

Assignment 5.3

DSC650-T302 Big Data (2235-1)

Jake Meyer

04/13/2023

```
In [1]: import keras  
keras.__version__
```

```
Out[1]: '2.11.0'
```

Predicting house prices: a regression example

This notebook contains the code samples found in Chapter 3, Section 6 of [Deep Learning with Python](#). Note that the original text features far more content, in particular further explanations and figures: in this notebook, you will only find source code and related comments.

In our two previous examples, we were considering classification problems, where the goal was to predict a single discrete label of an input data point. Another common type of machine learning problem is "regression", which consists of predicting a continuous value instead of a discrete label. For instance, predicting the temperature tomorrow, given meteorological data, or predicting the time that a software project will take to complete, given its specifications.

Do not mix up "regression" with the algorithm "logistic regression": confusingly, "logistic regression" is not a regression algorithm, it is a classification algorithm.

The Boston Housing Price dataset

We will be attempting to predict the median price of homes in a given Boston suburb in the mid-1970s, given a few data points about the suburb at the time, such as the crime rate, the local property tax rate, etc.

The dataset we will be using has another interesting difference from our two previous examples: it has very few data points, only 506 in total, split between 404 training samples and 102 test samples, and each "feature" in the input data (e.g. the crime rate is a feature)

has a different scale. For instance some values are proportions, which take a values between 0 and 1, others take values between 1 and 12, others between 0 and 100...

Let's take a look at the data:

```
In [2]: from keras.datasets import boston_housing  
  
(train_data, train_targets), (test_data, test_targets) = boston_housing.load_data()  
  
Downloading data from https://storage.googleapis.com/tensorflow/tf-keras-datasets/  
boston_housing.npz  
57026/57026 [=====] - 0s 1us/step
```

```
In [3]: train_data.shape
```

```
Out[3]: (404, 13)
```

```
In [4]: test_data.shape
```

```
Out[4]: (102, 13)
```

As you can see, we have 404 training samples and 102 test samples. The data comprises 13 features. The 13 features in the input data are as follow:

1. Per capita crime rate.
2. Proportion of residential land zoned for lots over 25,000 square feet.
3. Proportion of non-retail business acres per town.
4. Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
5. Nitric oxides concentration (parts per 10 million).
6. Average number of rooms per dwelling.
7. Proportion of owner-occupied units built prior to 1940.
8. Weighted distances to five Boston employment centres.
9. Index of accessibility to radial highways.
10. Full-value property-tax rate per \$10,000.
11. Pupil-teacher ratio by town.
12. $1000 * (Bk - 0.63)^2$ where Bk is the proportion of Black people by town.
13. % lower status of the population.

The targets are the median values of owner-occupied homes, in thousands of dollars:

```
In [5]: train_targets
```

```
Out[5]: array([15.2, 42.3, 50., 21.1, 17.7, 18.5, 11.3, 15.6, 15.6, 14.4, 12.1,
   17.9, 23.1, 19.9, 15.7, 8.8, 50., 22.5, 24.1, 27.5, 10.9, 30.8,
   32.9, 24., 18.5, 13.3, 22.9, 34.7, 16.6, 17.5, 22.3, 16.1, 14.9,
   23.1, 34.9, 25., 13.9, 13.1, 20.4, 20., 15.2, 24.7, 22.2, 16.7,
   12.7, 15.6, 18.4, 21., 30.1, 15.1, 18.7, 9.6, 31.5, 24.8, 19.1,
   22., 14.5, 11., 32., 29.4, 20.3, 24.4, 14.6, 19.5, 14.1, 14.3,
   15.6, 10.5, 6.3, 19.3, 19.3, 13.4, 36.4, 17.8, 13.5, 16.5, 8.3,
   14.3, 16., 13.4, 28.6, 43.5, 20.2, 22., 23., 20.7, 12.5, 48.5,
   14.6, 13.4, 23.7, 50., 21.7, 39.8, 38.7, 22.2, 34.9, 22.5, 31.1,
   28.7, 46., 41.7, 21., 26.6, 15., 24.4, 13.3, 21.2, 11.7, 21.7,
   19.4, 50., 22.8, 19.7, 24.7, 36.2, 14.2, 18.9, 18.3, 20.6, 24.6,
   18.2, 8.7, 44., 10.4, 13.2, 21.2, 37., 30.7, 22.9, 20., 19.3,
   31.7, 32., 23.1, 18.8, 10.9, 50., 19.6, 5., 14.4, 19.8, 13.8,
   19.6, 23.9, 24.5, 25., 19.9, 17.2, 24.6, 13.5, 26.6, 21.4, 11.9,
   22.6, 19.6, 8.5, 23.7, 23.1, 22.4, 20.5, 23.6, 18.4, 35.2, 23.1,
   27.9, 20.6, 23.7, 28., 13.6, 27.1, 23.6, 20.6, 18.2, 21.7, 17.1,
   8.4, 25.3, 13.8, 22.2, 18.4, 20.7, 31.6, 30.5, 20.3, 8.8, 19.2,
   19.4, 23.1, 23., 14.8, 48.8, 22.6, 33.4, 21.1, 13.6, 32.2, 13.1,
   23.4, 18.9, 23.9, 11.8, 23.3, 22.8, 19.6, 16.7, 13.4, 22.2, 20.4,
   21.8, 26.4, 14.9, 24.1, 23.8, 12.3, 29.1, 21., 19.5, 23.3, 23.8,
   17.8, 11.5, 21.7, 19.9, 25., 33.4, 28.5, 21.4, 24.3, 27.5, 33.1,
   16.2, 23.3, 48.3, 22.9, 22.8, 13.1, 12.7, 22.6, 15., 15.3, 10.5,
   24., 18.5, 21.7, 19.5, 33.2, 23.2, 5., 19.1, 12.7, 22.3, 10.2,
   13.9, 16.3, 17., 20.1, 29.9, 17.2, 37.3, 45.4, 17.8, 23.2, 29.,
   22., 18., 17.4, 34.6, 20.1, 25., 15.6, 24.8, 28.2, 21.2, 21.4,
   23.8, 31., 26.2, 17.4, 37.9, 17.5, 20., 8.3, 23.9, 8.4, 13.8,
   7.2, 11.7, 17.1, 21.6, 50., 16.1, 20.4, 20.6, 21.4, 20.6, 36.5,
   8.5, 24.8, 10.8, 21.9, 17.3, 18.9, 36.2, 14.9, 18.2, 33.3, 21.8,
   19.7, 31.6, 24.8, 19.4, 22.8, 7.5, 44.8, 16.8, 18.7, 50., 50.,
   19.5, 20.1, 50., 17.2, 20.8, 19.3, 41.3, 20.4, 20.5, 13.8, 16.5,
   23.9, 20.6, 31.5, 23.3, 16.8, 14., 33.8, 36.1, 12.8, 18.3, 18.7,
   19.1, 29., 30.1, 50., 50., 22., 11.9, 37.6, 50., 22.7, 20.8,
   23.5, 27.9, 50., 19.3, 23.9, 22.6, 15.2, 21.7, 19.2, 43.8, 20.3,
   33.2, 19.9, 22.5, 32.7, 22., 17.1, 19., 15., 16.1, 25.1, 23.7,
   28.7, 37.2, 22.6, 16.4, 25., 29.8, 22.1, 17.4, 18.1, 30.3, 17.5,
   24.7, 12.6, 26.5, 28.7, 13.3, 10.4, 24.4, 23., 20., 17.8, 7.,
   11.8, 24.4, 13.8, 19.4, 25.2, 19.4, 19.4, 29.1])
```

The prices are typically between 10,000 and 50,000. If that sounds cheap, remember this was the mid-1970s, and these prices are not inflation-adjusted.

Preparing the data

It would be problematic to feed into a neural network values that all take wildly different ranges. The network might be able to automatically adapt to such heterogeneous data, but it would definitely make learning more difficult. A widespread best practice to deal with such data is to do feature-wise normalization: for each feature in the input data (a column in the input data matrix), we will subtract the mean of the feature and divide by the standard deviation, so that the feature will be centered around 0 and will have a unit standard deviation. This is easily done in Numpy:

```
In [6]: mean = train_data.mean(axis=0)
train_data -= mean
std = train_data.std(axis=0)
train_data /= std

test_data -= mean
test_data /= std
```

Note that the quantities that we use for normalizing the test data have been computed using the training data. We should never use in our workflow any quantity computed on the test data, even for something as simple as data normalization.

Building our network

Because so few samples are available, we will be using a very small network with two hidden layers, each with 64 units. In general, the less training data you have, the worse overfitting will be, and using a small network is one way to mitigate overfitting.

```
In [7]: from keras import models
from keras import layers

def build_model():
    # Because we will need to instantiate
    # the same model multiple times,
    # we use a function to construct it.
    model = models.Sequential()
    model.add(layers.Dense(64, activation='relu',
                          input_shape=(train_data.shape[1],)))
    model.add(layers.Dense(64, activation='relu'))
    model.add(layers.Dense(1))
    model.compile(optimizer='rmsprop', loss='mse', metrics=['mae'])
    return model
```

Our network ends with a single unit, and no activation (i.e. it will be linear layer). This is a typical setup for scalar regression (i.e. regression where we are trying to predict a single continuous value). Applying an activation function would constrain the range that the output can take; for instance if we applied a `sigmoid` activation function to our last layer, the network could only learn to predict values between 0 and 1. Here, because the last layer is purely linear, the network is free to learn to predict values in any range.

Note that we are compiling the network with the `mse` loss function -- Mean Squared Error, the square of the difference between the predictions and the targets, a widely used loss function for regression problems.

We are also monitoring a new metric during training: `mae`. This stands for Mean Absolute Error. It is simply the absolute value of the difference between the predictions and the targets. For instance, a MAE of 0.5 on this problem would mean that our predictions are off by \$500 on average.

Validating our approach using K-fold validation

To evaluate our network while we keep adjusting its parameters (such as the number of epochs used for training), we could simply split the data into a training set and a validation set, as we were doing in our previous examples. However, because we have so few data points, the validation set would end up being very small (e.g. about 100 examples). A consequence is that our validation scores may change a lot depending on *which* data points we choose to use for validation and which we choose for training, i.e. the validation scores may have a high *variance* with regard to the validation split. This would prevent us from reliably evaluating our model.

The best practice in such situations is to use K-fold cross-validation. It consists of splitting the available data into K partitions (typically K=4 or 5), then instantiating K identical models, and training each one on K-1 partitions while evaluating on the remaining partition. The validation score for the model used would then be the average of the K validation scores obtained.

In terms of code, this is straightforward:

```
In [8]: import numpy as np

k = 4
num_val_samples = len(train_data) // k
num_epochs = 100
all_scores = []
for i in range(k):
    print('processing fold #', i)
    # Prepare the validation data: data from partition # k
    val_data = train_data[i * num_val_samples: (i + 1) * num_val_samples]
    val_targets = train_targets[i * num_val_samples: (i + 1) * num_val_samples]
```

```

# Prepare the training data: data from all other partitions
partial_train_data = np.concatenate(
    [train_data[:i * num_val_samples],
     train_data[(i + 1) * num_val_samples:]],
    axis=0)
partial_train_targets = np.concatenate(
    [train_targets[:i * num_val_samples],
     train_targets[(i + 1) * num_val_samples:]],
    axis=0)

# Build the Keras model (already compiled)
model = build_model()
# Train the model (in silent mode, verbose=0)
model.fit(partial_train_data, partial_train_targets,
           epochs=num_epochs, batch_size=1, verbose=0)
# Evaluate the model on the validation data
val_mse, val_mae = model.evaluate(val_data, val_targets, verbose=0)
all_scores.append(val_mae)

```

```

processing fold # 0
processing fold # 1
processing fold # 2
processing fold # 3

```

In [9]: `all_scores`

Out[9]: `[2.158801794052124, 2.768878221511841, 2.6661810874938965, 2.3802413940429688]`

In [10]: `np.mean(all_scores)`

Out[10]: `2.4935256242752075`

As you can notice, the different runs do indeed show rather different validation scores, from 2.1 to 2.9. Their average (2.4) is a much more reliable metric than any single of these scores -- that's the entire point of K-fold cross-validation. In this case, we are off by **2,400 on average, which is still significant considering that the prices range from 10,000 to \$50,000.**

Let's try training the network for a bit longer: 500 epochs. To keep a record of how well the model did at each epoch, we will modify our training loop to save the per-epoch validation score log:

In [11]: `from keras import backend as K`

```

# Some memory clean-up
K.clear_session()

```

In [14]: `num_epochs = 500`
`all_mae_histories = []`
`for i in range(k):`
 `print('processing fold #', i)`
 `# Prepare the validation data: data from partition # k`

```

val_data = train_data[i * num_val_samples: (i + 1) * num_val_samples]
val_targets = train_targets[i * num_val_samples: (i + 1) * num_val_samples]

# Prepare the training data: data from all other partitions
partial_train_data = np.concatenate(
    [train_data[:i * num_val_samples],
     train_data[(i + 1) * num_val_samples:]],
    axis=0)
partial_train_targets = np.concatenate(
    [train_targets[:i * num_val_samples],
     train_targets[(i + 1) * num_val_samples:]],
    axis=0)

# Build the Keras model (already compiled)
model = build_model()
# Train the model (in silent mode, verbose=0)
history = model.fit(partial_train_data, partial_train_targets,
                     validation_data=(val_data, val_targets),
                     epochs=num_epochs, batch_size=1, verbose=0)
mae_history = history.history['val_mae']
all_mae_histories.append(mae_history)

processing fold # 0
processing fold # 1
processing fold # 2
processing fold # 3

```

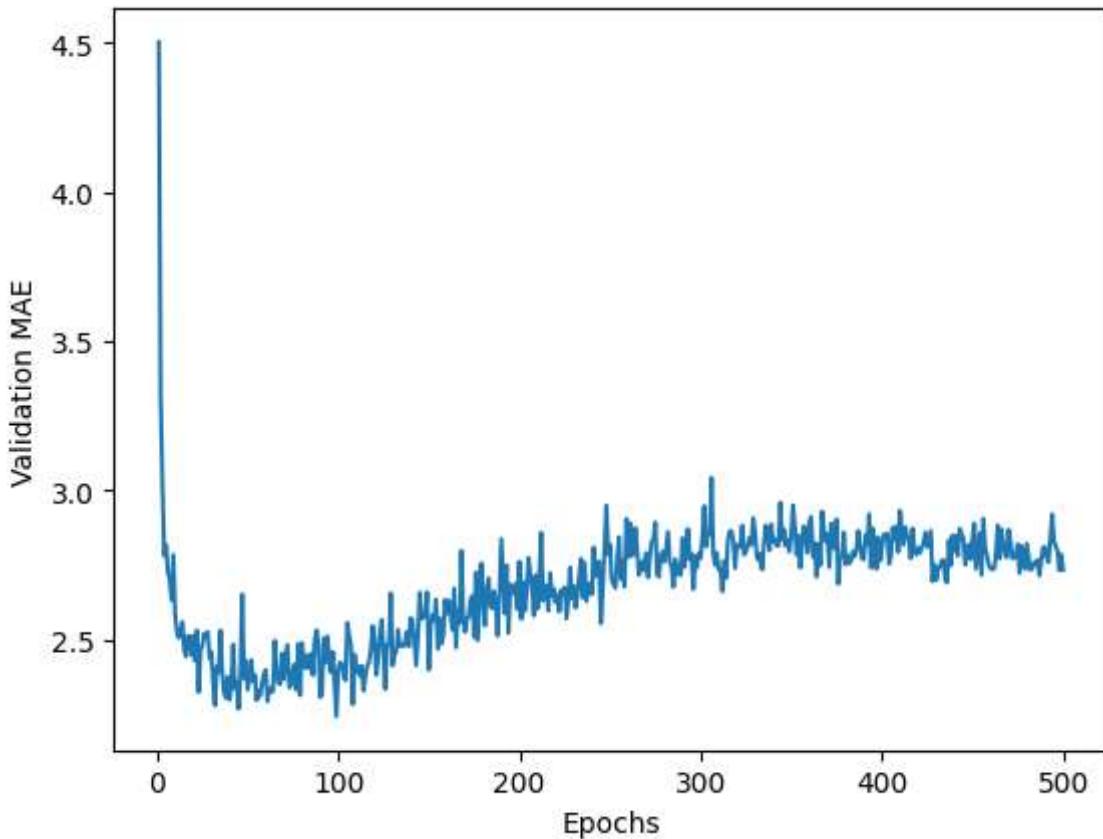
We can then compute the average of the per-epoch MAE scores for all folds:

```
In [15]: average_mae_history = [
    np.mean([x[i] for x in all_mae_histories]) for i in range(num_epochs)]
```

Let's plot this:

```
In [16]: import matplotlib.pyplot as plt

plt.plot(range(1, len(average_mae_history) + 1), average_mae_history)
plt.xlabel('Epochs')
plt.ylabel('Validation MAE')
plt.show()
```



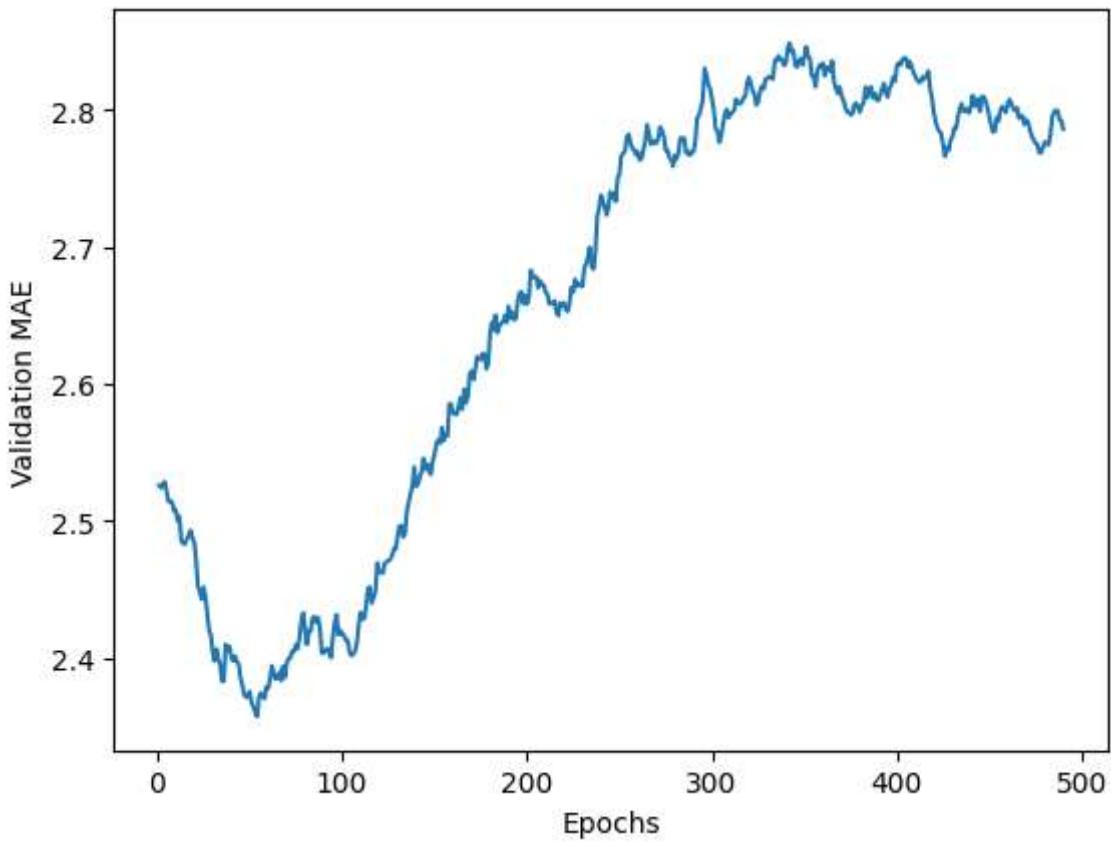
It may be a bit hard to see the plot due to scaling issues and relatively high variance. Let's:

- Omit the first 10 data points, which are on a different scale from the rest of the curve.
- Replace each point with an exponential moving average of the previous points, to obtain a smooth curve.

```
In [17]: def smooth_curve(points, factor=0.9):
    smoothed_points = []
    for point in points:
        if smoothed_points:
            previous = smoothed_points[-1]
            smoothed_points.append(previous * factor + point * (1 - factor))
        else:
            smoothed_points.append(point)
    return smoothed_points

smooth_mae_history = smooth_curve(average_mae_history[10:])

plt.plot(range(1, len(smooth_mae_history) + 1), smooth_mae_history)
plt.xlabel('Epochs')
plt.ylabel('Validation MAE')
plt.show()
```



According to this plot, it seems that validation MAE stops improving significantly after 80 epochs. Past that point, we start overfitting.

Once we are done tuning other parameters of our model (besides the number of epochs, we could also adjust the size of the hidden layers), we can train a final "production" model on all of the training data, with the best parameters, then look at its performance on the test data:

```
In [18]: # Get a fresh, compiled model.  
model = build_model()  
# Train it on the entirety of the data.  
model.fit(train_data, train_targets,  
          epochs=80, batch_size=16, verbose=0)  
test_mse_score, test_mae_score = model.evaluate(test_data, test_targets)  
  
4/4 [=====] - 0s 1ms/step - loss: 16.6546 - mae: 2.5694  
  
In [19]: test_mae_score  
  
Out[19]: 2.569382905960083
```

We are still off by about \$2,550.

Wrapping up

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

- Regression is done using different loss functions from classification; Mean Squared Error (MSE) is a commonly used loss function for

regression.

- Similarly, evaluation metrics to be used for regression differ from those used for classification; naturally the concept of "accuracy"

does not apply for regression. A common regression metric is Mean Absolute Error (MAE).

- When features in the input data have values in different ranges, each feature should be scaled independently as a preprocessing step.
- When there is little data available, using K-Fold validation is a great way to reliably evaluate a model.
- When little training data is available, it is preferable to use a small network with very few hidden layers (typically only one or two),

in order to avoid severe overfitting.

This example concludes our series of three introductory practical examples. You are now able to handle common types of problems with vector data input:

- Binary (2-class) classification.
- Multi-class, single-label classification.
- Scalar regression.

In the next chapter, you will acquire a more formal understanding of some of the concepts you have encountered in these first examples, such as data preprocessing, model evaluation, and overfitting.