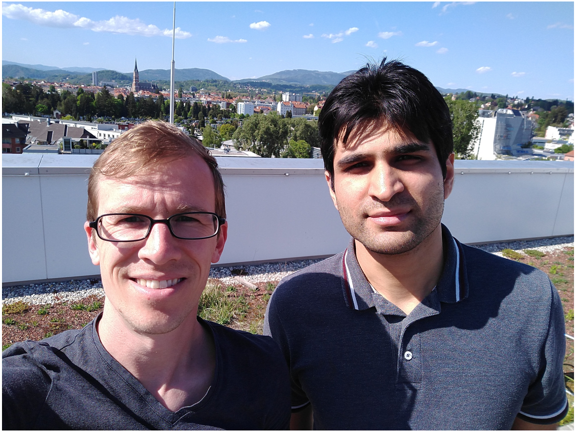
AUSTRIA'S LEADING RESEARCH CENTER
FOR DATA-DRIVEN BUSINESS AND BIG DATA ANALYTICS

Understanding Neural Networks with Information Theory

Bernhard C. Geiger
Joint Work with Rana Ali Amjad and Kairen Liu

Who are we?

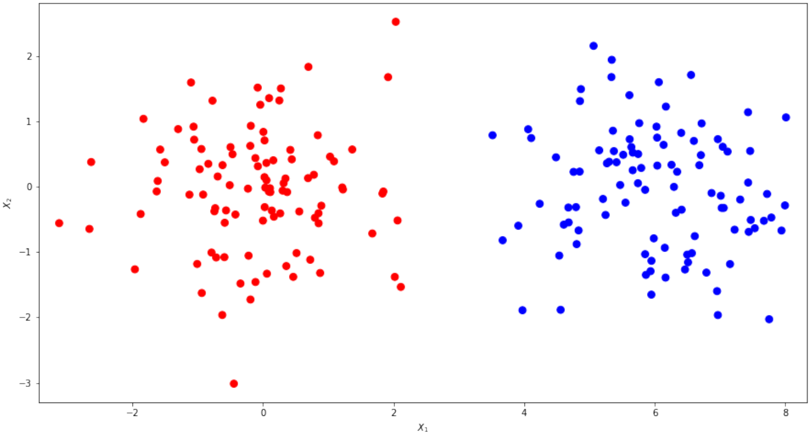


FWF
Der Wissenschaftsfonds.

Unterstützt von / Supported by
Alexander von Humboldt
Stiftung / Foundation

Overview

Binary Classification Task



Logistic Regression

- learn *class label* (red, blue) from *features* X_1 and X_2

Logistic Regression

- learn *class label* (red, blue) from *features* X_1 and X_2
- logistic regression is a linear model

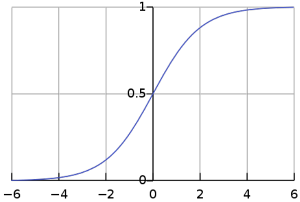
Logistic Regression

- learn *class label* (red, blue) from *features* X_1 and X_2
- logistic regression is a linear model
- logistic regression yields class probabilities:

  If $X_1 = x$ and $X_2 = x'$, then the probability that Y is red is p.

Logistic Regression (cont'd)

$$\mathbb{P}[Y = \text{red}] = \sigma(w_1 \cdot X_1 + w_2 \cdot X_2 + w_0)$$

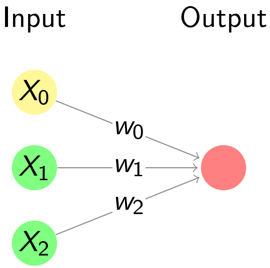


Public Domain by Qef.

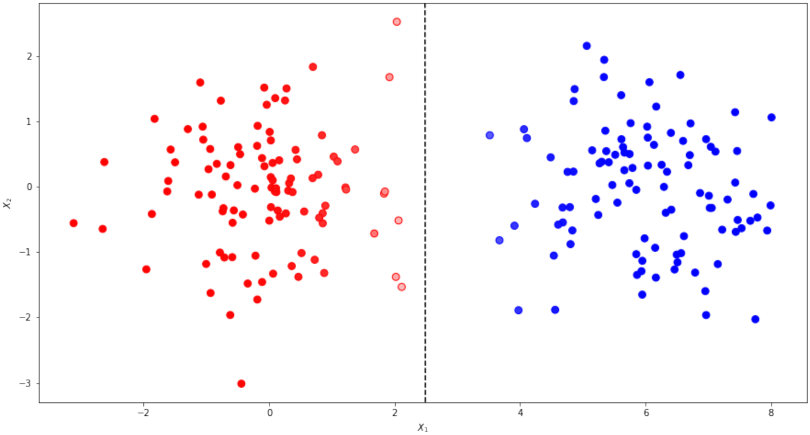
- $w_1 \cdot X_1 + w_2 \cdot X_2 + w_0 < 0$, then Y is more likely to be blue
- w_1, w_2, and w_0 define *decision boundary*
- **Task:** Learn w_1, w_2, and w_0 from data
- (typically: cross-entropy loss + L_2 regularization)

Logistic Regression (cont'd)

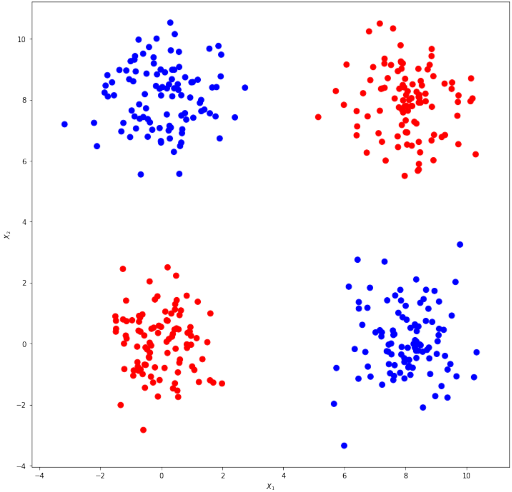
$$\mathbb{P}[Y = \text{red}] = \sigma(w_1 \cdot X_1 + w_2 \cdot X_2 + w_0)$$



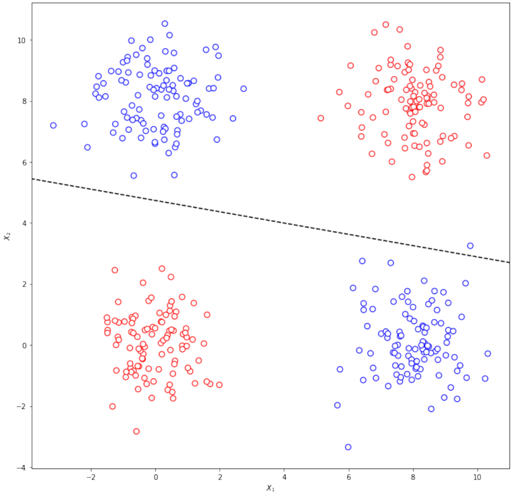
Binary Classification using Logistic Regression



Binary Classification (slightly more complicated)



Binary Classification (slightly more complicated)



Logistic Regression Fails. . .

. . . if the data is not linearly separable

Logistic Regression Fails. . .

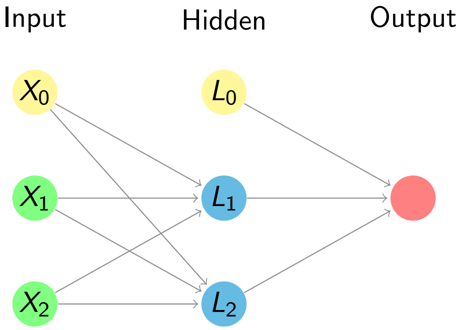
. . . if the data is not linearly separable

Idea: Stack multiple linear regression models on top of each other!

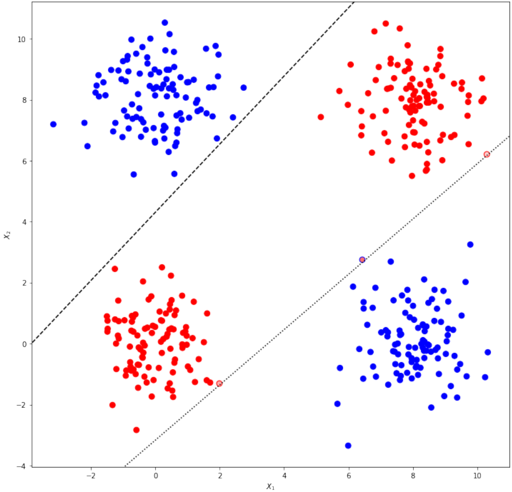
Logistic Regression Fails. . .

. . . if the data is not linearly separable

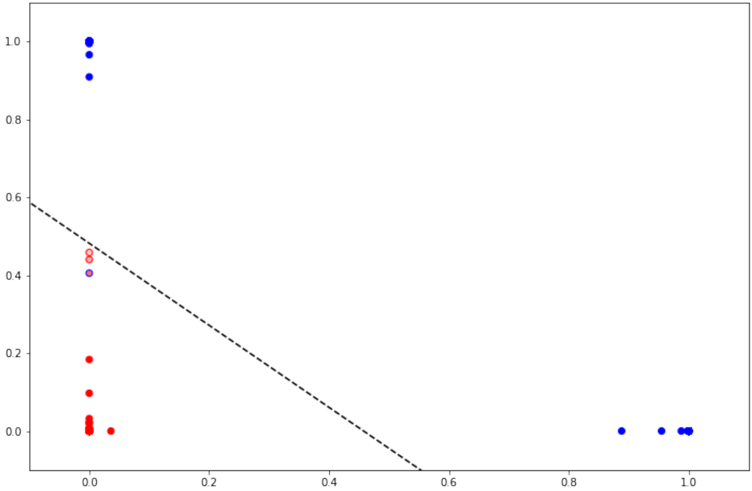
Idea: Stack multiple linear regression models on top of each other!



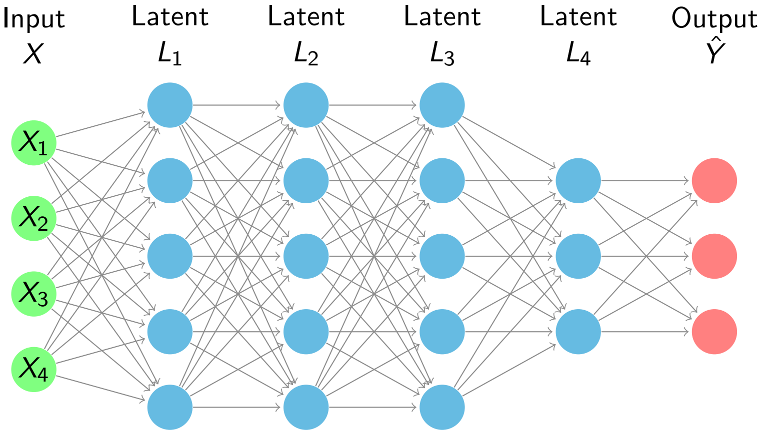
Binary Classification with a Neural Network



Binary Classification with a Neural Network



Binary Classification with Neural Networks



Binary Classification with Neural Networks

- Still easy to understand with two input features, hidden layers of width two (2D scatter plot)
- What happens for higher-dimensional input?
  - MNIST: input has 784 dimensions
  - CIFAR-10: input has 3×1024 dimensions
  - ...
- What happens for wider layers?
  - e.g., a $100-100$ MLP trained on MNIST?
  - ...

Two Approaches to Understand NNs

- Explainable/Interpretable AI:
  - What input features led to the decision?[1]
  - What training data was most influential for this decision?[2]
  - Simplified decision boundaries[3], extract decision procedure, etc.
  - ...
- How do NNs work internally?
  - Behavior during training
  - Why do NNs generalize so well?[4]
  - Importance of individual ("cat") neurons
  - ...

---

[1] Montavon, Samek, and Müller, "Methods for interpreting and understanding deep neural networks", 2018

[2] Koh and Liang, "Understanding Black-box Predictions via Influence Functions", 2017

[3] Ribeiro, Singh, and Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier", 2016

[4] Frankle and Carbin, "The Lottery Ticket Hypothesis: Training Pruned Neural Networks",

Two Approaches to Understand NNs

- Explainable/Interpretable AI:
  - What input features led to the decision?[1]
  - What training data was most influential for this decision?[2]
  - Simplified decision boundaries[3], extract decision procedure, etc.
  - ...
- How do NNs work internally?
  - Behavior during training
  - Why do NNs generalize so well?[4]
  - Importance of individual ("cat") neurons
  - ...

[1] Montavon, Samek, and Müller, "Methods for interpreting and understanding deep neural networks", 2018

[2] Koh and Liang, "Understanding Black-box Predictions via Influence Functions", 2017

[3] Ribeiro, Singh, and Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier", 2016

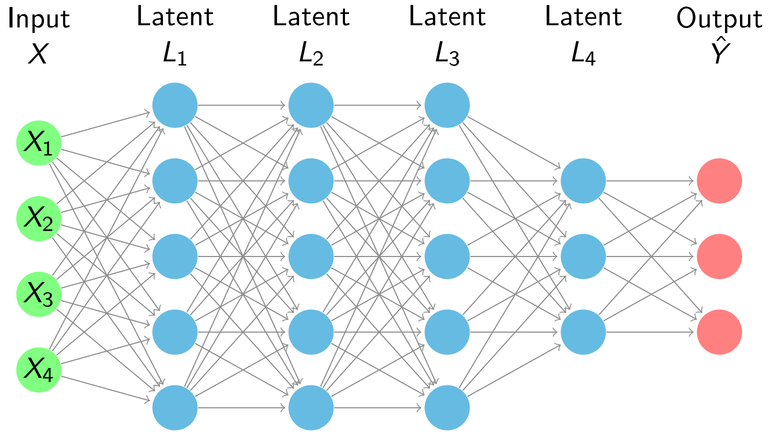
[4] Frankle and Carbin, "The Lottery Ticket Hypothesis: Training Pruned Neural Networks",

Prerequisite: Mutual Information

$$I(U;V)$$

- is defined for general random variables
- measures statistical dependence between U and V
- generalizes (linear) correlation
- is zero if and only if U and V are independent
- is invariant under invertible maps
- (can be difficult to estimate)

Information Plane Analyses



Information Plane Analyses (cont'd)

Intermediate representation L (NN layer) should

P1 contain sufficient info for classification

- e.g., L should suffice to determine whether X is a cat or a dog

P2 ...but not more info than necessary (compression)

- e.g., L should not contain information about the color of the fur, length of ears, etc.

[5] Alemi et al., "Deep Variational Information Bottleneck", 2017

[6] Kolchinsky, Tracey, and Wolpert, "Nonlinear Information Bottleneck", 2019

[7] Fischer, "The Conditional Entropy Bottleneck", 2020

Information Plane Analyses (cont'd)

Intermediate representation L (NN layer) should

P1 contain sufficient info for classification

- e.g., L should suffice to determine whether X is a cat or a dog

P2 ...but not more info than necessary (compression)

- e.g., L should not contain information about the color of the fur, length of ears, etc.

$$\text{P1} \Leftrightarrow \text{large } I(Y;L)$$
$$\text{P2} \Leftrightarrow \text{small } I(X;L)$$

[5] Alemi et al., "Deep Variational Information Bottleneck", 2017

[6] Kolchinsky, Tracey, and Wolpert, "Nonlinear Information Bottleneck", 2019

[7] Fischer, "The Conditional Entropy Bottleneck", 2020

Information Plane Analyses (cont'd)

Intermediate representation L (NN layer) should

P1 contain sufficient info for classification

- e.g., L should suffice to determine whether X is a cat or a dog

P2 ...but not more info than necessary (compression)

- e.g., L should not contain information about the color of the fur, length of ears, etc.

$$\text{P1} \Leftrightarrow \text{large } I(Y;L)$$
$$\text{P2} \Leftrightarrow \text{small } I(X;L)$$

Idea has been successfully applied in NN training[5,6,7]

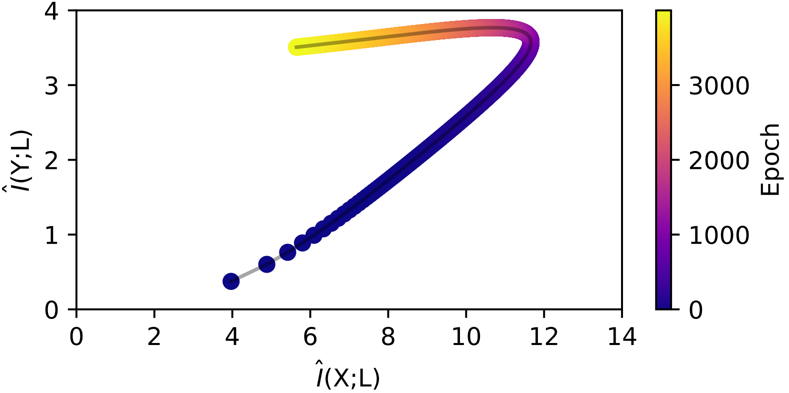
[5] Alemi et al., "Deep Variational Information Bottleneck", 2017

[6] Kolchinsky, Tracey, and Wolpert, "Nonlinear Information Bottleneck", 2019

[7] Fischer, "The Conditional Entropy Bottleneck", 2020

Information Plane Analyses (cont'd)

Estimate how $I(X;L)$ and $I(Y;L)$ evolve during NN training[8]:



[8]Shwartz-Ziv and Tishby, *Opening the Black Box of Deep Neural Networks via Information*, 2017

Information Plane Analyses (cont'd)

Hot Topic, but many open questions:

- requires estimating mutual information, which is problematic[9]
- connection to generalization not fully clear, e.g.[10]
- information plane appears to show geometric picture (clustering)[11]
- current results in the literature are inconsistent (is there a compression phase?, etc.)[12]
- ongoing debate

[9] Amjad and Geiger, "Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle", 2020

[10] Saxe et al., "On the Information Bottleneck Theory of Deep Learning", 2018

[11] Goldfeld et al., "Estimating Information Flow in Deep Neural Networks", 2019

[12] Geiger, *On Information Plane Analyses of Neural Network Classifiers – A Review*, 2020

Bounds on Generalization Gap

i.e., difference between expected and estimated loss as a function of size m of dataset $\mathcal{D} = \{D_1, \dots, D_m\}$

[13] Vera, Piantanida, and Vega, “The Role of the Information Bottleneck in Representation Learning”, 2018

[14] Shwartz-Ziv, Painsky, and Tishby, *Representation Compression and Generalization in Deep Neural Networks*, 2018

[15] Xu and Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms”, 2017

[16] Bu, Zou, and Veeravalli, “Tightening Mutual Information Based Bounds on Generalization Error”, 2019

[17] Pensia, Jog, and Loh, “Generalization Error Bounds for Noisy, Iterative Algorithms”, 2018

[18] Achille and Soatto, “Emergence of Invariance and Disentanglement in Deep Representations”, 2018

Bounds on Generalization Gap

i.e., difference between expected and estimated loss as a function of size m of dataset $\mathcal{D} = \{D_1, \ldots, D_m\}$

- $\propto \sqrt{I(X;L)}\frac{\log m}{\sqrt{m}}$, see[13]
- $\left(2^{I(X;L)} + \log(2/\delta)\right)/(2m)$ with probability $1-\delta$, see[14]

---

[13] Vera, Piantanida, and Vega, "The Role of the Information Bottleneck in Representation Learning", 2018

[14] Shwartz-Ziv, Painsky, and Tishby, *Representation Compression and Generalization in Deep Neural Networks*, 2018

[15] Xu and Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms", 2017

[16] Bu, Zou, and Veeravalli, "Tightening Mutual Information Based Bounds on Generalization Error", 2019

[17] Pensia, Jog, and Loh, "Generalization Error Bounds for Noisy, Iterative Algorithms", 2018

[18] Achille and Soatto, "Emergence of Invariance and Disentanglement in Deep Representations", 2018

Bounds on Generalization Gap

i.e., difference between expected and estimated loss as a function of size m of dataset $\mathcal{D} = \{D_1, \dots, D_m\}$

- $\propto \sqrt{I(X;L)}\frac{\log m}{\sqrt{m}}$, see[13]
- $\left(2^{I(X;L)} + \log(2/\delta)\right)/(2m)$ with probability $1-\delta$, see[14]
- $\propto \sqrt{\frac{1}{m} I(\mathcal{D}; A(\mathcal{D}))}$, see[15]
- $\propto \frac{1}{m}\sum_{i=1}^{m}\sqrt{I(D_i; A(\mathcal{D}))}$, see[16]
- extensions to SGD-type training[17]

---

[13] Vera, Piantanida, and Vega, "The Role of the Information Bottleneck in Representation Learning", 2018

[14] Shwartz-Ziv, Painsky, and Tishby, *Representation Compression and Generalization in Deep Neural Networks*, 2018

[15] Xu and Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms", 2017

[16] Bu, Zou, and Veeravalli, "Tightening Mutual Information Based Bounds on Generalization Error", 2019

[17] Pensia, Jog, and Loh, "Generalization Error Bounds for Noisy, Iterative Algorithms", 2018

[18] Achille and Soatto, "Emergence of Invariance and Disentanglement in Deep Representations", 2018

Bounds on Generalization Gap

i.e., difference between expected and estimated loss as a function of size m of dataset $\mathcal{D} = \{D_1, \dots, D_m\}$

- $\propto \sqrt{I(X;L)}\frac{\log m}{\sqrt{m}}$, see[13]
- $\left(2^{I(X;L)} + \log(2/\delta)\right)/(2m)$ with probability $1-\delta$, see[14]
- $\propto \sqrt{\frac{1}{m} I(\mathcal{D}; A(\mathcal{D}))}$, see[15]
- $\propto \frac{1}{m}\sum_{i=1}^{m} \sqrt{I(D_i; A(\mathcal{D}))}$, see[16]
- extensions to SGD-type training[17]
- see also[18]

[13] Vera, Piantanida, and Vega, "The Role of the Information Bottleneck in Representation Learning", 2018

[14] Shwartz-Ziv, Painsky, and Tishby, *Representation Compression and Generalization in Deep Neural Networks*, 2018

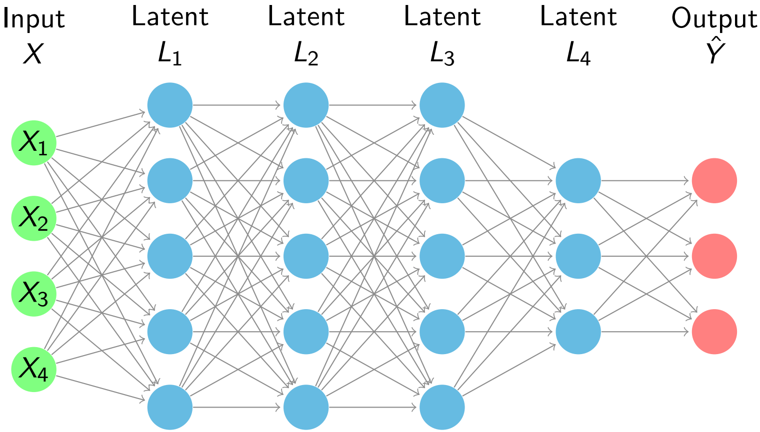
[15] Xu and Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms", 2017

[16] Bu, Zou, and Veeravalli, "Tightening Mutual Information Based Bounds on Generalization Error", 2019

[17] Pensia, Jog, and Loh, "Generalization Error Bounds for Noisy, Iterative Algorithms", 2018

[18] Achille and Soatto, "Emergence of Invariance and Disentanglement in Deep Representations", 2018

What about Individual Neurons?



What about Individual Neurons? (cont'd)

How important is the ℓ-th neuron in the i-th layer?

What about Individual Neurons? (cont'd)

How important is the ℓ-th neuron in the i-th layer?

- compute mutual information $I(Y; L_{i,\ell})$
- much easier to estimate than $I(Y; L_i)$ (whole layer) or $I(X; L_i)$ (X is high-dimensional/continuously distributed)
- **Hypothesis**: Large values indicate that the ℓ-th neuron in the i-th layer is important for the task

Information-Ordered Cumulative Ablation[19]

- **Ablation**: Turning off individual neurons, i.e., set $L_{i,\ell} = 0$

[19] Liu, Amjad, and Geiger, *Understanding Individual Neuron Importance Using Information Theory*, 2018

Information-Ordered Cumulative Ablation[19]

- **Ablation**: Turning off individual neurons, i.e., set $L_{i,\ell} = 0$
- **Cumulative** Ablation: Turn off more and more neurons and see how, e.g., classification accuracy is affected

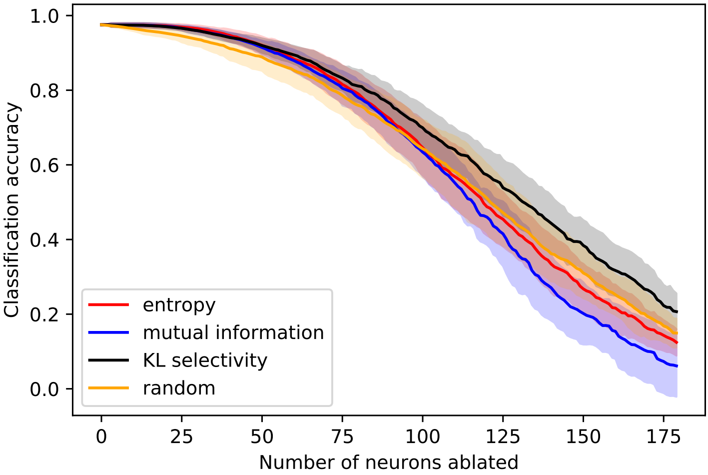
[19] Liu, Amjad, and Geiger, *Understanding Individual Neuron Importance Using Information Theory*, 2018

Information-Ordered Cumulative Ablation[19]

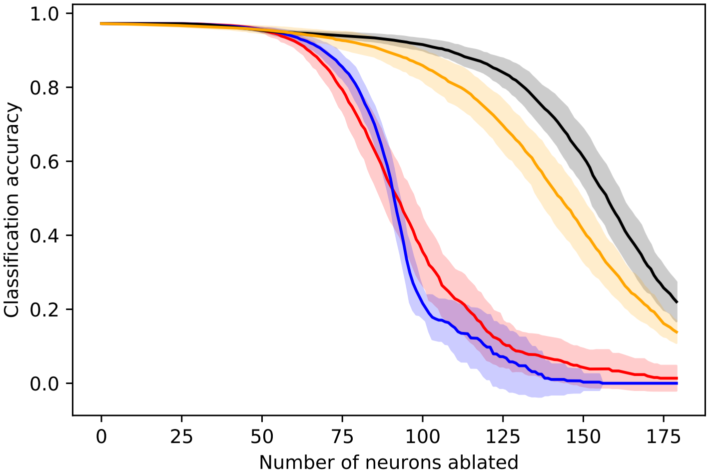
- **Ablation**: Turning off individual neurons, i.e., set $L_{i,\ell} = 0$
- **Cumulative** Ablation: Turn off more and more neurons and see how, e.g., classification accuracy is affected
- **Information-Ordering**: Turn off the k neurons with lowest (highest) mutual information and compare with turning off neurons randomly

[19] Liu, Amjad, and Geiger, *Understanding Individual Neuron Importance Using Information Theory*, 2018

MNIST $100 - 100$, L_2 regularization



MNIST $100 - 100$, Dropout



What about Individual Neurons? (cont'd)

How important is the ℓ-th neuron in the i-th layer?

- it seems as if neurons with *high* mutual information are *not useful/hurting* classification performance
- reproduces results from[20]

[20] Morcos et al., *On the importance of single directions for generalization*, 2018

What about Individual Neurons? (cont'd)

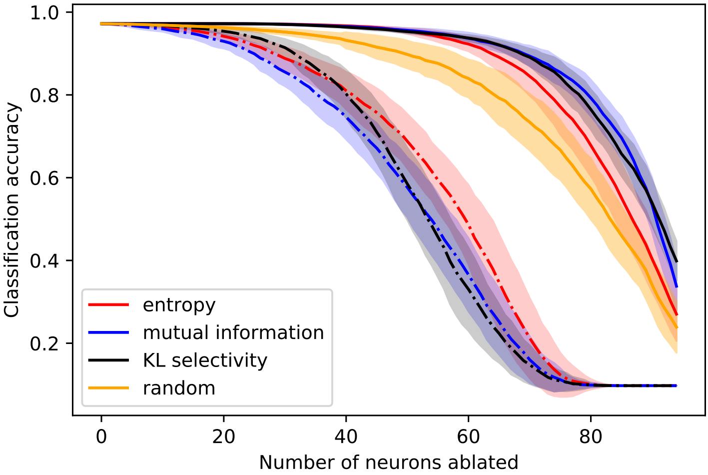
How important is the ℓ-th neuron in the i-th layer?

- it seems as if neurons with *high* mutual information are *not useful/hurting* classification performance
- reproduces results from[20]

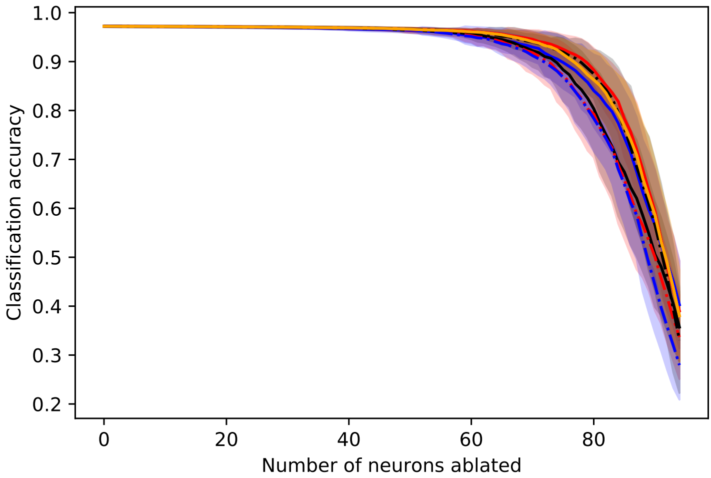
Let's take a closer look!

[20] Morcos et al., *On the importance of single directions for generalization*, 2018

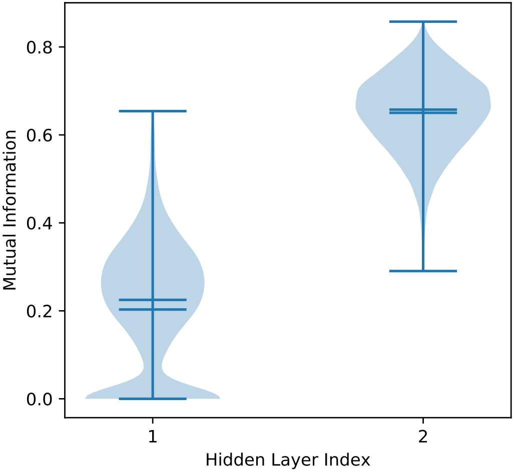
MNIST $100 - 100$, Dropout, Layer 1



MNIST $100 - 100$, Dropout, Layer 2



MNIST $100 - 100$, Dropout



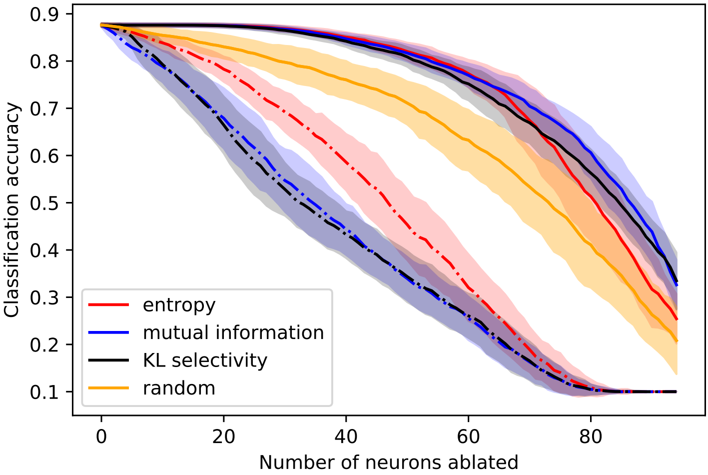
What about Individual Neurons? (cont'd)

How important is the ℓ-th neuron in the i-th layer?

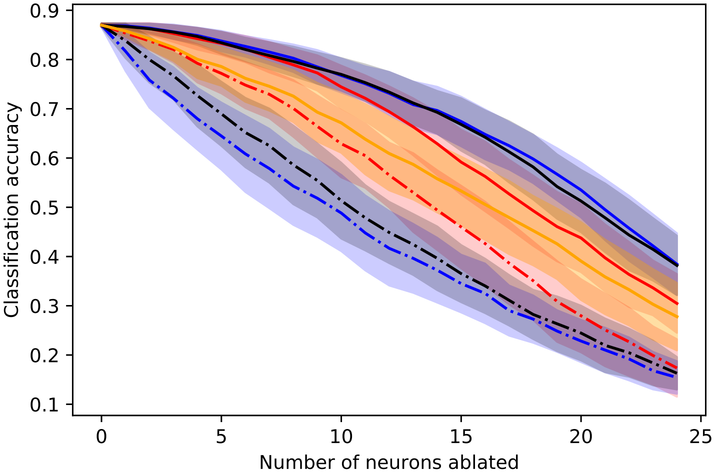
- it seems as if neurons with *high* mutual information are *not useful/hurting* classification performance[21]
- **BUT:** neurons with *high* mutual information are *useful* within a given layer
- layers have different distribution of mutual information values
- $\Rightarrow$ Simpson's paradox

[21] Morcos et al., *On the importance of single directions for generalization*, 2018

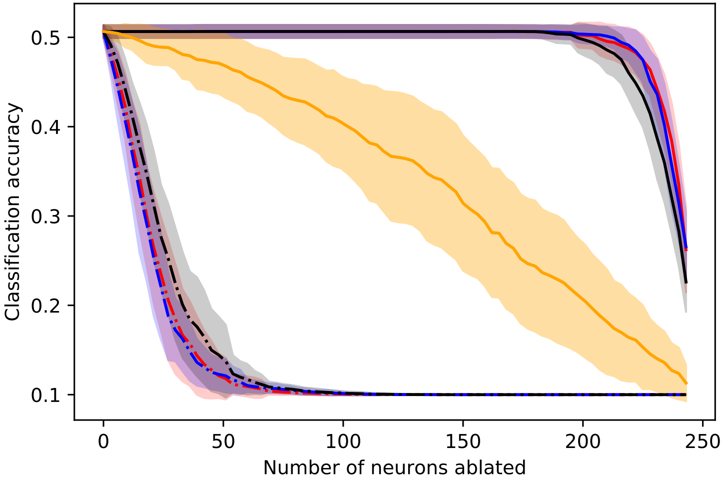
FashionMNIST $100 - 100$, L_2, Layer 1



FashionMNIST $30-30$, L_2, Layer 1



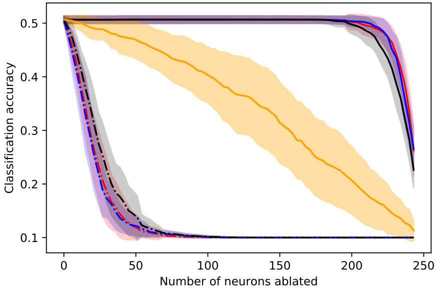
CIFAR-10 $250 - 500 - 250 - 500$, L_2, **Layer 3**



Information-Ordered Cumulative Ablation

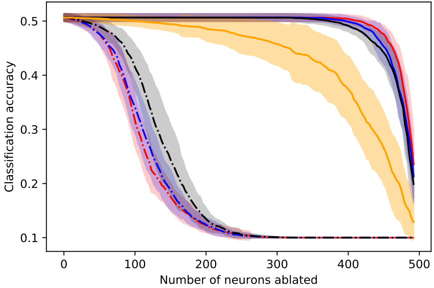
What else can we learn?

CIFAR-10 $250 - 500 - 250 - 500$, L_2, Layer 2



- 40 neurons with highest mutual information suffice
- removing 60 neurons with highest mutual information destroy performance
- $\approx$ 200 neurons are **inactive**

CIFAR-10 $250-500-250-500$, L_2, Layer 4



- 100 neurons with highest mutual information suffice
- removing 250 neurons with highest mutual information destroy performance
- $\approx$ 250 neurons are inactive
- $\approx$ 50-150 neurons are **redundant**

More Insights?

- beyond mutual information
- beyond ReLU activation functions
- beyond L_2 regularization
- effects of quantization
- ...

arXiv:1804.06679v3 [cs.LG]

Conclusion

NNs are difficult to understand, but

information theory is powerful:

- Bounds on the generalization error
- Investigating learning behavior
- Interplay between learning and geometric compression
- Importance of individual neurons via ordered cumulative ablation
  - neurons with large mutual information (**within a layer**) are important for classification
  - mutual information values differ between layers
  - cumulative ablation reveals inactive, redundant, and synergistic neurons

Conclusion

NNs are difficult to understand, but

information theory is powerful:

- Bounds on the generalization error
- Investigating learning behavior
- Interplay between learning and geometric compression
- Importance of individual neurons via ordered cumulative ablation
  - neurons with large mutual information (**within a layer**) are important for classification
  - mutual information values differ between layers
  - cumulative ablation reveals inactive, redundant, and synergistic neurons

Thanks for your attention!