Review

We wish to create a model to predict the stock movement based on financial news articles. The training data is composed of a list of news associated with the following targets: 1 if the daily return is > +1%, 0 if the daily return is between -1% and 1% and -1 if the daily return is < 1%

Preprocessing the data

The dataset is composed of N news of different lengths. We split the dataset into N_train samples for the training, N_validation samples for the validation, and N_test samples for the testing.

Sentence 1 $\quad w_1^1 \quad w_1^2 \quad w_1^3 \quad \cdots \quad w_1^{T_1}$

Sentence 2 $\quad w_2^1 \quad w_2^2 \quad w_2^3 \quad \cdots \quad w_2^{T_2}$

Sentence 3 $\quad w_3^1 \quad w_3^2 \quad w_3^3 \quad \cdots \quad w_3^{T_3}$

$$\vdots \qquad \vdots \qquad \vdots$$

Sentence N $\quad w_N^1 \quad w_N^2 \quad w_N^3 \quad \cdots \quad w_N^{T_N}$

1. Describe the first preprocessing step, which consists in transforming the list of sentences into a list of lists of integers.

2. Explain briefly why we split the dataset into train-validation-test data instead of train-test data?

_____

_____

_____

_____
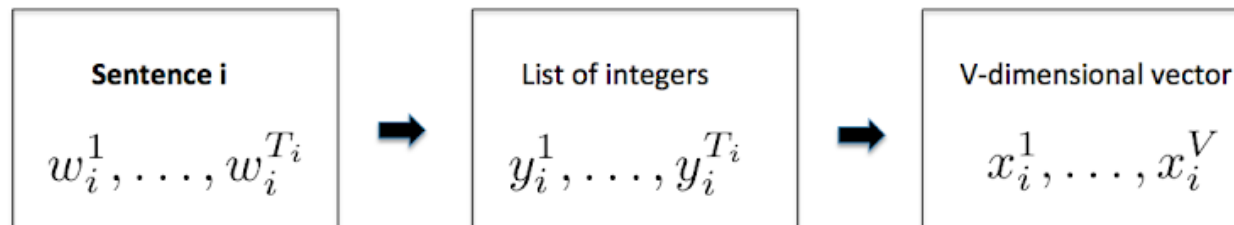
3. Let V be the vocabulary size. How can each list of integers (representing a sentence) be encoded into a V-dimensional vector?

Sentence i		List of integers		V-dimensional vector
$w_i^1, \ldots, w_i^{T_i}$	➡	$y_i^1, \ldots, y_i^{T_i}$	➡	$x_i^1, \ldots, x_i^V$

_____

_____

_____

_____

4. What would be the shape of the targets after the one-hot encoding process?
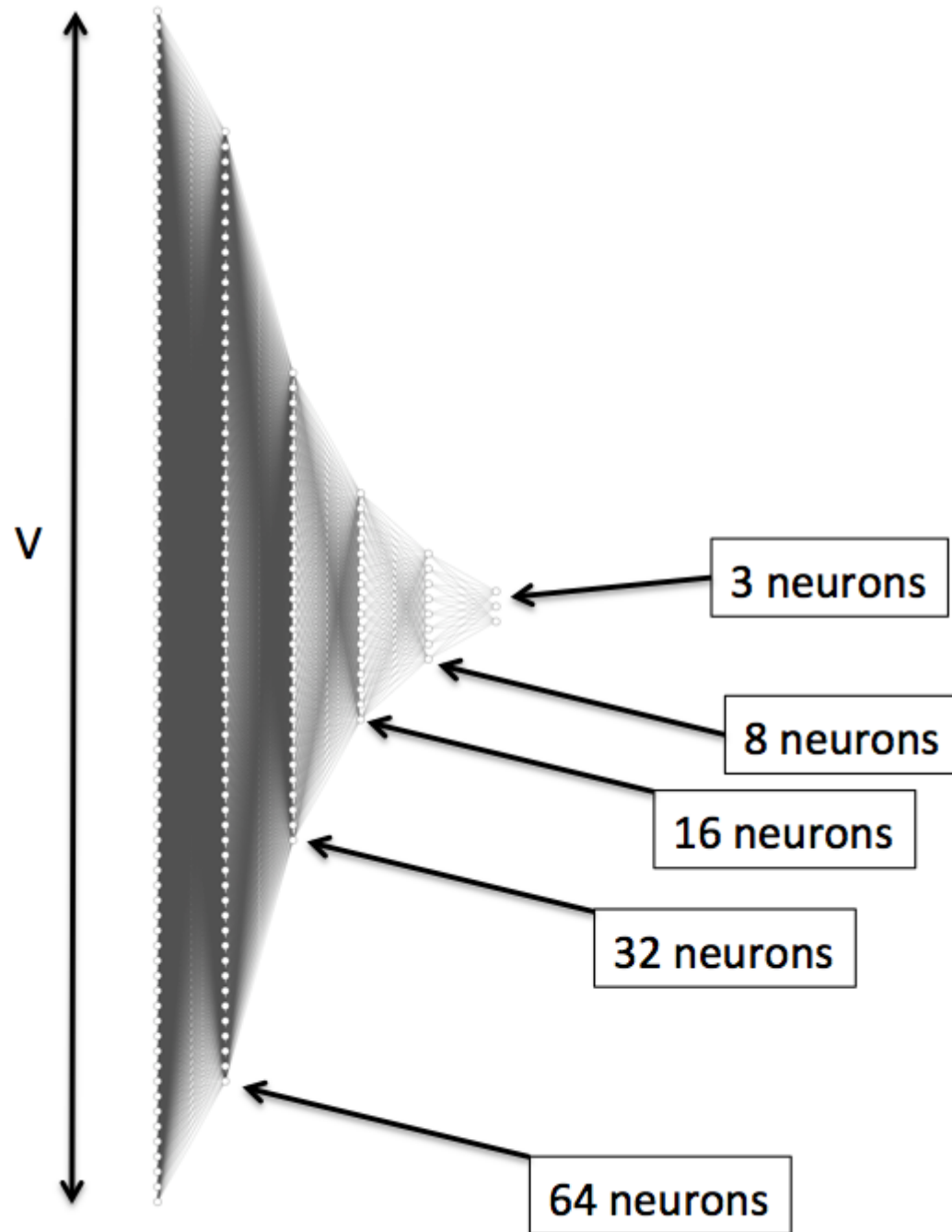
_____

_____

_____

_____

_____

A Feedforward Neural Network

We wish to use a feedforward neural network to classify the news using the following network:

Let V = 20000. We use 5 dense layers of lengths: 64, 32, 16, 8 and 3.

V

3 neurons

8 neurons

16 neurons

32 neurons

64 neurons

5. Describe the evolution of the data shape after each layer transformation.
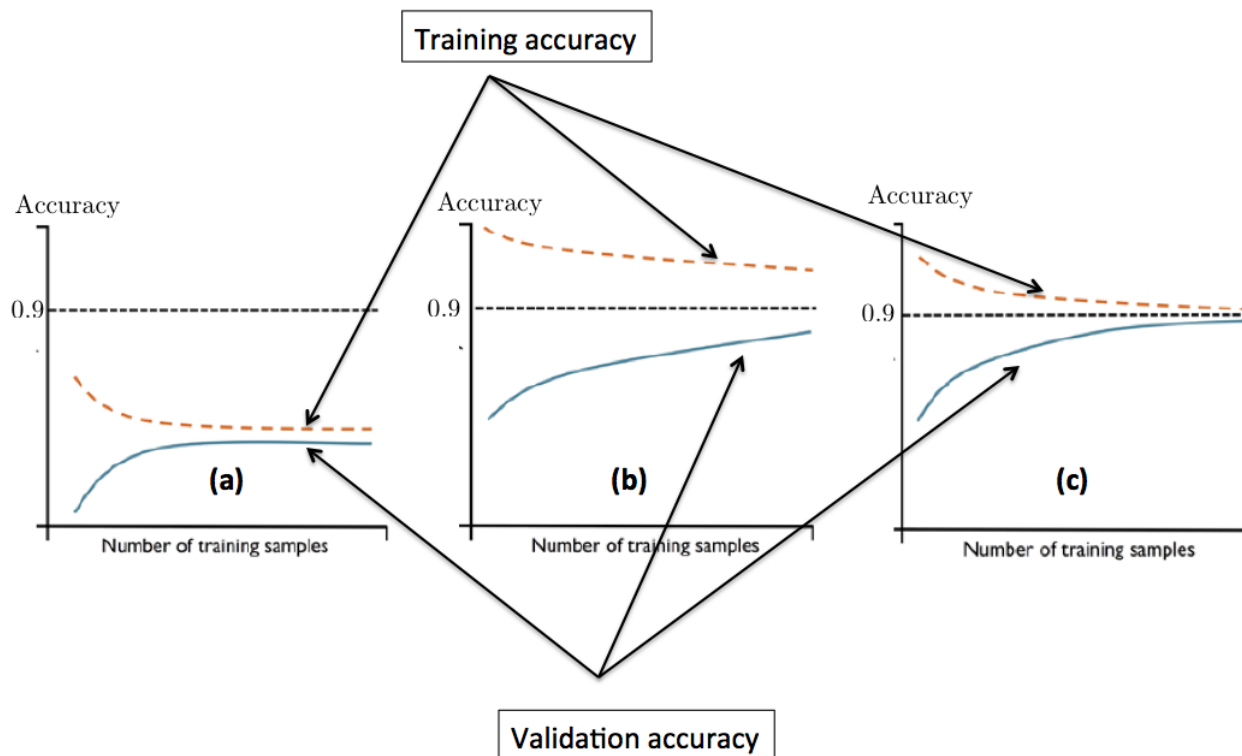
_____

_____

_____

_____

_____

We wish to compare 3 different architectures: model (a), model (b), and model (c). To that end, we plot the training and validation accuracy with respect to the number of samples for each model, we obtain the following curves:

6. Which model would you prefer and what issues do the other models exhibit?

_____

_____

_____

_____

_____

Google Forms