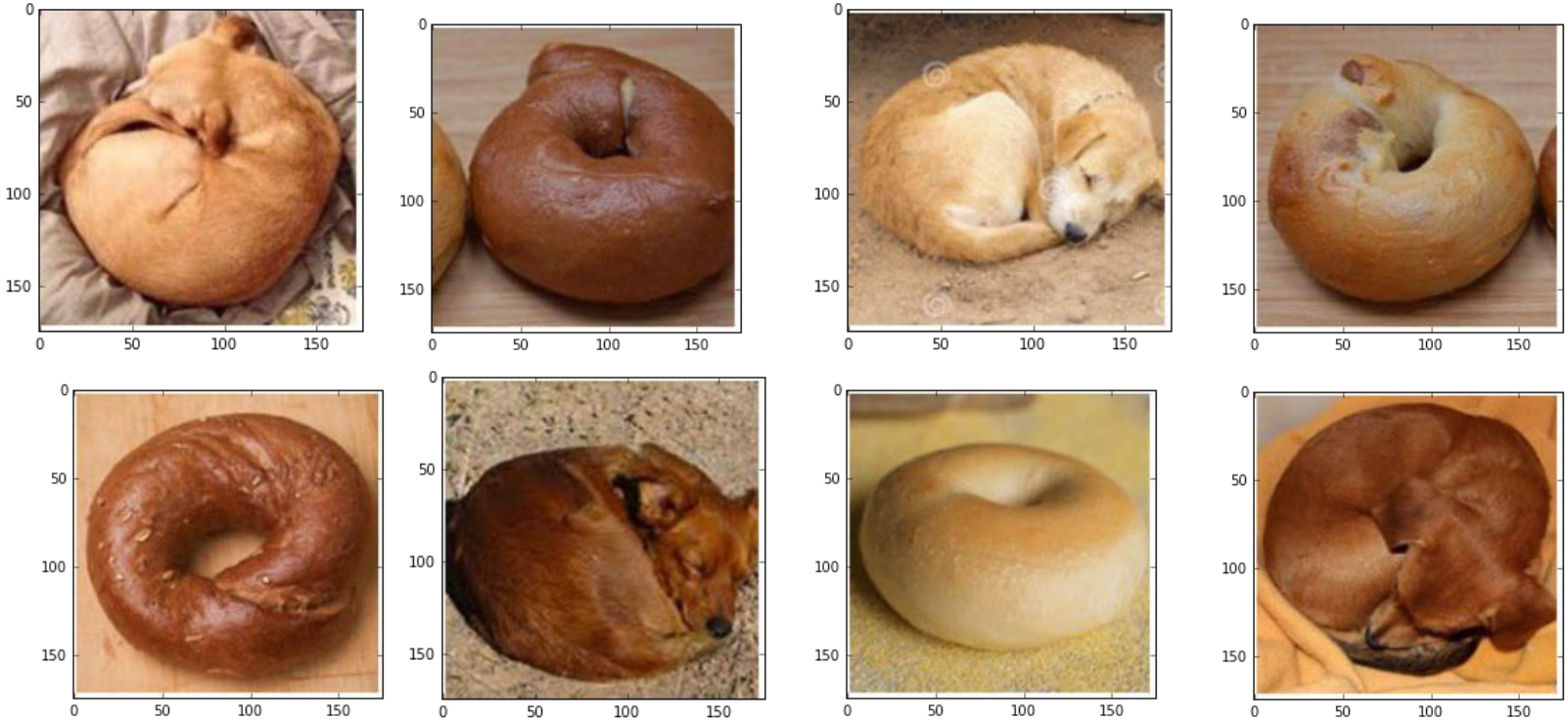


Kitchen sink feature sets

Adversarial ML

- Discriminator - determine whether data is from generative network or not
- Generative - generate new data based on existing training data

Adversarial ML



https://github.com/yskmt/dog_recognition/blob/master/dog_bagel.ipynb

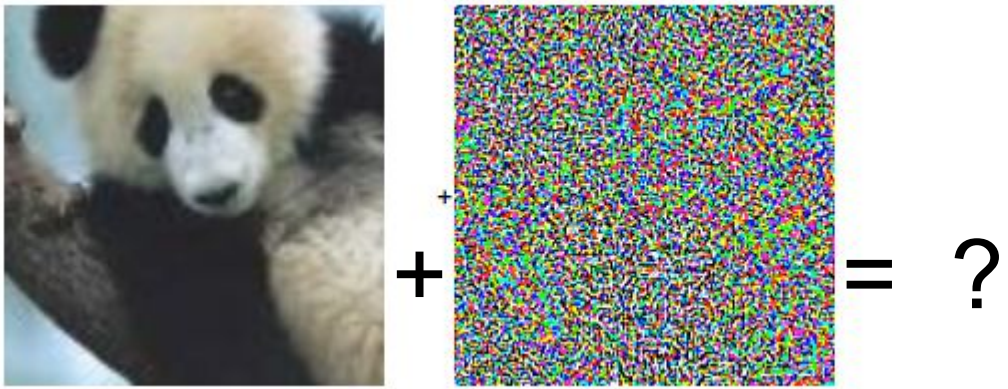
Intriguing properties of neural networks

Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus

Adversarial Training

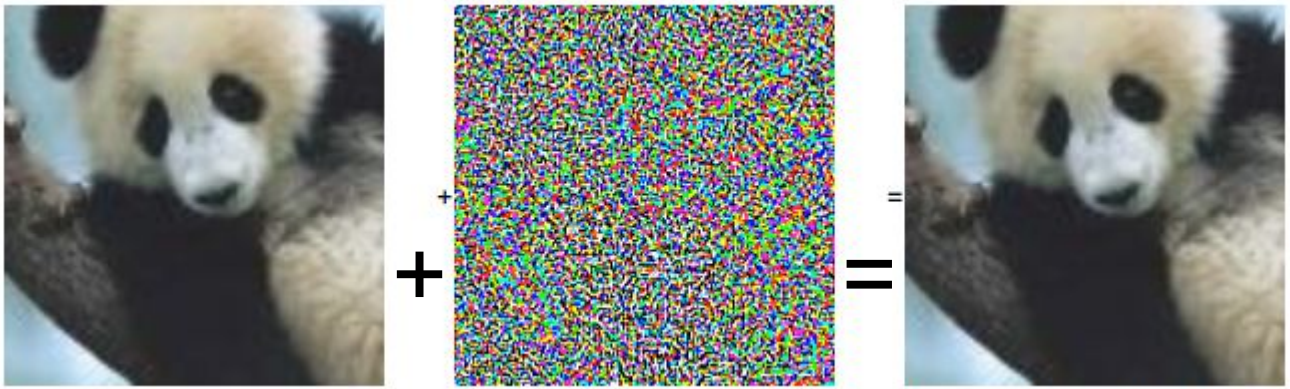
Panda

Adversarial Training



Panda Adjustment

Adversarial Training

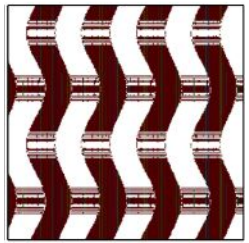


Panda Adjustment Gibbon

Adversarial training

Deep neural networks are easily fooled

- Nguyen, Yosinski, Clune

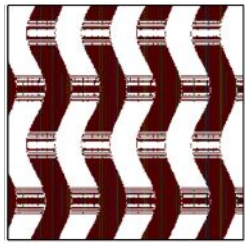


Adversarial training

Deep neural networks are easily fooled

- Nguyen, Yosinski, Clune

African Grey parrot



Guitar

Discussion of Techniques

Ability to interact with model directly and many times

vs.

Limited access to interacting with model.

Sock Puppet Audit

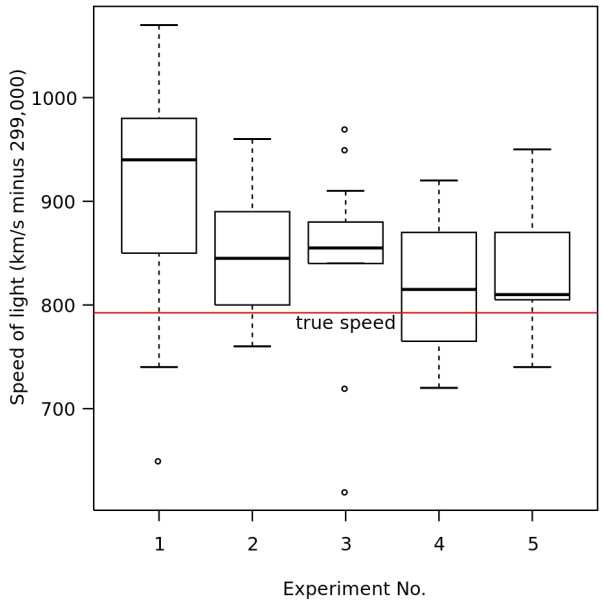
My favorite methods: Scrapy/requests, Beautiful Soup, Selenium, Chrome Developer console, JSON

Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function

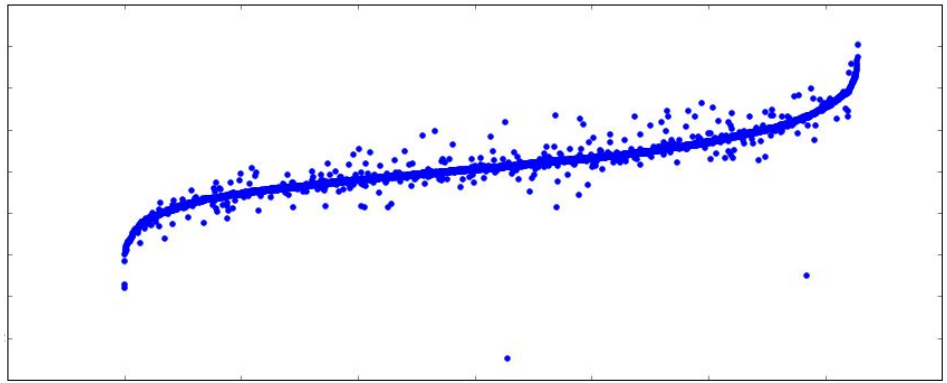
Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



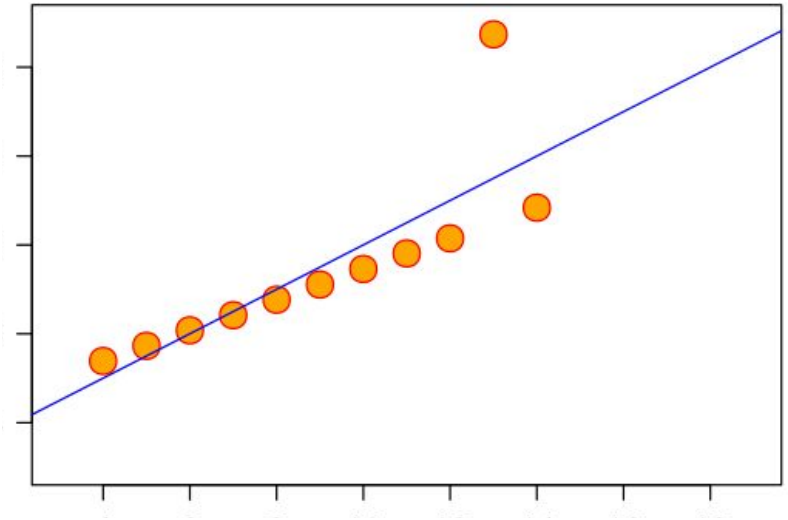
Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



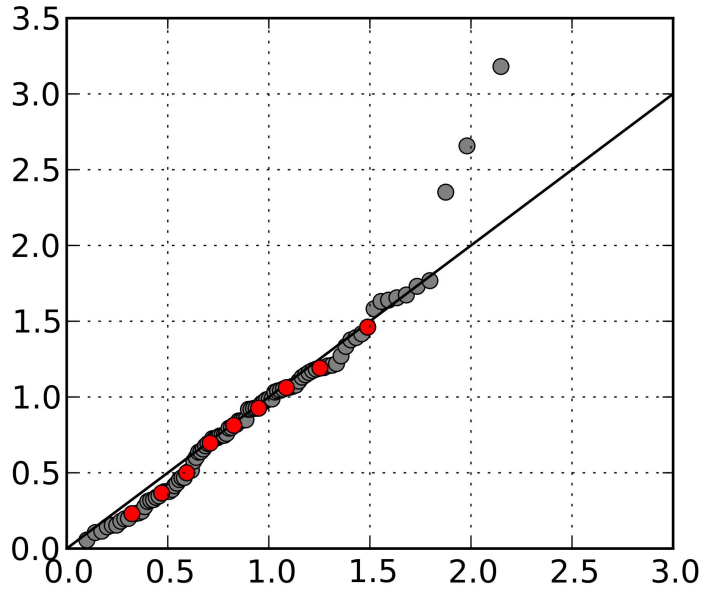
Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



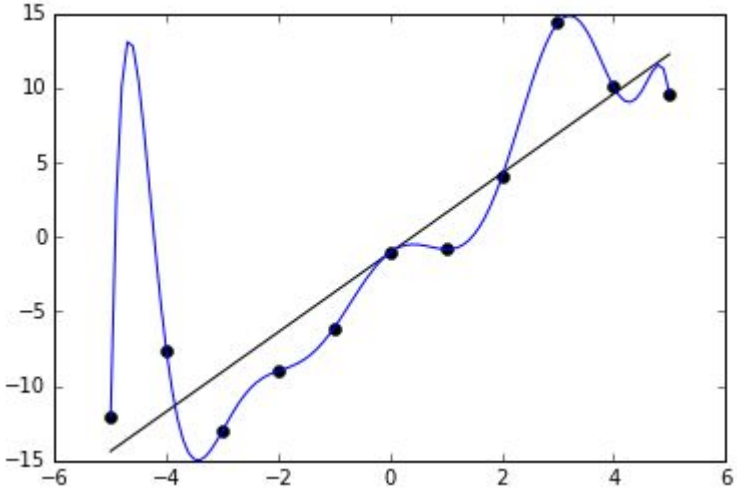
Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



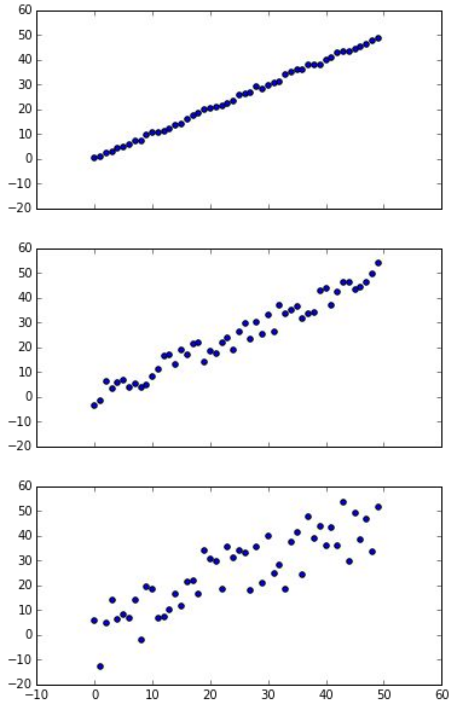
Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



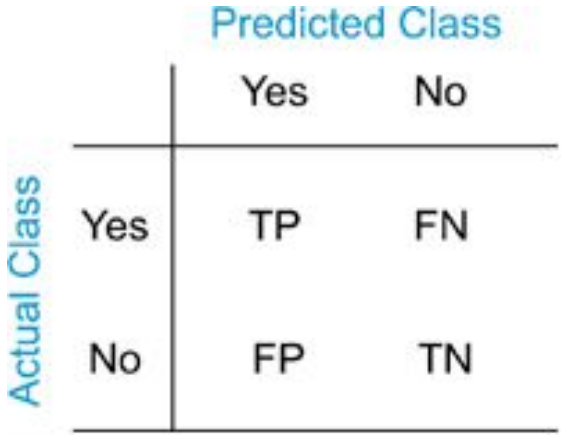
Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



Think like an ML researcher

- Data exploration / summary statistics
- Data cleansing
- Overfitting
- CV and other metrics
- Objective function



Bayesian Poisoning

- In spam filtering (assume Naive Bayes), attempt to degrade effectiveness:

$$P(spam|prince) = \frac{P(prince|spam) \cdot P(spam)}{P(prince)}$$

Bayesian Poisoning

- In spam filtering (assume Naive Bayes), attempt to degrade effectiveness:

$$P(Email \in \neg spam) = \prod P(word_i | \neg spam)$$

$$P(Email \in spam) = \prod_i P(word_i | spam)$$

Bayesian Poisoning

- In spam filtering (assume Naive Bayes), attempt to degrade effectiveness:

Hi Kyle,

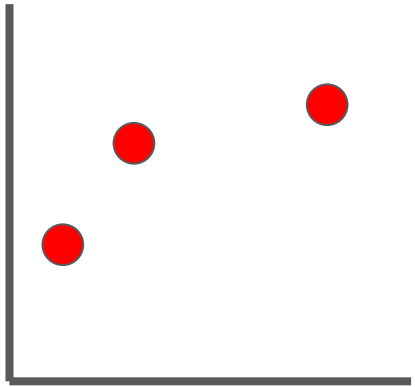
A common way people might try to manipulate a spam filter is by attempting to cause false positives (type i errors) by adding spam-like content. FYI!

Amazon.com Amazon.com Amazon.com
Also, V@agria, Mail order Brid3s, 0dayT0rreNTz

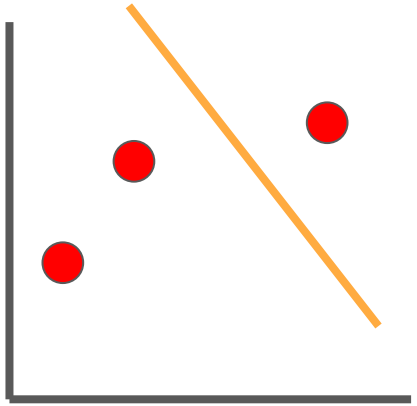
Your selection of current executive positions
Your selection of current executive openings. Find new opportunities and high-end positions chosen just for you. Updated regularly, so save your favorites!

Hi Uncle Kyle how are you doing I lost a tooth!!!!!!

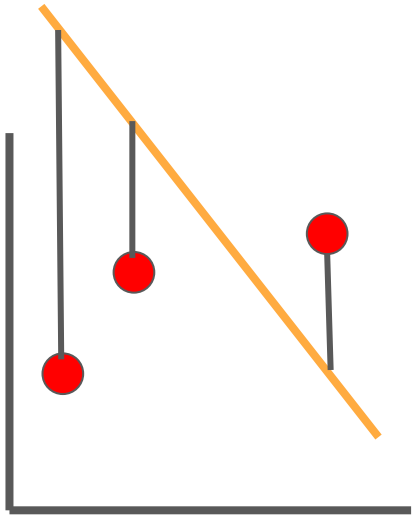
Regression



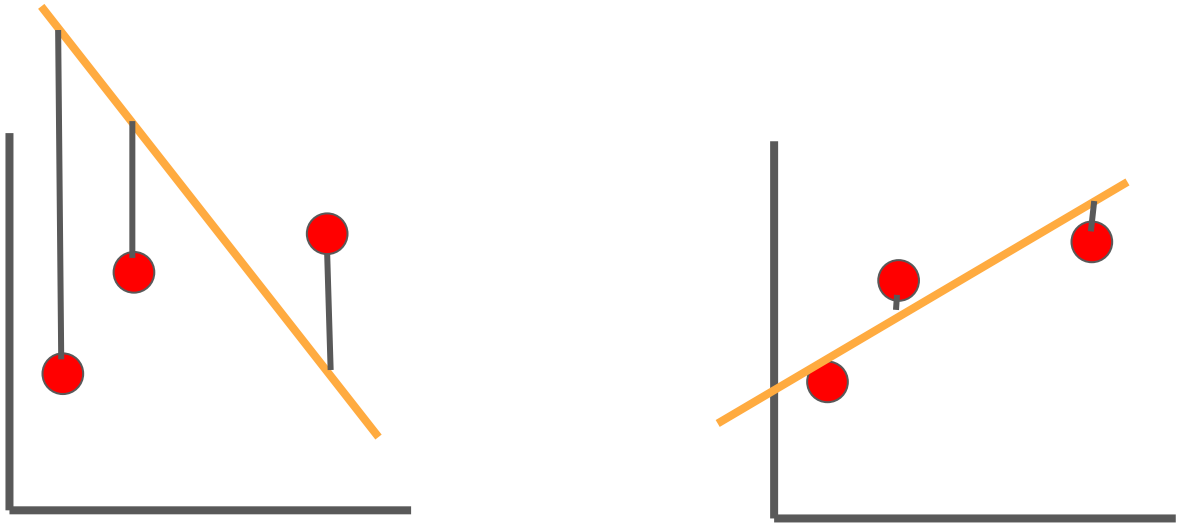
Regression



Regression

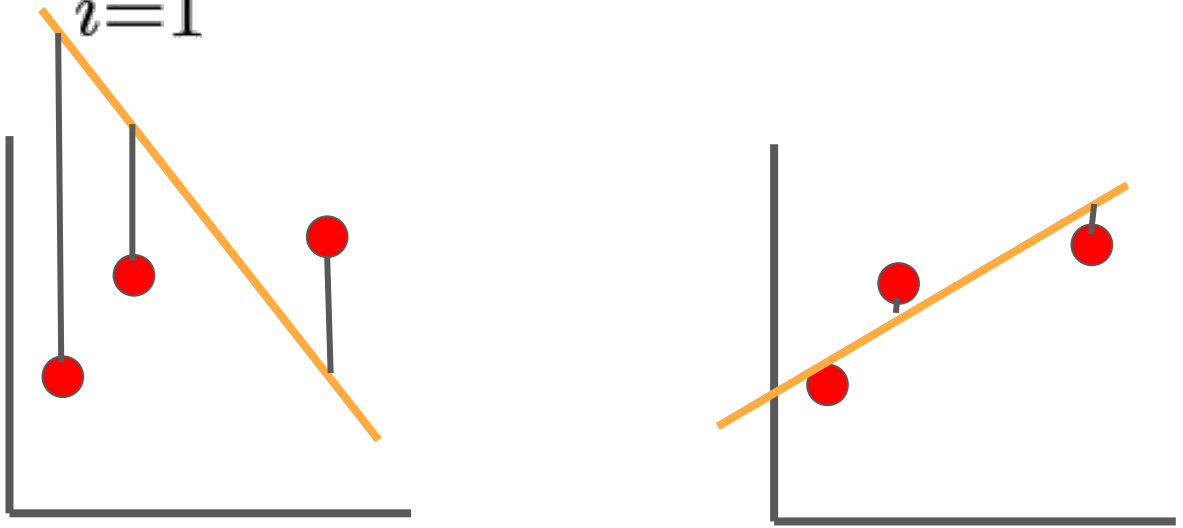


Regression



Regression

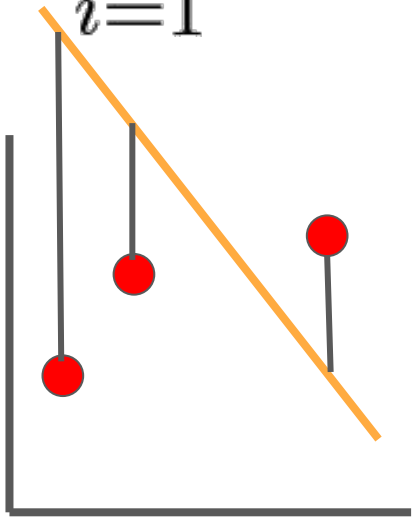
$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

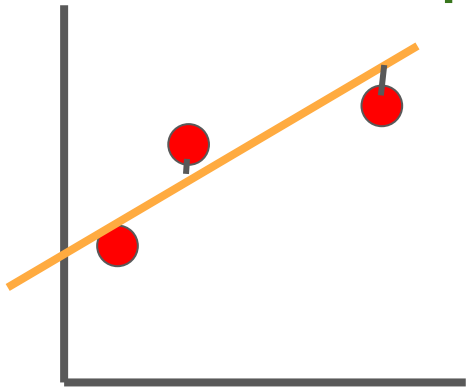


Regression

$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

(prediction$_i$ is ith prediction or x$_i$ * Beta$_i$)

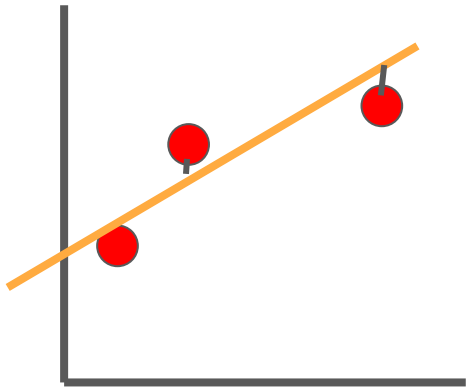




Regression

$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

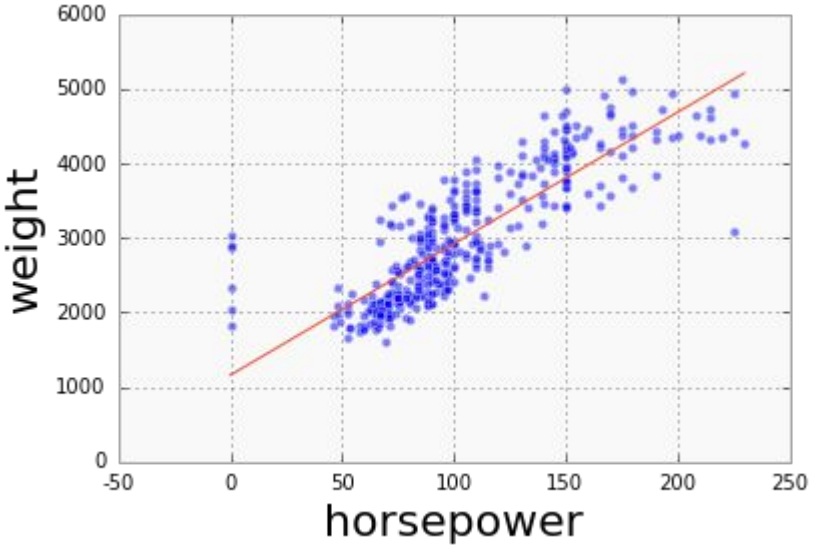
$$\underset{\beta}{\operatorname{argmin}} \Big(L(\beta) \Big)$$



Regression

$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

$$\underset{\beta}{\operatorname{argmin}} \Big(L(\beta) \Big)$$

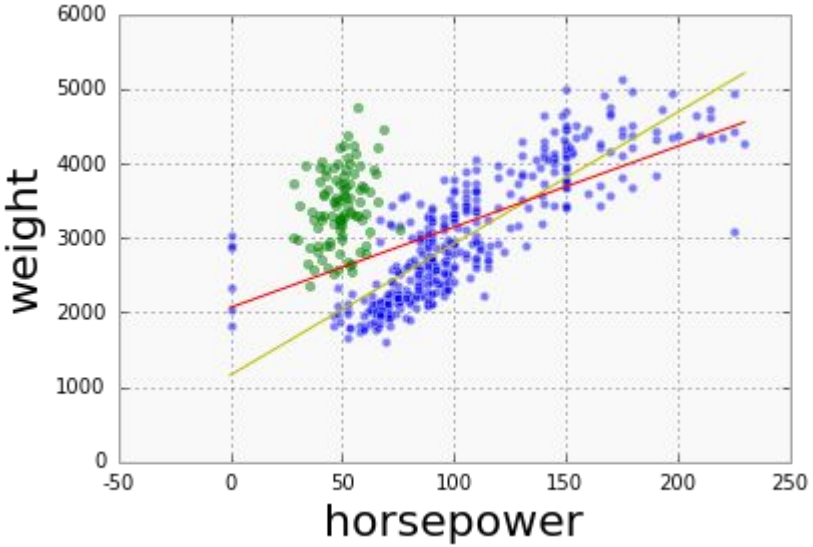


Weight = 17.6 * horsepower + 1157.1

Regression

$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

$$\underset{\beta}{\operatorname{argmin}} \Big(L(\beta) \Big)$$



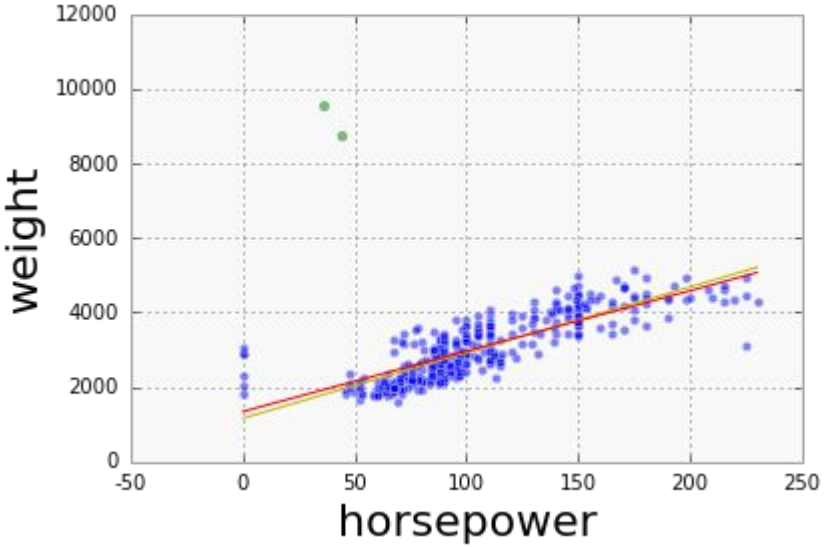
Weight = 17.6 * horsepower + 1157.1

Weight = 10.8 * horsepower + 2063.6

Regression

$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

$$\underset{\beta}{\operatorname{argmin}} \Big(L(\beta) \Big)$$



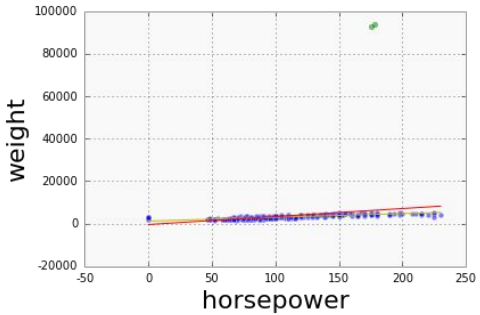
Weight = 17.6 * horsepower + 1157.1

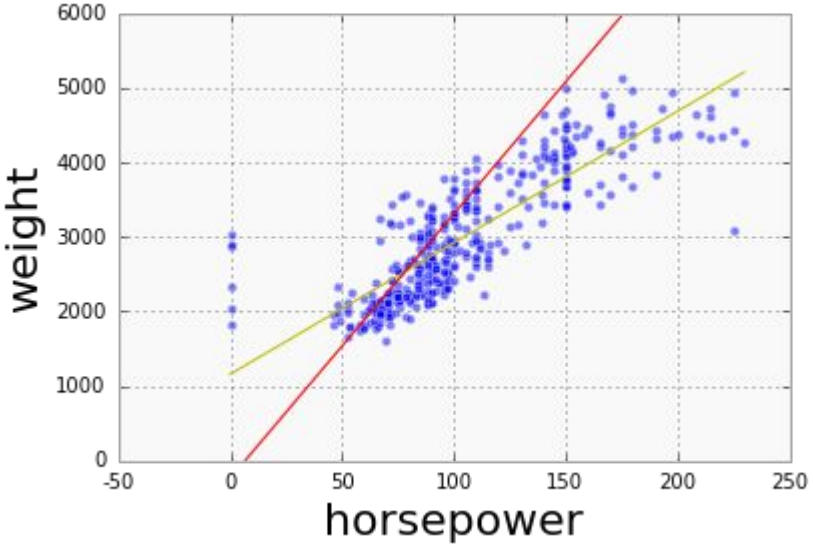
Weight = 16.2 * horsepower + 1339.1

Regression

$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2$$

$$\underset{\beta}{\operatorname{argmin}} \Big(L(\beta) \Big)$$





Weight = 17.6 * horsepower + 1157.1

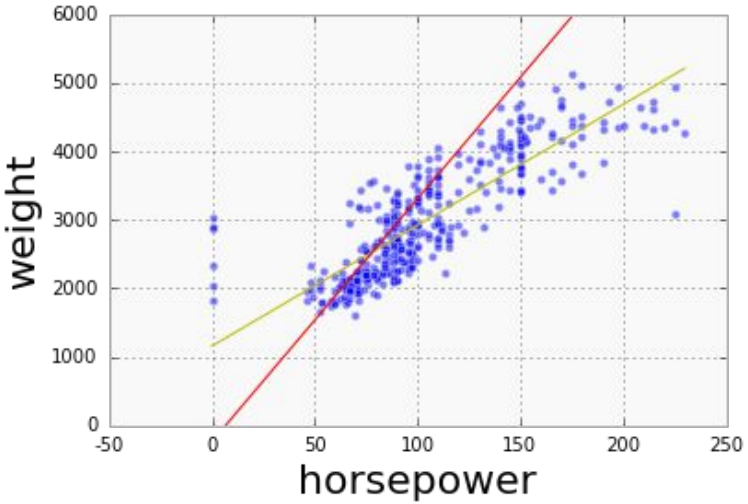
Weight = 35.4 * horsepower + -237.9

Regression, Regularization

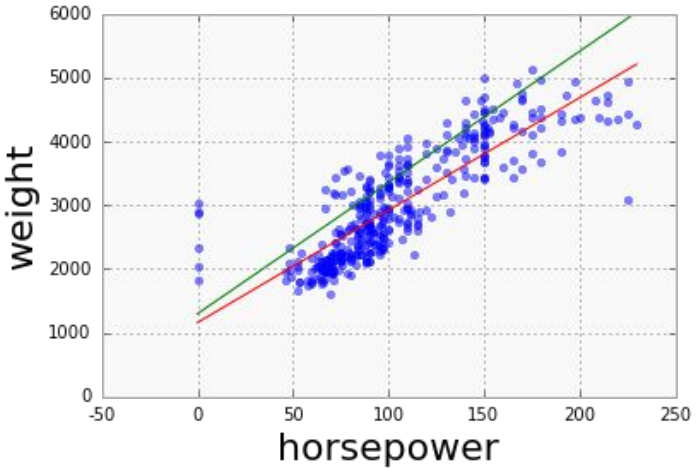
$$L(\beta) = \sum_{i=1}^{n} (actual_i - prediction_i)^2 \textbf{ + Penalty}$$

$$\underset{\beta}{\operatorname{argmin}} \Big(L(\beta) \Big)$$

Unregularized



Regularized



Regularization

Additional constraints to help solve an ill-posed problem or prevent overfitting

$$\underset{\beta}{\operatorname{argmin}}\left(L(\beta) + \lambda_1 ||\beta||_1\right)$$

L1 (Ridge) - Penalize for many parameters

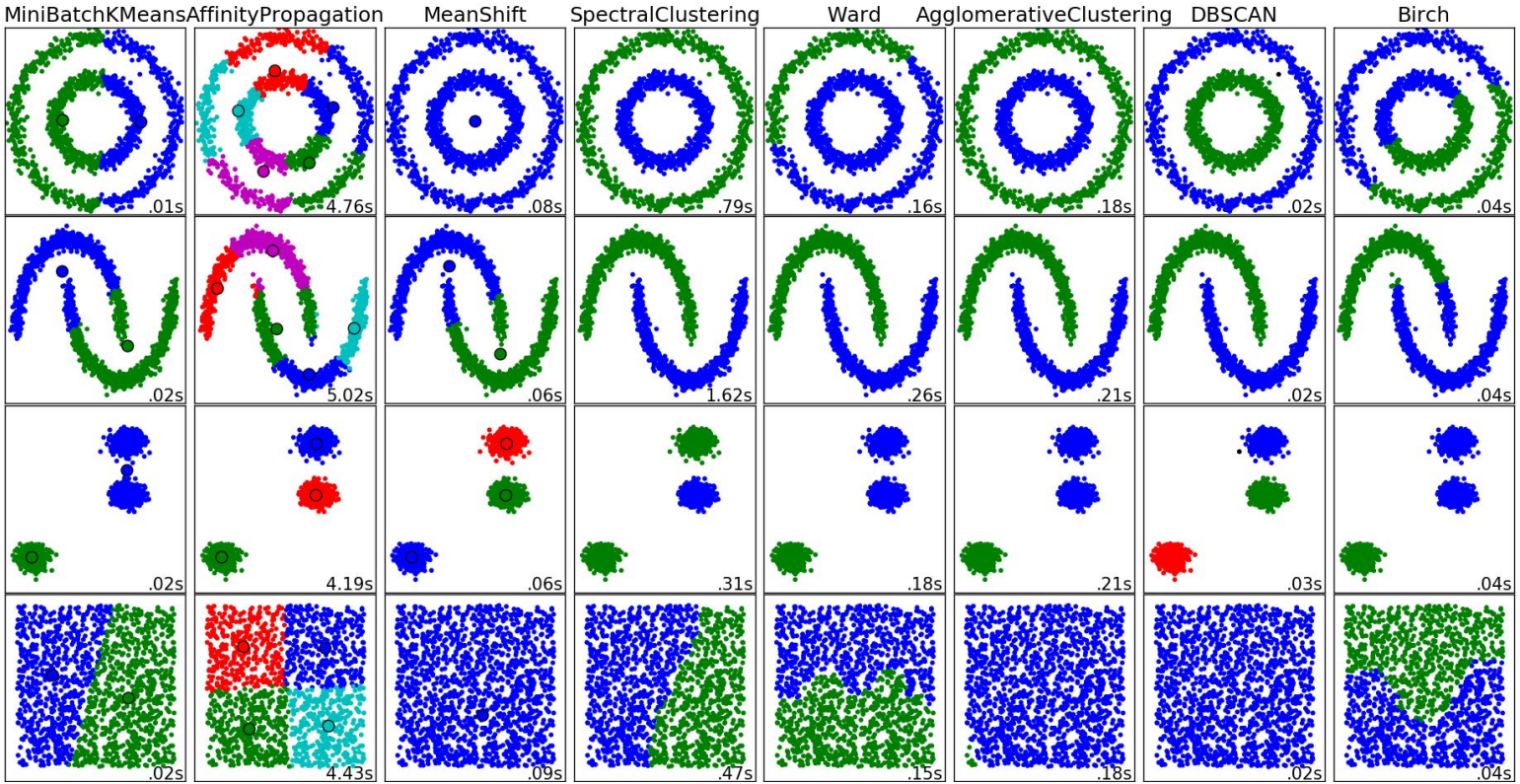
$$\underset{\beta}{\operatorname{argmin}}\left(L(\beta) + \frac{\lambda_2}{2}||\beta||^2\right)$$

L2 (Lasso) - Penalize for large values

Sentiment example

Let’s go to Jupyter

Geo Clustering



Some reminders

- Direct-ish access to model?
  - Reverse engineer, typically via search or inspection
- Partial access to model?
  - Reverse engineer, typically via (guided) trial and error
  - Feature vector “hacking”
- Ability to contribute to training dataset?
  - Introduce bias
  - Add noise

Why again?

- Discriminatory (not nice kind) algorithms
- ML community can benefit from challenges
- Same reason as why you hack
- Let’s consider bug bounties?
- Hack responsibly!

Thanks

@DataSkeptic

dataskeptic.com

Q&A