Reusability of Holdout Set Using Differential Privacy
The Magic Broker Scam

Based on Aaron Roth’s slide
The Magic Broker Scam
Based on Aaron Roth’s slide
The Magic Broker Scam
Based on Aaron Roth’s slide

Day 1
Tomorrow Stock X

Day 2
Tomorrow Stock X
The Magic Broker Scam
Based on Aaron Roth’s slide

Day 1
Tomorrow Stock X

Day 2
Tomorrow Stock X

Day 3
Tomorrow Stock X
The Magic Broker Scam
Based on Aaron Roth’s slide

Day 1
Tomorrow Stock X

Day 2
Tomorrow Stock X

Day 3
Tomorrow Stock X

..............

Day 10
Tomorrow Stock X
The Magic Broker Scam
Based on Aaron Roth’s slide

- Day 1: Tomorrow Stock X
- Day 2: Tomorrow Stock X
- Day 3: Tomorrow Stock X
- Day 10: Tomorrow Stock X
- Day 11: Pay me I will help you invest
The Magic Broker Scam

Based on Aaron Roth’s slide

- Day 11: Pay me I will help you invest
- What are the chances that someone guesses randomly and gets correct?
The Magic Broker Scam

Based on Aaron Roth’s slide

- Day 1: Tomorrow Stock X
- Day 2: Tomorrow Stock X
- Day 3: Tomorrow Stock X
- Day 10: Tomorrow Stock X

- Day 11: Pay me I will help you invest

- What are the chances that some one guesses randomly and gets correct?
  - 1 in 1024
The Magic Broker Scam

Based on Aaron Roth’s slide

- Day 1: Tomorrow Stock X
  - Thumb up

- Day 2: Tomorrow Stock X
  - Thumb down

- Day 3: Tomorrow Stock X
  - Thumb down

- Day 10: Tomorrow Stock X
  - Thumb down

- Day 11: Pay me I will help you invest

- What are the chances that someone guesses randomly and gets correct?
  - 1 in 1024

- But the broker will always scam people, why?
The Magic Broker Scam

Based on Aaron Roth’s slide

Day 1
Tomorrow Stock X

Day 2
Tomorrow Stock X

Day 3
Tomorrow Stock X

Day 10
Tomorrow Stock X
The Magic Broker Scam
Based on Aaron Roth’s slide

Day 1: send email to a million people (1/2 + and 1/2 -)

• Day 1: send email to a million people (1/2 + and 1/2 -)
The Magic Broker Scam

Based on Aaron Roth’s slide

- Day 1: send email to a million people (1/2 + and 1/2 -)
- Day 2: only send email to 500,000 people we got right (1/2 + and 1/2 -)
The Magic Broker Scam
Based on Aaron Roth’s slide

- Day 1: send email to a million people (1/2 + and 1/2 -)
- Day 2: only send email to 500,000 people we got right (1/2 + and 1/2 -)
- …….
The Magic Broker Scam

Based on Aaron Roth’s slide

- Day 1: send email to a million people (1/2 + and 1/2 -)

- Day 2: only send email to 500,000 people we got right (1/2 + and 1/2 -)

- ..... 

- Day 11: We are left with $\frac{1000000}{1024} \approx 1000$ people we can scam
Reproducibility problem
Ideal ML Experiment
Ideal ML Experiment

Data
Ideal ML Experiment

Data → Training → Holdout
Ideal ML Experiment

Data → Training → Holdout → Reuse many times
Ideal ML Experiment

Data → Training → Holdout → Reuse many times → Form hypothesis
Ideal ML Experiment

Data → Training → Holdout → Form hypothesis

Reuse many times

Use holdout set only once and report accuracy
• We are not allowed to form hypothesis based on data we used to test: age old statistics
Ideal ML Experiment

• We are not allowed to form hypothesis based on data we used to test: age old statistics

• But too tempting to form more informed opinion
Ideal ML Experiment

- We are not allowed to form hypothesis based on data we used to test: age old statistics
- But too tempting to form more informed opinion
- We do a train/validation/test set split
Ideal ML Experiment

- We are not allowed to form hypothesis based on data we used to test: age old statistics
- But too tempting to form more informed opinion
- We do a train/validation/test set split
- But this means we can’t reuse datasets over time
ML Today
ML Today

• Benchmark dataset from MNIST to IMAGENET
ML Today

• Benchmark dataset from MNIST to IMAGENET

• Competitions run with feedback from test set to competitors ;)
ML Today

• Benchmark dataset from MNIST to IMAGENET

• Competitions run with feedback from test set to competitors ;)

• Effort to collect public dataset to share....
ML Today

- Benchmark dataset from MNIST to IMAGENET
- Competitions run with feedback from test set to competitors ;)
- Effort to collect public dataset to share….
- All great but we need to be careful!
ML Today

• Benchmark dataset from MNIST to IMAGENET

• Competitions run with feedback from test set to competitors ;)

• Effort to collect public dataset to share....

• All great but we need to be careful!

  • Old fix: pre-register experiment, many many examples of people fudging this....
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

- Data: 20,000 data points in 10000 dimensions drawn randomly
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

- Data: 20,000 data points in 10000 dimensions drawn randomly
- Labels: random +1 or -1 (no correlation with input)
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

- Data: 20,000 data points in 10000 dimensions drawn randomly
- Labels: random +1 or -1 (no correlation with input)
- Data-scientist runs:
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

- Data: 20,000 data points in 10000 dimensions drawn randomly

- Labels: random +1 or -1 (no correlation with input)

- Data-scientist runs:
  - Split data into train and test/holdout of equal size
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

• Data: 20,000 data points in 10000 dimensions drawn randomly

• Labels: random +1 or -1 (no correlation with input)

• Data-scientist runs:
  • Split data into train and test/holdout of equal size
  • Select best k features on training data (using magnitude of correlation)
An Example

Code at:
https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

• Data: 20,000 data points in 10000 dimensions drawn randomly

• Labels: random +1 or -1 (no correlation with input)

• Data-scientist runs:
  • Split data into train and test/holdout of equal size
  • Select best k features on training data (using magnitude of correlation)
  • Drop features that don’t have same sign of correlation on holdout
An Example

Data: 20,000 data points in 10000 dimensions drawn randomly

Labels: random +1 or -1 (no correlation with input)

Data-scientist runs:

• Split data into train and test/holdout of equal size
• Select best k features on training data (using magnitude of correlation)
• Drop features that don’t have same sign of correlation on holdout
• Build linear predictor out of these selected variables (on training set)

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

• Data: 20,000 data points in 10000 dimensions drawn randomly

• Labels: random +1 or -1 (no correlation with input)

• Data-scientist runs:
  • Split data into train and test/holdout of equal size
  • Select best k features on training data (using magnitude of correlation)
  • Drop features that don’t have same sign of correlation on holdout
  • Build linear predictor out of these selected variables (on training set)
  • Find best k = 10,20,30,40,….
An Example

Number of features
An Example

Code at: https://github.com/isofer/thresholdout-experiments/blob/master/Thresholdout%20experiments.ipynb

• Data: 20,000 data points in 10000 dimensions drawn randomly

• Labels: random +1 or -1 (no correlation with input)

• Data-scientist runs:
  • Split data into train and test/holdout of equal size
  • Select best k features on training data (using magnitude of correlation)
  • Drop features that don’t have same sign of correlation on holdout
  • Build linear predictor out of these selected variables (on training set)
  • Find best k = 10,20,30,40,….
Can we reuse holdout set?
Maybe if we do it right...
Maybe if we do it right...

• Idea: when we report back accuracies from the dataset, we add noise so as to not to leak too much information
Maybe if we do it right…

• Idea: when we report back accuracies from the dataset, we add noise so as to not to leak too much information

• Eg. Report back accuracies on holdout set only when training and test accuracy are significantly different
Threshold-out Algorithm

Input:
Data $S$, holdout $H$, threshold $T > 0$, tolerance $\sigma > 0$

Given function $q$:

Sample $\eta, \eta'$ from $N(0,\sigma^2)$

If $|\text{avg}_H[q] - \text{avgs}[q]| > T + \eta$:
  output $\text{avg}_H[q] + \eta'$

Otherwise:
  output $\text{avgs}[q]$
The Example
The Guarantee

• Say we have a holdout set of size n

• Rough statement: Thresholdout gives $\Delta$-accurate estimates for any sequence of adaptively chosen queries until $O(\Delta^4 n^2)$ we report holdout accuracies.
Threshold-out Algorithm

Input:
Data $S$, holdout $H$, threshold $T > 0$, tolerance $\sigma > 0$

Given function $q$:

Sample $\eta$, $\eta'$ from $N(0,\sigma^2)$

If $|\text{avg}_H[q] - \text{avgs}[q]| > T + \eta$:
  output $\text{avg}_H[q] + \eta'$

Otherwise:
  output $\text{avgs}[q]$
Differential Privacy

- Post-processing: If $A$ is a differentially private algorithm, and $f$ is any mapping on the outcome space of $A$, then the algorithm that maps from sample $S$ to $f(A(S))$ is also differentially private.

- Composability: $\text{Algo}(S)$ is given by the following procedure:
  
  - For $i$ in 1 to $k$
    
    - If we choose any epsilon differentially private algorithm $A_i$ based on outcomes $O_1, \ldots, O_{i-1}$
    
    - Set $O_i = A_i(S)$
  
  - Return $O_1, \ldots, O_n$

- The above algorithm is $O(\varepsilon \sqrt{k})$ differentially private.
Differential Privacy

- Differential privacy implies generalization (via stability)

- Threshold-out is differentially private. Hence holdout set accuracy reported and population accuracy are close