W207– Applied Machine Learning

Cornelia Ilin, PhD

School of Information
UC Berkeley

Feedforward NN

Last week

- Evaluation metrics for classification tasks
- Dealing with class imbalance
- Multi-class Logistic Regression (extend the wine dataset to 3 classes)

Last week

- Evaluation metrics for classification tasks

- Dealing with class imbalance

- Multi-class Logistic Regression (extend the wine dataset to 3 classes)

Input: Today is a great day.

Translation (in Romanian): Astazi e o zi frumoasa.

How do you
evaluate model performance?

Last week

- Evaluation metrics for classification tasks
- Dealing with class imbalance
- Multi-class Logistic Regression (extend the wine dataset to 3 classes)

Input: Today is a great day.

Translation (in Romanian): Astazi e o zi frumoasa.

Translation (in Romanian): Este o zi frumoasa astazi.

How do you
evaluate model performance?

Today’s learning objectives

- General concepts: Feedforward Neural Networks
- Training, validation, and test datasets
- Application: Detect Diabetic Retinopathy using image data

Neural Networks

Shallow networks (no hidden layers) can only capture linear decision boundaries

Neural Networks

Shallow networks (no hidden layers) can only capture linear decision boundaries

What if CT scans data?

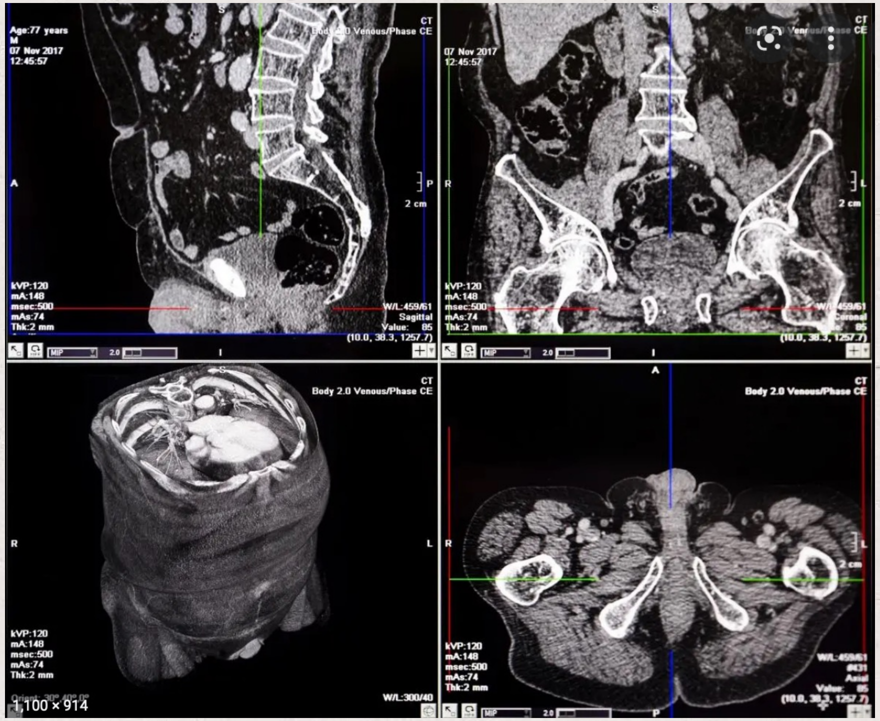


Image source: https://www.medicalnewstoday.com/articles/153201#what-is-a-CT-scan

Neural Networks

Shallow networks (no hidden layers) can only capture linear decision boundaries

What if video data?

Image source: https://web.stanford.edu/class/biods220/lectures/lecture5.pdf

Neural Networks

Shallow networks (no hidden layers) can only capture linear decision boundaries

What if audio data? (e.g., speech recognition)

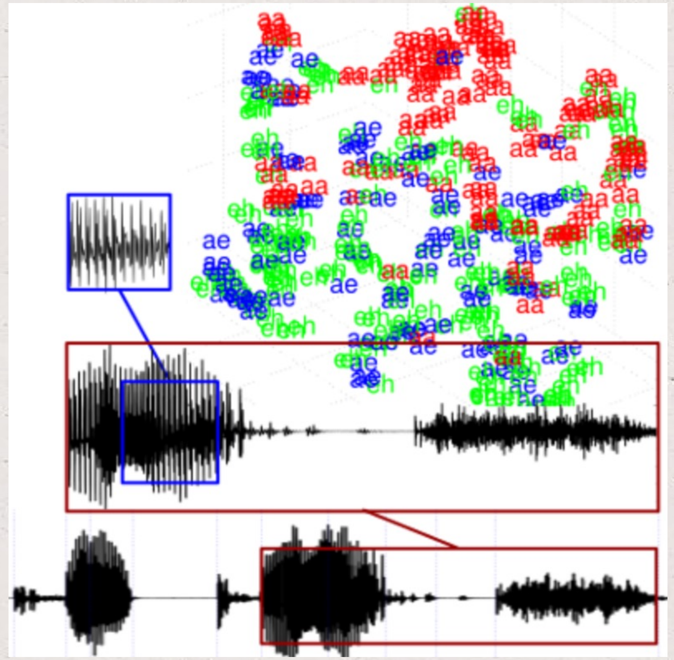


Image source: https://cbmm.mit.edu/research/projects-thrust/theoretical-frameworks-intelligence/invariant-representation-learning

Neural Networks

Shallow networks (no hidden layers) can only capture linear decision boundaries

What if CT scans data?

What if video data?

What if audio data? (e.g., speech recognition)

unstructured data

Neural Networks

Shallow networks (no hidden layers) can only capture linear decision boundaries

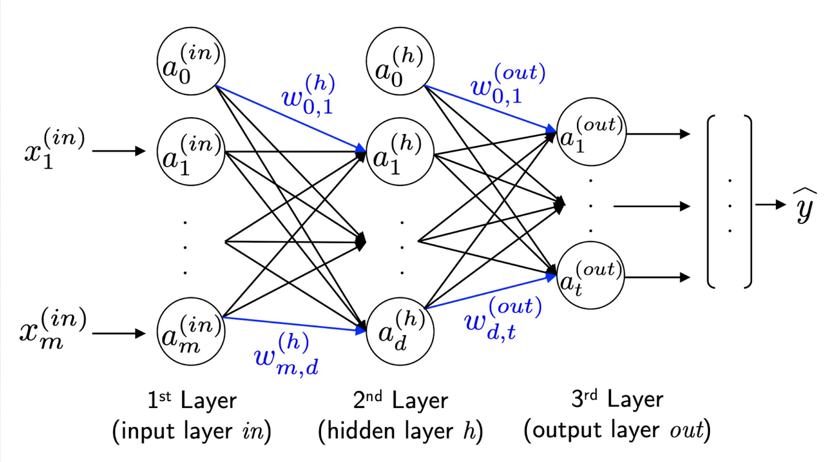
What if CT scans data?

What if video data?

What if audio data? (e.g., speech recognition)

unstructured data

Deep NN



a = activation function

Neural Networks

ML real world problems: we don’t know how large the network should be a priori!

Small network

underfit

Network cannot learn the underlying structure of complex datasets

Very large network

overfit

Network memorizes training data

Neural Networks

ML real world problems: we don't know how large the network should be a priori!

Small network

underfit

Network cannot learn the underlying structure of complex datasets

Solution: build a network with a relatively higher capacity (slightly higher than necessary) to do well on training

Very large network

overfit

Network memorizes training data

Neural Networks

ML real world problems: we don’t know how large the network should be a priori!

Small network

underfit

Network cannot learn the underlying structure of complex datasets

Solution: build a network with a relatively higher capacity (slightly higher than necessary) to do well on training

Prevent overfitting by applying one or more regularization schemes

dropout in NN

L_1, L_2 for others

Very large network

overfit

Network memorizes training data

Training, validation, and test datasets

All training = 80%

Training (50%)	Validation (30%)	Test (20%)

Held-out evaluation set for selecting best hyperparameters during training

Use only for final evaluation

Other splits: 60/20/20 is also popular.

Training, validation, and test datasets

All training (80%)	Test (20%)

Done with hyperparameter selection using the validation set?

Common to merge training and validation sets to train a final model using chosen hyperparameters.

Use only for final evaluation

K-fold cross validation

All training (80%)	Test (20%)

Small labeled dataset? Very common in healthcare… K-fold cross validation (can be computationally expensive!) may be worthwhile.

K-fold cross validation

All training (80%)	Test (20%)

use different fold for validation

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Test (20%)

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Test (20%)

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Test (20%)

⋮

Small labeled dataset? Very common in healthcare… K-fold cross validation may be worthwhile.

K-fold cross validation

All training (80%)	Test (20%)

use different fold for validation

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Test (20%)

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Test (20%)

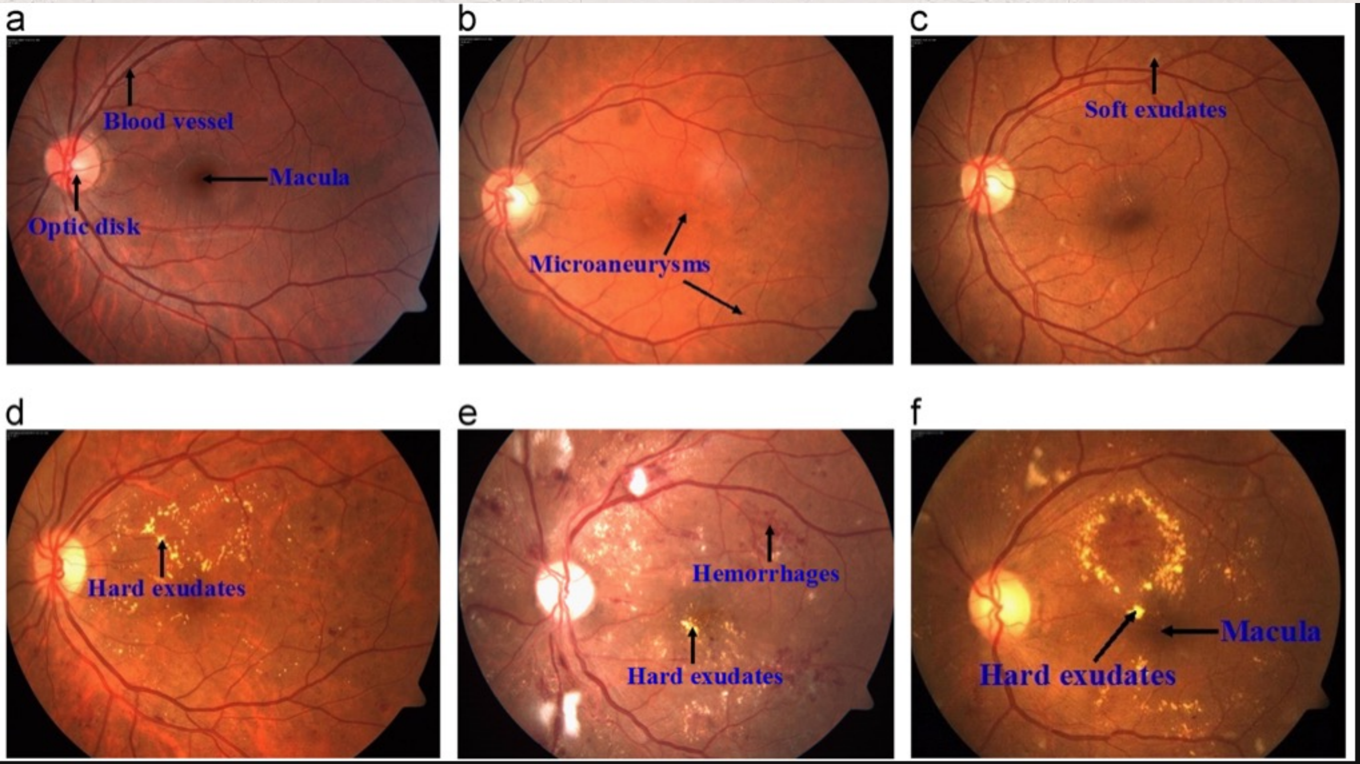
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Test (20%)

- Train model K times with a different fold as the validation set each time; then average the validation set results.

Smart approach: allows more data to be used for each training of the model, without compromising validation results

OK to apply same concept to test-time evaluation.

Application: detect diabetic retinopathy



Typical fundus images: (a) Normal; (b) Mild DR; (c) Moderate DR; (d) Severe DR; (e) Prolific DR; (f) Macular edema.

Image source: https://www.sciencedirect.com/science/article/pii/S0010482513002862#f0005