Morse Code Datasets for Machine Learning

Sourya Dey, Keith Chugg, Peter Beerel

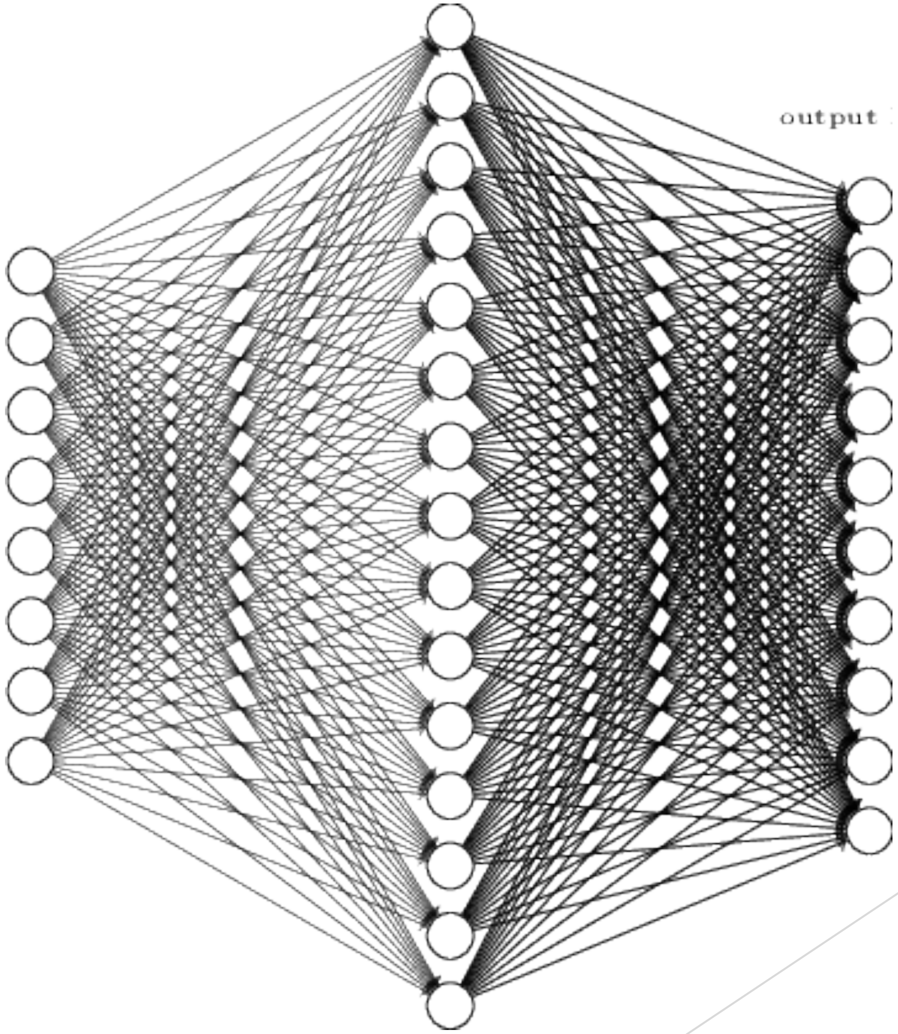
9th International Conference on Computing, Communication and Networking Technologies

July 2018

USC University of Southern California

Machine Learning and Neural Networks

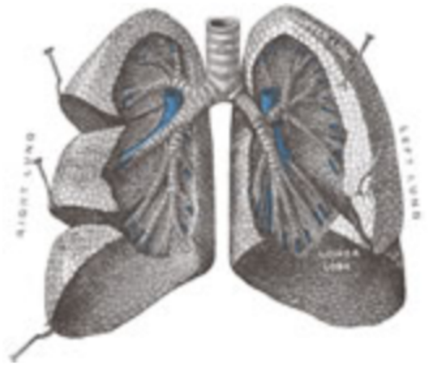
An algorithm to learn from data and classify it



Need a lot of data for good performance

Issues with Natural Data

- Most data is naturally collected and labeled by humans
- Labeling is time-consuming (e.g. Imagenet[1])
- Data can have missing features (e.g. Lung cancer dataset[2])



1: http://www.image-net.org/
2: http://archive.ics.uci.edu/ml/datasets/Lung+Cancer

Synthetic data as a Solution

- **Synthetic data** generated and labeled using algorithms
- Can be mass-produced cheaply without missing features

- Algorithm can be tuned to:
  - *Adjust difficulty*
  - Get any distribution

Overview of our Work

- Algorithm to generate Morse code classification datasets of varying difficulty
- Metrics to evaluate difficulty of a dataset

Morse code is a system of communication to encode characters as dots and dashes

\+   • ▬ • ▬ •

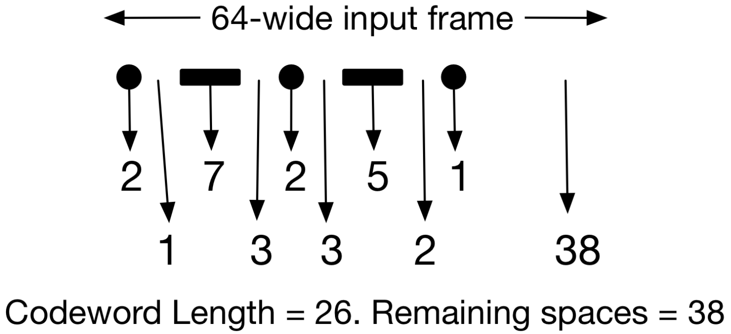
64 character classes

The Algorithm

Step 1:

Frame length: 64

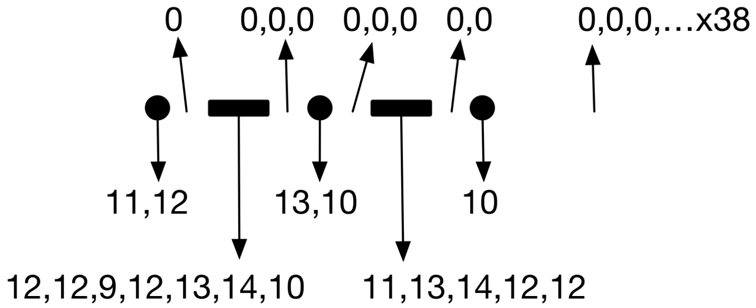
Dot: 1-3
Dash: 4-9
Intermediate space: 1-3
Leading spaces: None
Trailing spaces: Remaining at end



Step 2:

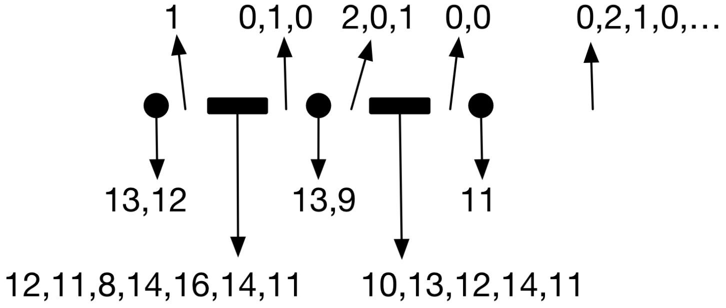
Expected value range = [0,16]

Dot, dash = *Normal*(12,4/3)
Space = 0



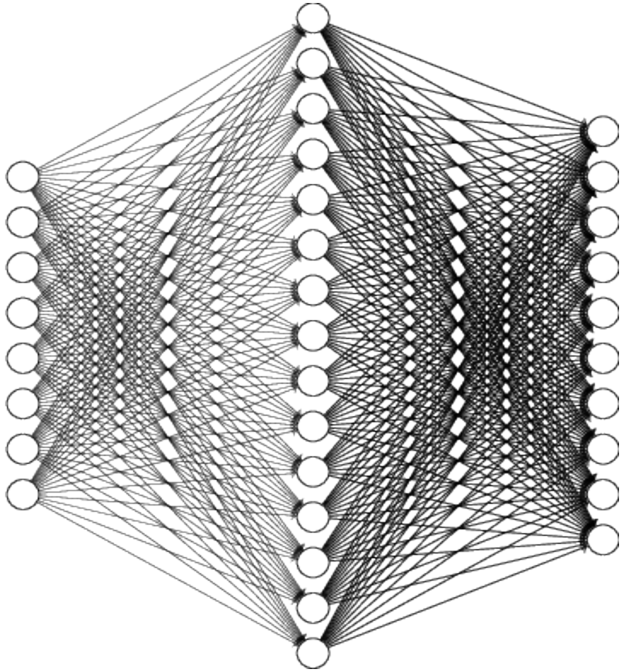
Step 3:

Additive Noise = *Normal*$(0,\boldsymbol{\sigma})$
(For this case, $\boldsymbol{\sigma}=1$)



The Neural Network

64 input neurons = Frame length of each Morse codeword

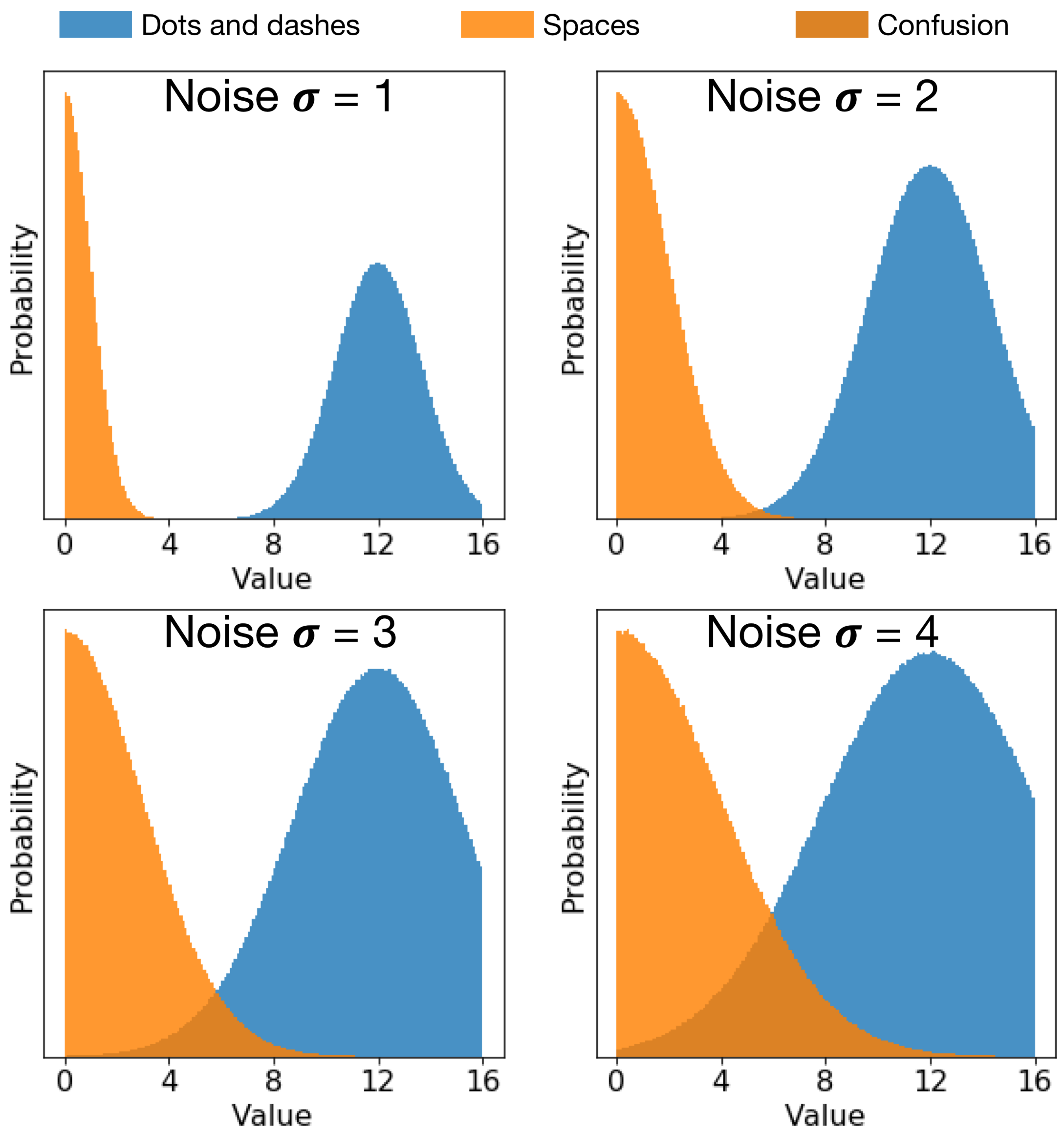


64 output neurons = Number of character classes

1024 hidden neurons

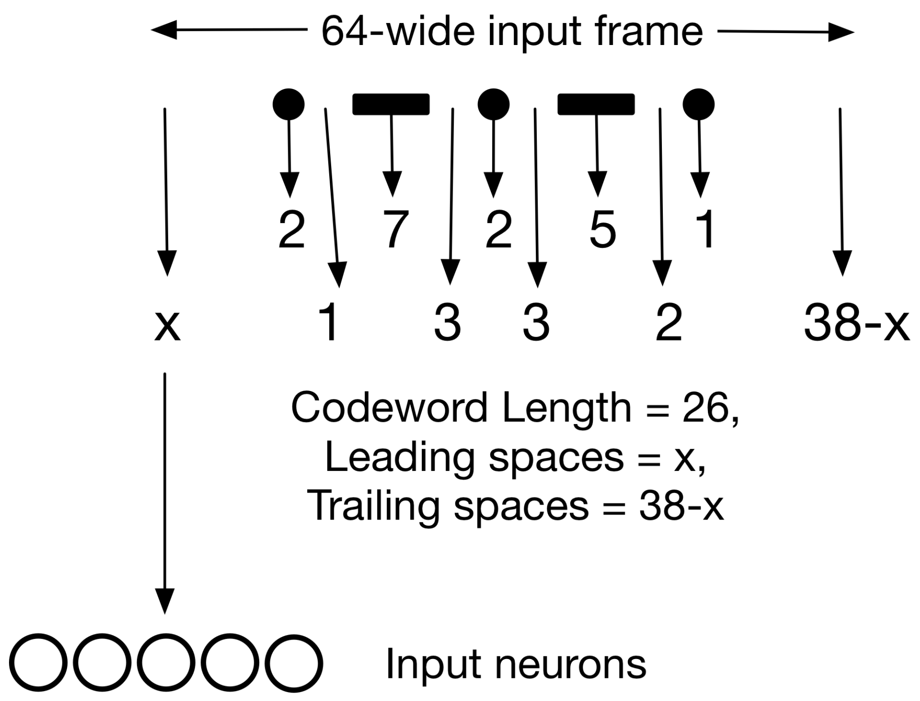
Variations and Difficulty Scaling - 1

Increasing σ of noise leads to confusion between dots, dashes and spaces



Variations and Difficulty Scaling - 2

Distribute remaining spaces randomly between leading and trailing



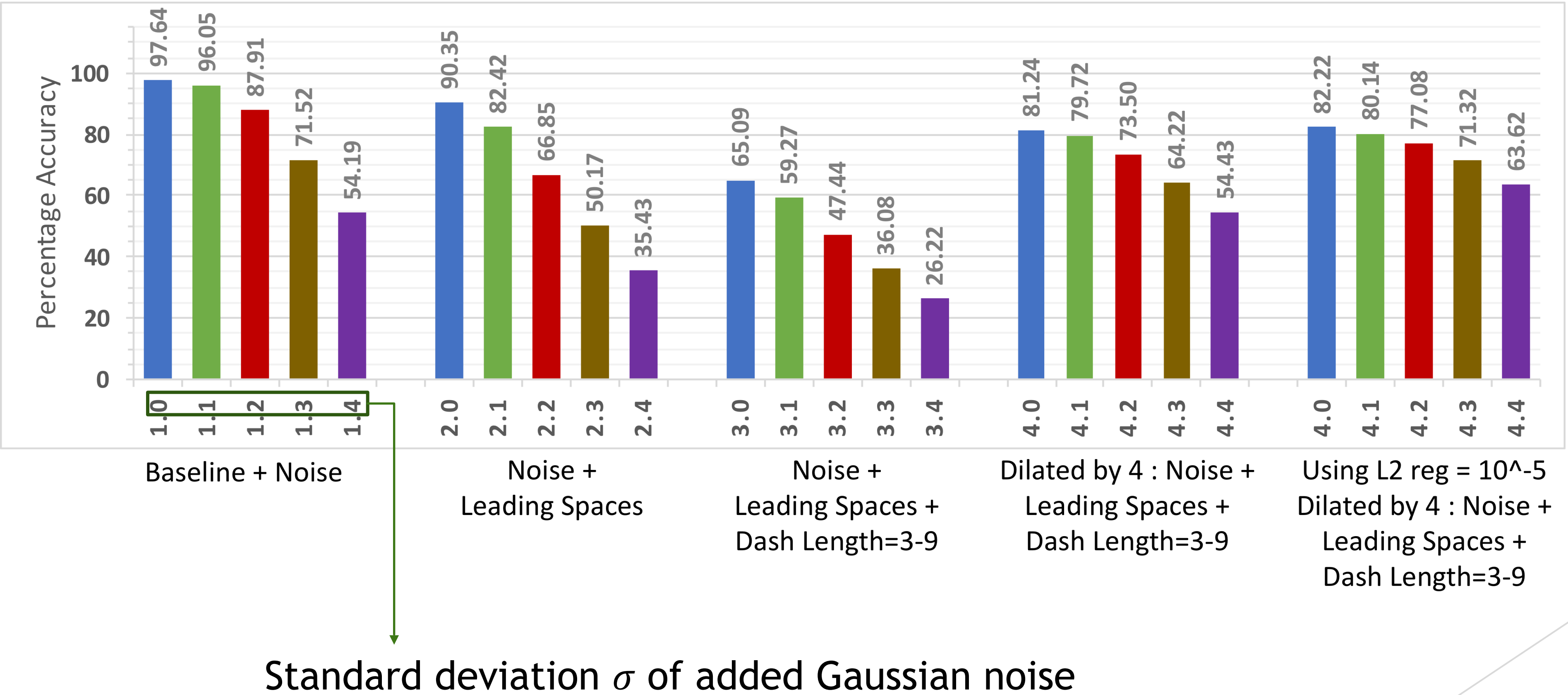
Variations and Difficulty Scaling - 3, 4

Dash length is 3-9, can be confused with dots and spaces

Dilate inputs by 4x

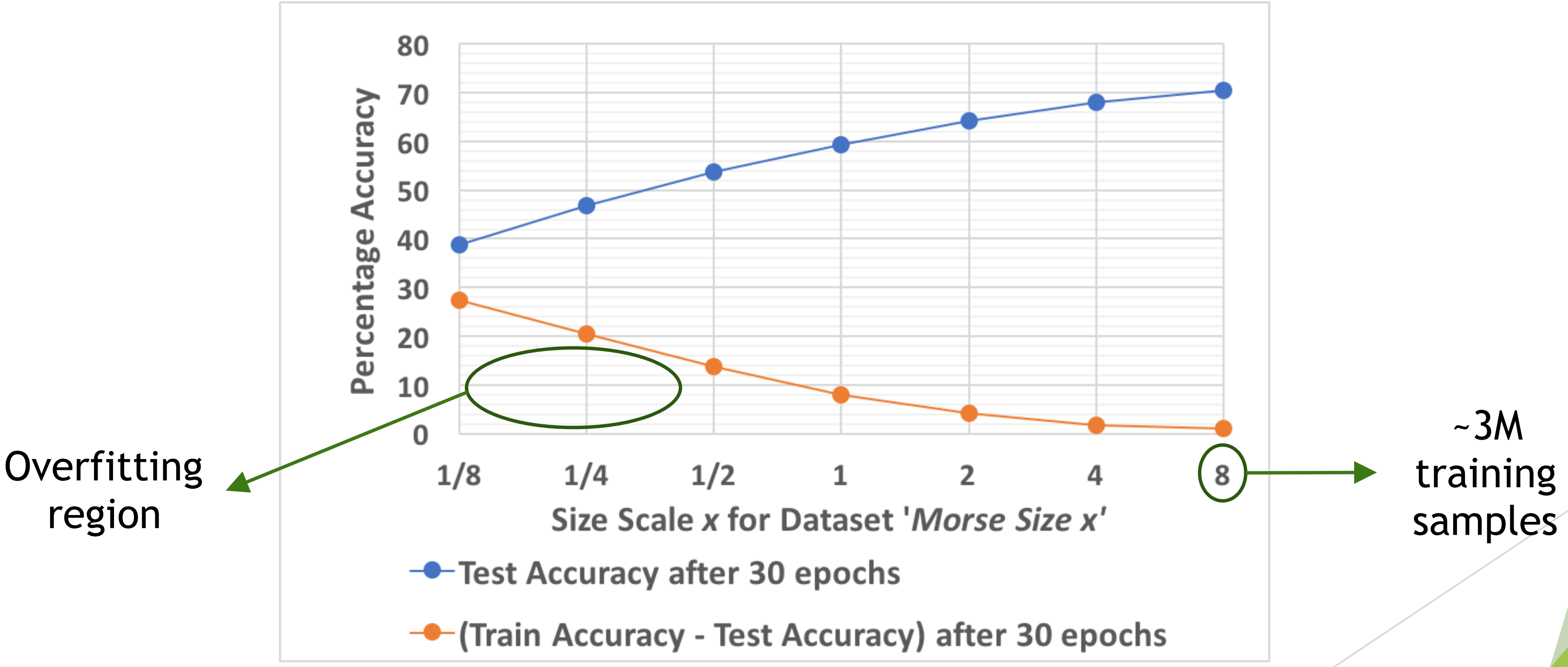
Property	Before Dilation	After Dilation
Frame length (= Number of inputs)	64	256
Space	1-3	4-12
Dot	1-3	4-12
Dash	3-9	12-36

Classification Accuracy on Test Data



Increasing Dataset Size

Unlimited amounts of data can be easily generated using computer algorithms



Dataset Evaluating Metrics

Difficult datasets have increased probability of classification errors

$$\sum_{m=1}^{M} P(m) \left[\max_{\substack{j \in \{1,2,\cdots,M\} \\ j \neq m}} P_{PW}(j|m) \right] \leq P(E)$$

$$\leq \sum_{m=1}^{M} P(m) \sum_{\substack{j=1 \\ j \neq m}}^{M} P_{PW}(j|m)$$

Dataset Evaluating Metrics

Difficult datasets have increased probability of classification errors

$$L = \sum_{m=1}^{M} P(m) Q\left(\sqrt{\frac{d_{min}(m)^2}{4{\sigma_m}^2}}\right) \leftarrow \sum_{m=1}^{M} P(m)\left[\max_{\substack{j\in\{1,2,\cdots,M\}\\ j\neq m}} P_{PW}\left(j|m\right)\right] \leq P(E)$$

$$\leq \sum_{m=1}^{M} P(m) \sum_{\substack{j=1\\ j\neq m}}^{M} P_{PW}\left(j|m\right)$$

Dataset Evaluating Metrics

Difficult datasets have increased probability of classification errors

$$L = \sum_{m=1}^{M} P(m) Q\left(\sqrt{\frac{d_{min}(m)^2}{4{\sigma_m}^2}}\right)$$

$$\sum_{m=1}^{M} P(m)\left[\max_{\substack{j \in \{1,2,\cdots,M\} \\ j \neq m}} P_{PW}\left(j|m\right)\right] \leq P(E)$$

$$\leq \sum_{m=1}^{M} P(m) \sum_{\substack{j=1 \\ j \neq m}}^{M} P_{PW}\left(j|m\right)$$

$$U = \sum_{m=1}^{M} P(m) \sum_{\substack{j=1 \\ j \neq m}}^{M} Q\left(\sqrt{\frac{d(m,j)^2}{4{\sigma_m}^2}}\right)$$

Dataset Evaluating Metrics

Difficult datasets have increased probability of classification errors

$$L = \sum_{m=1}^{M} P(m) Q\left(\sqrt{\frac{d_{min}(m)^2}{4{\sigma_m}^2}}\right)$$

$$D = \frac{\sum_{m=1}^{M} \frac{\sigma_m}{d_{min}(m)}}{M}$$

$$\sum_{m=1}^{M} P(m) \left[\max_{\substack{j\in\{1,2,\cdots,M\}\\ j\neq m}} P_{PW}(j|m)\right] \leq P(E)$$

$$\leq \sum_{m=1}^{M} P(m) \sum_{\substack{j=1\\ j\neq m}}^{M} P_{PW}(j|m)$$

$$U = \sum_{m=1}^{M} P(m) \sum_{\substack{j=1\\ j\neq m}}^{M} Q\left(\sqrt{\frac{d(m,j)^2}{4{\sigma_m}^2}}\right)$$

Dataset Evaluating Metrics

Difficult datasets have increased probability of classification errors

$$L = \sum_{m=1}^{M} P(m) Q\left(\sqrt{\frac{d_{min}(m)^2}{4{\sigma_m}^2}}\right)$$

$$\sum_{m=1}^{M} P(m) \left[\max_{\substack{j \in \{1,2,\cdots,M\} \\ j \neq m}} P_{PW}(j|m) \right] \leq P(E)$$

$$\leq \sum_{m=1}^{M} P(m) \sum_{\substack{j=1 \\ j \neq m}}^{M} P_{PW}(j|m)$$

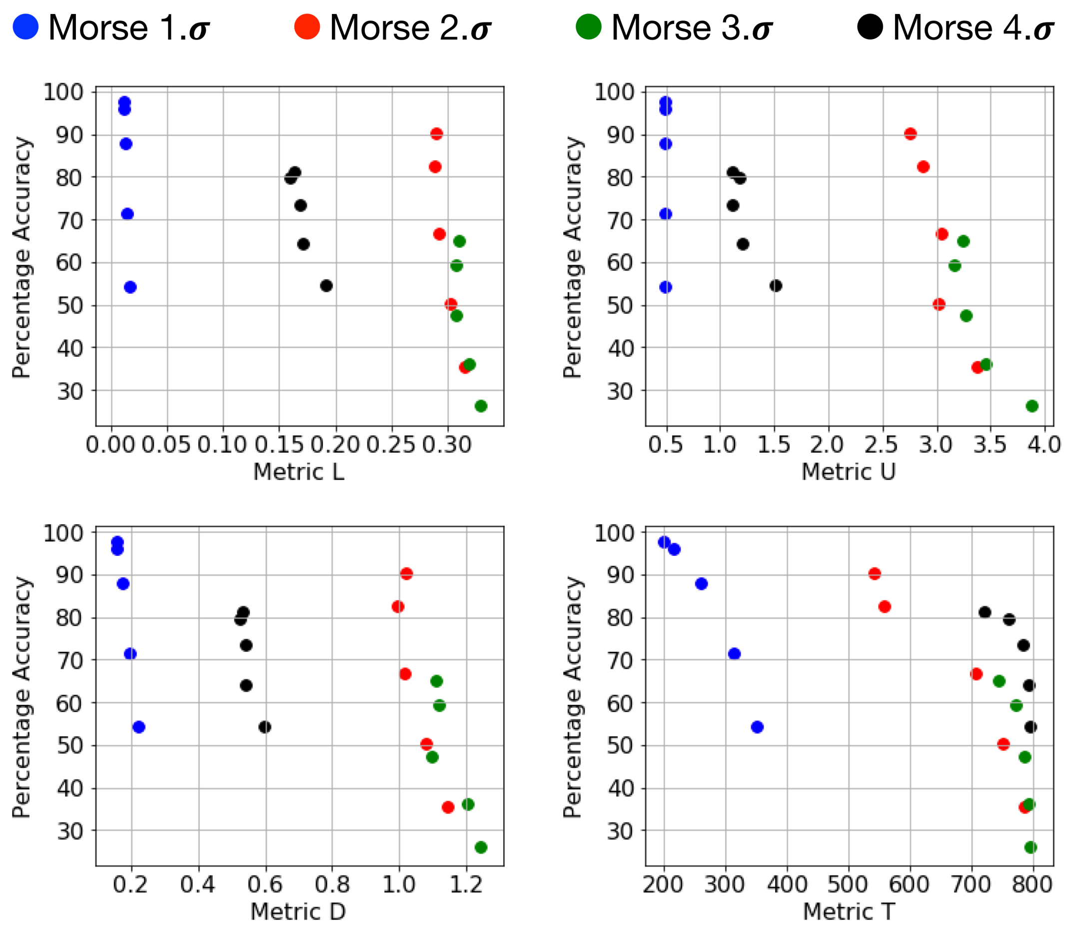
$$D = \frac{\sum_{m=1}^{M} \frac{\sigma_m}{d_{min}(m)}}{M}$$

$$U = \sum_{m=1}^{M} P(m) \sum_{\substack{j=1 \\ j \neq m}}^{M} Q\left(\sqrt{\frac{d(m,j)^2}{4{\sigma_m}^2}}\right)$$

$$T = \sum_{m=1}^{M} \sum_{\substack{j=1 \\ j \neq m}}^{M} \mathbb{I}\left(d_1(m,j) < 0.05\right)$$

Performance of the Metrics

Harder datasets have lower accuracy and higher metric values



Metric	-ρ
L	0.59
U	**0.64**
D	0.63
T	**0.64**

Conclusion

- Algorithm to generate machine learning datasets of tunable difficulty
- Synthetic data to solve challenges associated with natural data
- Metrics to evaluate dataset difficulty prior to training

Thank you!

Questions?

Contact: souryade@usc.edu