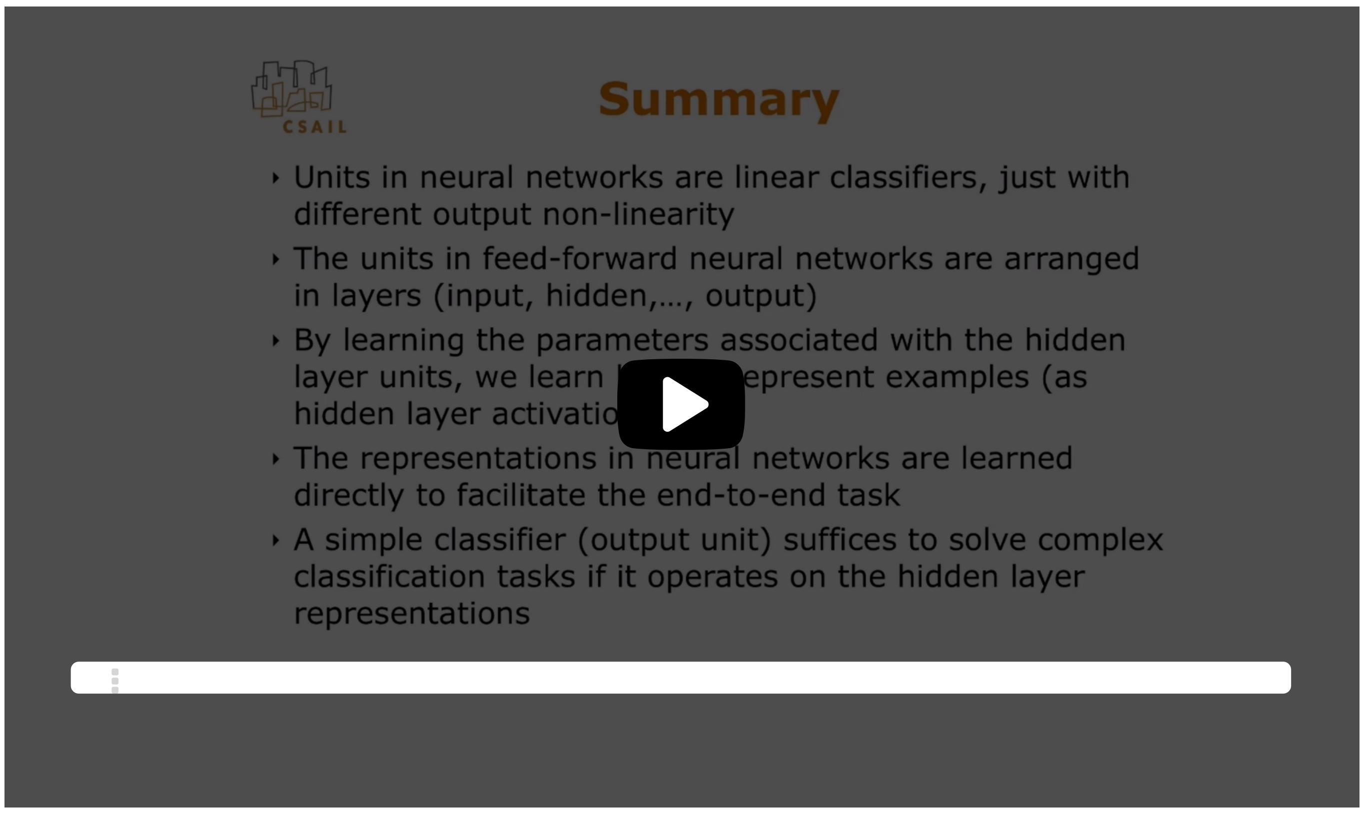
6. Hidden Layer Models

Models with Hidden Layer



18:01 / 18:01 1.0x

is their prediction that we actually want.

And the role of the hidden layers

is really to adjust their transformation,

adjust their computation in such a way

that the output layer will have an easier task

to solve the problem.

The problem now becomes actually learning these representations

together with the final classifier.

And we will talk about that in the next lecture.

End of transcript. Skip to the start.

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

For the following set of problems, let's consider a simple 2-dimensional classification task. The training set is made up of 4 points listed below:

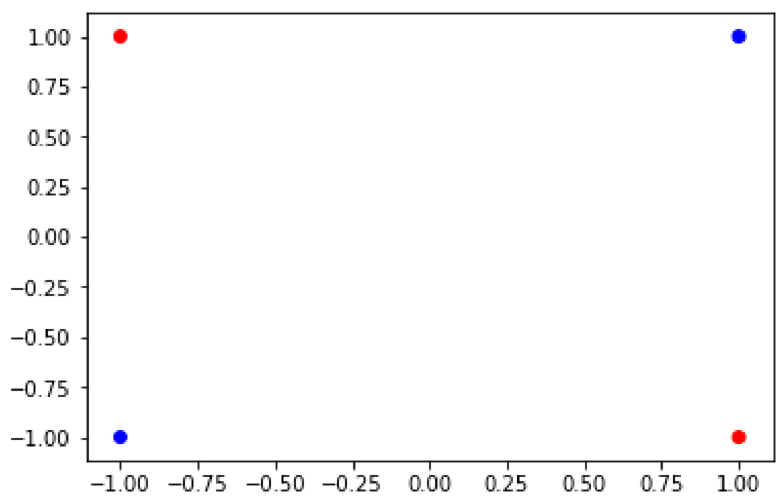
$$x^{(1)} = (-1,-1) \quad , \quad y^{(1)} = 1$$

$$x^{(2)} = (1,-1) \quad , \quad y^{(2)} = -1$$

$$x^{(3)} = (-1,1) \quad , \quad y^{(3)} = -1$$

$$x^{(4)} = (1,1) \quad , \quad y^{(4)} = 1$$

The dataset is illustrated below (blue - positive, red - negative)

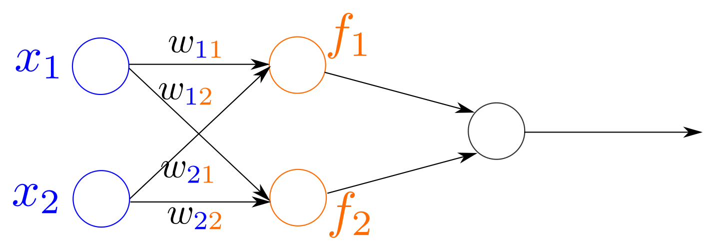


For simplicity, assume that we are only interested in binary classification problems for now. That is, $y^{(i)}$ can be either 1 or -1.

Linear Separability After First Layer

1/1 point (graded)

For this problem, let us focus on a network with one hidden layer and two units in that layer:



Let $f_1^{(i)}, f_2^{(i)}$ denote the output of the two units in the hidden layer corresponding to the input $x^{(i)}$ respectively, i.e.

$$\begin{aligned} f_1^{(i)} &= f\left(w_{01} + \left(w_{11}x_1^{(i)} + w_{21}x_2^{(i)}\right)\right) \\ f_2^{(i)} &= f\left(w_{02} + \left(w_{12}x_1^{(i)} + w_{22}x_2^{(i)}\right)\right) \end{aligned}$$

Consider the set $D' = \left\{ \left(\left[f_1^{(i)}, f_2^{(i)} \right], y^{(i)} \right), \quad i = 1,2,3,4 \right\}$

Assume that f is the linear activation function given by $f(z) = 2z - 3$.

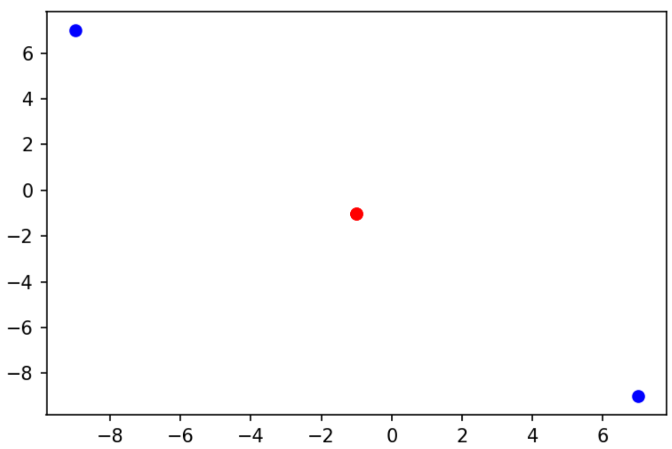
For which of the following values of weights would the set D' be linearly separable? (Select all that apply.)

- [ ] $w_{11} = w_{21} = 0, w_{12} = w_{22} = 0, w_{01} = w_{02} = 0$
- [ ] $w_{11} = w_{21} = 2, w_{12} = w_{22} = -2, w_{01} = w_{02} = 1$
- [ ] $w_{11} = w_{21} = -2, w_{12} = w_{22} = 2, w_{01} = w_{02} = 1$
- [x] None of the above ✔

✔

Solution:

First of all note that from the figure in the text above that D is clearly not linearly separable.
Also, $f(z) = 2z - 3$ is a linear activation function.
Any linear transformation of the feature space of a linearly in-separable classification problem would still continue to remain linearly inseparable. For this question, one can compute the feature representations of all the data points and verify visually.
For example, the result of the second answer is plotted here:



Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

Non-linear Activation Functions

1/1 point (graded)

Again, let's focus on a network with one hidden layer with two units and use the same training set as above. The weights of the network are given as follows:

$$w_{11} = 1, w_{21} = -1, w_{01} = 1$$
$$w_{12} = -1, w_{22} = 1, w_{02} = 1$$

Let f_1, f_2 be the outputs of the first and second unit respectively.

Consider the set $D' = \{([f_1^{(i)}, f_2^{(i)}], y^{(i)}), \quad i = 1,2,3,4\}$

For which of the following functions f, would the set D' be linearly separable? (Select one or more that apply.)

- [ ] $f(z) = 5z - 2$
- [x] $f(z) = \text{ReLU}(z)$ ✔
- [x] $f(z) = \tanh(z)$ ✔
- [ ] $f(z) = z$

✔

Solution:

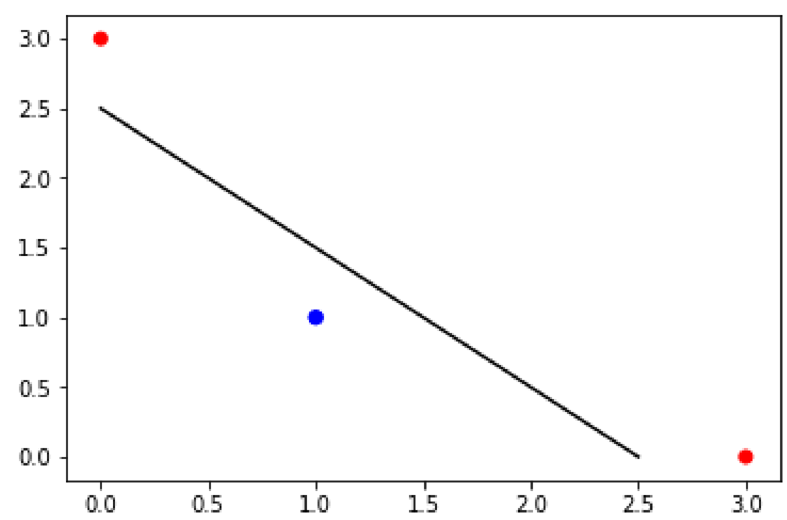
From the above problem, we note that any linear transformation of the feature space of a linearly in-separable classification problem would still continue to remain linearly inseparable. Hence we rule out the two linear functions.
For all of the parts below, note that

$$\begin{aligned} f_1^{(i)} &= f(w_{01} + (w_{11}x_1^{(i)} + w_{21}x_2^{(i)})) \\ f_2^{(i)} &= f(w_{02} + (w_{12}x_1^{(i)} + w_{22}x_2^{(i)})) \end{aligned}$$

- $f(z) = \text{ReLU}(z)$: Substituting for ReLU into f in the above equation gives the following results:

$$(1,1), (3,0), (0,3), (1,1)$$

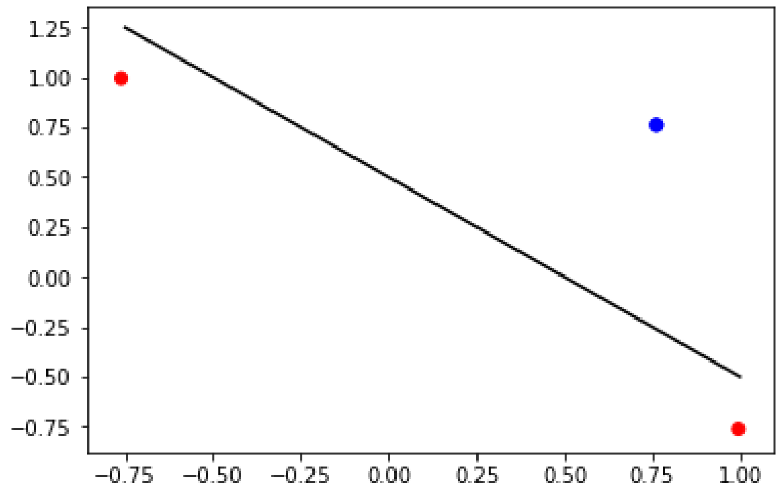
The following figure plots these points and a potential linear classifier:



- $f(x) = \tanh(x)$: Substituting for $tanh$ into f in the above equation gives the following results:

$$(0.76, 0.76), (0.99, -0.76), (-0.76, 0.99), (0.76, 0.76)$$

The following figure plots these points and a potential linear classifier:



Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Neural Network Learned parameters

1/1 point (graded)
Given a neural network with one hidden layer for classification, we can view the hidden layer as a feature representation, and the output layer as a classifier using the learned feature representation.

There're also other parameters that will affect the learning process and the performance of the model, such as the learning rate and parameters that control the network architecture (e.g. number of hidden units/layers) etc. These are often called hyper-paramters.

Which of the following is/are optimized during the training process? Check all that apply.

- [ ] The dimension of the feature representation
- [x] The weights that control the feature representation ✔
- [ ] The hyper-parameters
- [x] The weights for the classifier ✔

✔

Solution:

Similar to the linear classifiers that we covered in previous lectures, we need to learn the parameters for the classifier. However, in this case we also learn the parameters that generate a representation for the data.
The dimensions and the hyper-parameters are decided with the structure of the model and are not optimized directly during the learning process but can be chosen by performing a grid search with the evaluation data or by more advanced techniques (such as meta-learning).

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 3 Neural networks (2.5 weeks):Lecture 8. Introduction to Feedforward Neural Networks / 6. Hidden Layer Models