

```
In [ ]: import keras  
keras.__version__
```

Classifying newswires: a multi-class classification example

The Reuters dataset We will be working with the Reuters dataset, a set of short newswires and their topics, published by Reuters in 1986. It's a very simple, widely used toy dataset for text classification. There are 46 different topics; some topics are more represented than others, but each topic has at least 10 examples in the training set. Like IMDB and MNIST, the Reuters dataset comes packaged as part of Keras. Let's take a look right away:

```
In [1]: from keras.datasets import reuters  
(train_data, train_labels), (test_data, test_labels) = reuters.load_data(num_wor  
  
Downloading data from https://storage.googleapis.com/tensorflow/tf-keras-dataset  
s/reuters.npz  
2113536/2110848 [=====] - 0s 0us/step  
/opt/conda/lib/python3.8/site-packages/tensorflow/python/keras/datasets/reuters.  
py:148: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequen  
ces (which is a list-or-tuple of lists-or-tuples-or ndarrays with different leng  
ths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=o  
bject' when creating the ndarray  
    x_train, y_train = np.array(xs[:idx]), np.array(labels[:idx])  
/opt/conda/lib/python3.8/site-packages/tensorflow/python/keras/datasets/reuters.  
py:149: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequen  
ces (which is a list-or-tuple of lists-or-tuples-or ndarrays with different leng  
ths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=o  
bject' when creating the ndarray  
    x_test, y_test = np.array(xs[idx:]), np.array(labels[idx:])
```

Like with the IMDB dataset, the argument num_words=10000 restricts the data to the 10,000 most frequently occurring words found in the data. We have 8,982 training examples and 2,246 test examples:

```
In [2]: len(train_data)
```

```
Out[2]: 8982
```

```
In [3]: len(test_data)
```

```
Out[3]: 2246
```

As with the IMDB reviews, each example is a list of integers (word indices):

```
In [4]: train_data[10]
```

```
Out[4]: [1,  
        245,  
        273,  
        207,  
        156,  
        53,
```

```
74,  
160,  
26,  
14,  
46,  
296,  
26,  
39,  
74,  
2979,  
3554,  
14,  
46,  
4689,  
4329,  
86,  
61,  
3499,  
4795,  
14,  
61,  
451,  
4329,  
17,  
12]
```

Here's how you can decode it back to words, in case you are curious:

```
In [5]:  
word_index = reuters.get_word_index()  
reverse_word_index = dict([(value, key) for (key, value) in word_index.items()])  
# Note that our indices were offset by 3  
# because 0, 1 and 2 are reserved indices for "padding", "start of sequence", and  
decoded_newswire = ' '.join([reverse_word_index.get(i - 3, '?') for i in train_d
```

```
Downloading data from https://storage.googleapis.com/tensorflow/tf-keras-dataset  
s/reuters_word_index.json  
557056/550378 [=====] - 0s 0us/step
```

```
In [6]:  
decoded_newswire
```

```
Out[6]: '? ? ? said as a result of its december acquisition of space co it expects earni  
ngs per share in 1987 of 1 15 to 1 30 dlr per share up from 70 cts in 1986 the  
company said pretax net should rise to nine to 10 mln dlr from six mln dlr in  
1986 and rental operation revenues to 19 to 22 mln dlr from 12 5 mln dlr it sa  
id cash flow per share this year should be 2 50 to three dlr reuter 3'
```

The label associated with an example is an integer between 0 and 45: a topic index.

```
In [7]:  
train_labels[10]
```

```
Out[7]: 3
```

Preparing the data We can vectorize the data with the exact same code as in our previous example:

```
In [10]:  
import numpy as np  
def vectorize_sequences(sequences, dimension=10000):  
    results = np.zeros((len(sequences), dimension))  
    for i, sequence in enumerate(sequences):
```

```

        results[i, sequence] = 1.
    return results
# Our vectorized training data
x_train = vectorize_sequences(train_data)
# Our vectorized test data
x_test = vectorize_sequences(test_data)

```

To vectorize the labels, there are two possibilities: we could just cast the label list as an integer tensor, or we could use a “one-hot” encoding. One-hot encoding is a widely used format for categorical data, also called “categorical encoding”. For a more detailed explanation of one-hot encoding, you can refer to Chapter 6, Section 1. In our case, one-hot encoding of our labels consists in embedding each label as an all-zero vector with a 1 in the place of the label index, e.g.:

```
In [11]: def to_one_hot(labels, dimension=46):
    results = np.zeros((len(labels), dimension))
    for i, label in enumerate(labels):
        results[i, label] = 1.
    return results
# Our vectorized training labels
one_hot_train_labels = to_one_hot(train_labels)
# Our vectorized test labels
one_hot_test_labels = to_one_hot(test_labels)
```

Note that there is a built-in way to do this in Keras, which you have already seen in action in our MNIST example:

```
In [12]: from keras.utils.np_utils import to_categorical
one_hot_train_labels = to_categorical(train_labels)
one_hot_test_labels = to_categorical(test_labels)
```

Building our network This topic classification problem looks very similar to our previous movie review classification problem: in both cases, we are trying to classify short snippets of text. There is however a new constraint here: the number of output classes has gone from 2 to 46, i.e. the dimensionality of the output space is much larger. In a stack of Dense layers like what we were using, each layer can only access information present in the output of the previous layer. If one layer drops some information relevant to the classification problem, this information can never be recovered by later layers: each layer can potentially become an “information bottleneck”. In our previous example, we were using 16-dimensional intermediate layers, but a 16-dimensional space may be too limited to learn to separate 46 different classes: such small layers may act as information bottlenecks, permanently dropping relevant information. For this reason we will use larger layers. Let’s go with 64 units:

```
In [13]: from keras import models
from keras import layers
model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(46, activation='softmax'))
```

There are two other things you should note about this architecture:

- We are ending the network with a Dense layer of size 46. This means that for each input sample, our network will output a 46-dimensional vector. Each entry in this vector (each dimension) will encode a different output class.
- The last layer uses a softmax activation. You have already seen this pattern in the MNIST example. It means that the network will output a probability distribution over the 46 different output classes, i.e. for every input sample, the network will produce a 46-dimensional output vector where $\text{output}[i]$ is the probability that the sample belongs to class i . The 46 scores will sum to 1. The best loss function to use in this case is categorical_crossentropy. It measures the distance between two probability distributions: in our case, between the probability distribution output by our network, and the true distribution of the labels. By minimizing the distance between these two distributions, we train our network to output something as close as possible to the true labels.

```
In [14]: model.compile(optimizer='rmsprop',
                    loss='categorical_crossentropy',
                    metrics=['accuracy'])
```

Validating our approach Let's set apart 1,000 samples in our training data to use as a validation set:

```
In [15]: x_val = x_train[:1000]
partial_x_train = x_train[1000:]
y_val = one_hot_train_labels[:1000]
partial_y_train = one_hot_train_labels[1000:]
```

Now let's train our network for 20 epochs:

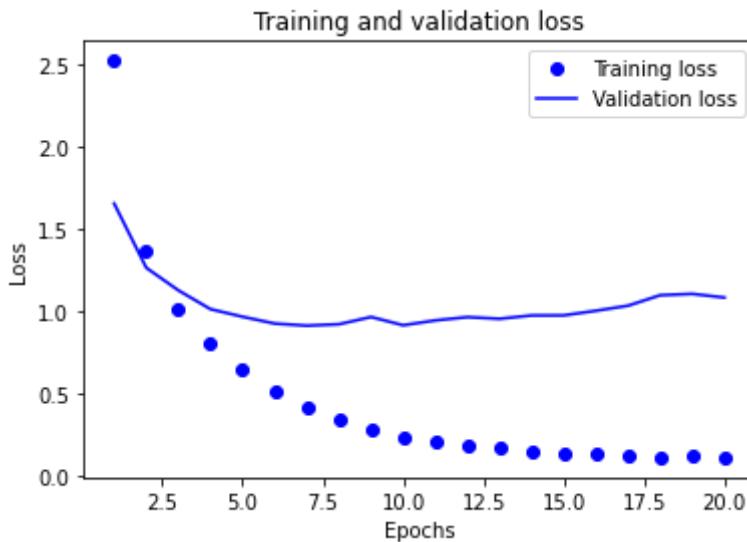
```
In [16]: history = model.fit(partial_x_train,
                        partial_y_train,
                        epochs=20,
                        batch_size=512,
                        validation_data=(x_val, y_val))
```

```
Epoch 1/20
16/16 [=====] - 1s 52ms/step - loss: 3.0305 - accuracy: 0.4064 - val_loss: 1.6530 - val_accuracy: 0.6490
Epoch 2/20
16/16 [=====] - 0s 18ms/step - loss: 1.4527 - accuracy: 0.6967 - val_loss: 1.2629 - val_accuracy: 0.7240
Epoch 3/20
16/16 [=====] - 0s 26ms/step - loss: 1.0512 - accuracy: 0.7714 - val_loss: 1.1259 - val_accuracy: 0.7390
Epoch 4/20
16/16 [=====] - 0s 16ms/step - loss: 0.8317 - accuracy: 0.8203 - val_loss: 1.0119 - val_accuracy: 0.7960
Epoch 5/20
16/16 [=====] - 0s 15ms/step - loss: 0.6550 - accuracy: 0.8657 - val_loss: 0.9657 - val_accuracy: 0.7890
Epoch 6/20
16/16 [=====] - 0s 17ms/step - loss: 0.5116 - accuracy: 0.8913 - val_loss: 0.9244 - val_accuracy: 0.8060
Epoch 7/20
16/16 [=====] - 0s 15ms/step - loss: 0.4090 - accuracy:
```

```
0.9177 - val_loss: 0.9118 - val_accuracy: 0.8100
Epoch 8/20
16/16 [=====] - 0s 15ms/step - loss: 0.3358 - accuracy: 0.9338 - val_loss: 0.9203 - val_accuracy: 0.8120
Epoch 9/20
16/16 [=====] - 0s 15ms/step - loss: 0.2795 - accuracy: 0.9387 - val_loss: 0.9644 - val_accuracy: 0.7930
Epoch 10/20
16/16 [=====] - 0s 14ms/step - loss: 0.2338 - accuracy: 0.9482 - val_loss: 0.9142 - val_accuracy: 0.8080
Epoch 11/20
16/16 [=====] - 0s 16ms/step - loss: 0.1910 - accuracy: 0.9563 - val_loss: 0.9436 - val_accuracy: 0.8130
Epoch 12/20
16/16 [=====] - 0s 16ms/step - loss: 0.1743 - accuracy: 0.9535 - val_loss: 0.9636 - val_accuracy: 0.8040
Epoch 13/20
16/16 [=====] - 0s 16ms/step - loss: 0.1565 - accuracy: 0.9571 - val_loss: 0.9535 - val_accuracy: 0.8110
Epoch 14/20
16/16 [=====] - 0s 17ms/step - loss: 0.1352 - accuracy: 0.9581 - val_loss: 0.9739 - val_accuracy: 0.8120
Epoch 15/20
16/16 [=====] - 0s 15ms/step - loss: 0.1211 - accuracy: 0.9622 - val_loss: 0.9743 - val_accuracy: 0.8090
Epoch 16/20
16/16 [=====] - 0s 20ms/step - loss: 0.1182 - accuracy: 0.9630 - val_loss: 1.0021 - val_accuracy: 0.8150
Epoch 17/20
16/16 [=====] - 0s 23ms/step - loss: 0.1027 - accuracy: 0.9663 - val_loss: 1.0329 - val_accuracy: 0.8010
Epoch 18/20
16/16 [=====] - 1s 45ms/step - loss: 0.1062 - accuracy: 0.9592 - val_loss: 1.0965 - val_accuracy: 0.8090
Epoch 19/20
16/16 [=====] - 0s 25ms/step - loss: 0.1123 - accuracy: 0.9584 - val_loss: 1.1046 - val_accuracy: 0.7980
Epoch 20/20
16/16 [=====] - 1s 36ms/step - loss: 0.0961 - accuracy: 0.9625 - val_loss: 1.0821 - val_accuracy: 0.8070
```

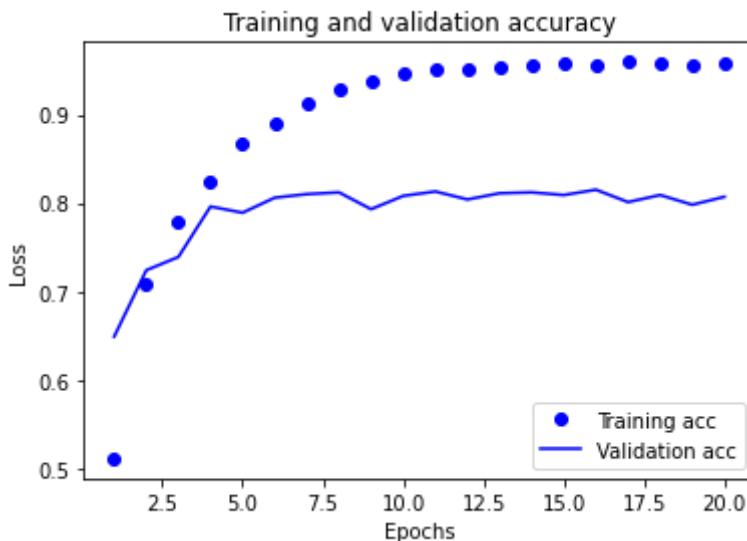
In [17]:

```
import matplotlib.pyplot as plt
loss = history.history['loss']
val_loss = history.history['val_loss']
epochs = range(1, len(loss) + 1)
plt.plot(epochs, loss, 'bo', label='Training loss')
plt.plot(epochs, val_loss, 'b', label='Validation loss')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```



In [18]:

```
plt.clf() # clear figure
acc = history.history['accuracy']
val_acc = history.history['val_accuracy']
plt.plot(epochs, acc, 'bo', label='Training acc')
plt.plot(epochs, val_acc, 'b', label='Validation acc')
plt.title('Training and validation accuracy')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()
```



It seems that the network starts overfitting after 8 epochs. Let's train a new network from scratch for 8 epochs, then let's evaluate it on the test set:

In [19]:

```
model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(64, activation='relu'))
model.add(layers.Dense(46, activation='softmax'))
model.compile(optimizer='rmsprop',
              loss='categorical_crossentropy',
              metrics=['accuracy'])
```

```

model.fit(partial_x_train,
          partial_y_train,
          epochs=8,
          batch_size=512,
          validation_data=(x_val, y_val))
results = model.evaluate(x_test, one_hot_test_labels)

```

```

Epoch 1/8
16/16 [=====] - 1s 34ms/step - loss: 3.0907 - accuracy: 0.4218 - val_loss: 1.7076 - val_accuracy: 0.6240
Epoch 2/8
16/16 [=====] - 0s 17ms/step - loss: 1.5064 - accuracy: 0.6785 - val_loss: 1.2864 - val_accuracy: 0.7240
Epoch 3/8
16/16 [=====] - 0s 16ms/step - loss: 1.0764 - accuracy: 0.7689 - val_loss: 1.1256 - val_accuracy: 0.7640
Epoch 4/8
16/16 [=====] - 0s 15ms/step - loss: 0.8352 - accuracy: 0.8244 - val_loss: 1.0192 - val_accuracy: 0.7950
Epoch 5/8
16/16 [=====] - 0s 17ms/step - loss: 0.6641 - accuracy: 0.8648 - val_loss: 0.9489 - val_accuracy: 0.7980
Epoch 6/8
16/16 [=====] - 0s 16ms/step - loss: 0.5354 - accuracy: 0.8908 - val_loss: 0.9256 - val_accuracy: 0.8010
Epoch 7/8
16/16 [=====] - 0s 16ms/step - loss: 0.4195 - accuracy: 0.9135 - val_loss: 0.9038 - val_accuracy: 0.8140
Epoch 8/8
16/16 [=====] - 0s 16ms/step - loss: 0.3578 - accuracy: 0.9260 - val_loss: 0.9083 - val_accuracy: 0.8060
71/71 [=====] - 0s 2ms/step - loss: 1.0059 - accuracy: 0.7814

```

In [20]: results

Out[20]: [1.005916953086853, 0.7813891172409058]

Our approach reaches an accuracy of ~78%. With a balanced binary classification problem, the accuracy reached by a purely random classifier would be 50%, but in our case it is closer to 19%, so our results seem pretty good, at least when compared to a random baseline:

```

In [21]: import copy
test_labels_copy = copy.copy(test_labels)
np.random.shuffle(test_labels_copy)
float(np.sum(np.array(test_labels) == np.array(test_labels_copy))) / len(test_la

```

Out[21]: 0.1745325022261799

Generating predictions on new data We can verify that the predict method of our model instance returns a probability distribution over all 46 topics. Let's generate topic predictions for all of the test data:

In [23]: predictions = model.predict(x_test)

Each entry in predictions is a vector of length 46:

```
In [24]: predictions[0].shape
```

```
Out[24]: (46,)
```

The coefficients in this vector sum to 1:

```
In [25]: np.sum(predictions[0])
```

```
Out[25]: 0.9999999
```

The largest entry is the predicted class, i.e. the class with the highest probability:

```
In [26]: np.argmax(predictions[0])
```

```
Out[26]: 3
```

A different way to handle the labels and the loss We mentioned earlier that another way to encode the labels would be to cast them as an integer tensor, like such:

```
In [27]: y_train = np.array(train_labels)
y_test = np.array(test_labels)
```

The only thing it would change is the choice of the loss function. Our previous loss, categorical_crossentropy, expects the labels to follow a categorical encoding. With integer labels, we should use sparse_categorical_crossentropy:

```
In [28]: model.compile(optimizer='rmsprop', loss='sparse_categorical_crossentropy', metri
```

This new loss function is still mathematically the same as categorical_crossentropy; it just has a different interface.

On the importance of having sufficiently large intermediate layers We mentioned earlier that since our final outputs were 46-dimensional, we should avoid intermediate layers with much less than 46 hidden units. Now let's try to see what happens when we introduce an information bottleneck by having intermediate layers significantly less than 46-dimensional, e.g. 4-dimensional.

```
In [29]: model = models.Sequential()
model.add(layers.Dense(64, activation='relu', input_shape=(10000,)))
model.add(layers.Dense(4, activation='relu'))
model.add(layers.Dense(46, activation='softmax'))
model.compile(optimizer='rmsprop',
              loss='categorical_crossentropy',
              metrics=['accuracy'])
model.fit(partial_x_train,
          partial_y_train,
          epochs=20,
          batch_size=128,
          validation_data=(x_val, y_val))
```

```
Epoch 1/20
63/63 [=====] - 1s 14ms/step - loss: 3.2875 - accuracy: 0.2245 - val_loss: 2.0773 - val_accuracy: 0.5900
Epoch 2/20
63/63 [=====] - 0s 7ms/step - loss: 1.8454 - accuracy: 0.6294 - val_loss: 1.5580 - val_accuracy: 0.6160
Epoch 3/20
63/63 [=====] - 0s 7ms/step - loss: 1.3933 - accuracy: 0.6480 - val_loss: 1.4191 - val_accuracy: 0.6260
Epoch 4/20
63/63 [=====] - 1s 8ms/step - loss: 1.1805 - accuracy: 0.6780 - val_loss: 1.3439 - val_accuracy: 0.6770
Epoch 5/20
63/63 [=====] - 0s 7ms/step - loss: 1.0636 - accuracy: 0.7094 - val_loss: 1.3204 - val_accuracy: 0.6780
Epoch 6/20
63/63 [=====] - 0s 7ms/step - loss: 0.9386 - accuracy: 0.7341 - val_loss: 1.3092 - val_accuracy: 0.6830
Epoch 7/20
63/63 [=====] - 0s 7ms/step - loss: 0.8552 - accuracy: 0.7515 - val_loss: 1.3306 - val_accuracy: 0.6910
Epoch 8/20
63/63 [=====] - 0s 7ms/step - loss: 0.8187 - accuracy: 0.7786 - val_loss: 1.3279 - val_accuracy: 0.7050
Epoch 9/20
63/63 [=====] - 0s 7ms/step - loss: 0.7532 - accuracy: 0.7994 - val_loss: 1.3495 - val_accuracy: 0.7010
Epoch 10/20
63/63 [=====] - 0s 7ms/step - loss: 0.6923 - accuracy: 0.8148 - val_loss: 1.3534 - val_accuracy: 0.7050
Epoch 11/20
63/63 [=====] - 0s 7ms/step - loss: 0.6409 - accuracy: 0.8264 - val_loss: 1.3854 - val_accuracy: 0.7060
Epoch 12/20
63/63 [=====] - 0s 7ms/step - loss: 0.6051 - accuracy: 0.8354 - val_loss: 1.4120 - val_accuracy: 0.7020
Epoch 13/20
63/63 [=====] - 0s 7ms/step - loss: 0.5547 - accuracy: 0.8525 - val_loss: 1.4480 - val_accuracy: 0.7040
Epoch 14/20
63/63 [=====] - 0s 7ms/step - loss: 0.5276 - accuracy: 0.8520 - val_loss: 1.5040 - val_accuracy: 0.7010
Epoch 15/20
63/63 [=====] - 0s 7ms/step - loss: 0.4925 - accuracy: 0.8641 - val_loss: 1.5289 - val_accuracy: 0.7020
Epoch 16/20
63/63 [=====] - 0s 6ms/step - loss: 0.5034 - accuracy: 0.8528 - val_loss: 1.5808 - val_accuracy: 0.7080
Epoch 17/20
63/63 [=====] - 0s 7ms/step - loss: 0.4593 - accuracy: 0.8658 - val_loss: 1.6337 - val_accuracy: 0.7030
Epoch 18/20
63/63 [=====] - 0s 7ms/step - loss: 0.4242 - accuracy: 0.8785 - val_loss: 1.6955 - val_accuracy: 0.7060
Epoch 19/20
63/63 [=====] - 0s 7ms/step - loss: 0.4285 - accuracy: 0.8771 - val_loss: 1.7835 - val_accuracy: 0.7010
Epoch 20/20
63/63 [=====] - 0s 7ms/step - loss: 0.4001 - accuracy: 0.8787 - val_loss: 1.7727 - val_accuracy: 0.7020
```

Out[29]: <tensorflow.python.keras.callbacks.History at 0x7eff405b3bb0>

Our network now seems to peak at ~71% test accuracy, a 8% absolute drop. This drop is mostly due to the fact that we are now trying to compress a lot of information (enough information to

recover the separation hyperplanes of 46 classes) into an intermediate space that is too low dimensional. The network is able to cram most of the necessary information into these 8-dimensional representations, but not all of it.

Further experiments

- Try using larger or smaller layers: 32 units, 128 units... • We were using two hidden layers. Now try to use a single hidden layer, or three hidden layers.

Wrapping up Here's what you should take away from this example:

- If you are trying to classify data points between N classes, your network should end with a Dense layer of size N.
- In a single-label, multi-class classification problem, your network should end with a softmax activation, so that it will output a probability distribution over the N output classes.
- Categorical crossentropy is almost always the loss function you should use for such problems. It minimizes the distance between the probability distributions output by the network, and the true distribution of the targets.
- There are two ways to handle labels in multi-class classification: **Encoding the labels via “categorical encoding” (also known as “one-hot encoding”) and using categorical_crossentropy as your loss function.** Encoding the labels as integers and using the sparse_categorical_crossentropy loss function.
- If you need to classify data into a large number of categories, then you should avoid creating information bottlenecks in your network by having intermediate layers that are too small.

In []: