exercise_1

October 29, 2018

```
In [377]: import numpy as np
          import _pickle as cPickle
          import os
          import gzip
```

1 First exercise: Classifying MNIST with MLPs

In this exercise you will implement a Neural Network (or MLP) and classify the MNIST digits with it. MNIST is a "well hung" dataset that has been used a lot over the years to benchmark different classification algorithms. To learn more about it have a look here: http://yann.lecun.com/exdb/mnist/ .

2 Data Loading

We first define a function for downloading and loading MNIST. **WARNING**: Executing it will obviously use up some space on your machine ;).

```
In [378]: def mnist(datasets_dir='./data'):
              if not os.path.exists(datasets_dir):
                  os.mkdir(datasets_dir)
              data_file = os.path.join(datasets_dir, 'mnist.pkl.gz')
              if not os.path.exists(data_file):
                  print('... downloading MNIST from the web')
                  try:
                      import urllib
                      urllib.urlretrieve('http://google.com')
                  except AttributeError:
                      import urllib.request as urllib
                  url = 'http://www.iro.umontreal.ca/~lisa/deep/data/mnist/mnist.pkl.gz'
                  urllib.urlretrieve(url, data_file)

              print('... loading data')
              # Load the dataset
              f = gzip.open(data_file, 'rb')
              try:
                  train_set, valid_set, test_set = cPickle.load(f, encoding="latin1")
              except TypeError:
```

```
        train_set, valid_set, test_set = cPickle.load(f)
    f.close()

    test_x, test_y = test_set
    test_x = test_x.astype('float32')
    test_x = test_x.astype('float32').reshape(test_x.shape[0], 1, 28, 28)
    test_y = test_y.astype('int32')
    valid_x, valid_y = valid_set
    valid_x = valid_x.astype('float32')
    valid_x = valid_x.astype('float32').reshape(valid_x.shape[0], 1, 28, 28)
    valid_y = valid_y.astype('int32')
    train_x, train_y = train_set
    train_x = train_x.astype('float32').reshape(train_x.shape[0], 1, 28, 28)
    train_y = train_y.astype('int32')
    rval = [(train_x, train_y), (valid_x, valid_y), (test_x, test_y)]
    print('... done loading data')
    return rval
```

3 Neural Network Layers

We now define "bare bone" neural network layers. The parts marked with **TODO** are where you should finish the implementation! Conceptually we will implement the layers as follows:

Each layer has a constructor that takes an input layer plus some additional arguments such as layer size and the activation function name. The layer then uses the provided input layer to compute the layer dimensions, weight shapes, etc. and setup all auxilliary variables.

Each layer then has to provide three functions (as defined in the Layer class below): *output_shape()*, *fprop()* and *brop()*. The output_shape function is used to figure out the shape for the next layer and the *fprop()/bprop()* functions are used to compute forward and backward passes through the network.

```
In [379]: # start by defining simple helpers
          def sigmoid(x):
              return 1.0/(1.0+np.exp(-x))

          def sigmoid_d(x):
              return sigmoid(x) * (1-sigmoid(x))

          def tanh(x):
              return np.tanh(x)

          def tanh_d(x):
              return 1-np.square(tanh(x))

          def relu(x):
              return np.maximum(0.0, x)

          def relu_d(x):
```

```python
    dx=np.zeros_like(x)
    dx[x>0]=1
    return dx

def softmax(x, axis=1):
    # to make the softmax a "safe" operation we will
    # first subtract the maximum along the specified axis
    # so that np.exp(x) does not blow up!
    # Note that this does not change the output.
    x_max = np.max(x, axis=axis, keepdims=True)
    x_safe = x - x_max
    e_x = np.exp(x_safe)
    return e_x / np.sum(e_x, axis=axis, keepdims=True)

def one_hot(labels):
    """this creates a one hot encoding from a flat vector:
    i.e. given y = [0,2,1]
     it creates y_one_hot = [[1,0,0], [0,0,1], [0,1,0]]
    """
    classes = np.unique(labels)
    n_classes = classes.size
    one_hot_labels = np.zeros(labels.shape + (n_classes,))
    for c in classes:
        one_hot_labels[labels == c, c] = 1
    return one_hot_labels

def unhot(one_hot_labels):
    """ Invert a one hot encoding, creating a flat vector """
    return np.argmax(one_hot_labels, axis=-1)

# then define an activation function class
class Activation(object):

    def __init__(self, tname):
        if tname == 'sigmoid':
            self.act = sigmoid
            self.act_d = sigmoid_d
        elif tname == 'tanh':
            self.act = tanh
            self.act_d = tanh_d
        elif tname == 'relu':
            self.act = relu
            self.act_d = relu_d
        else:
            raise ValueError('Invalid activation function.')

    def fprop(self, input):
        # we need to remember the last input
```

```python
        # so that we can calculate the derivative with respect
        # to it later on
        self.last_input = input
        return self.act(input)

    def bprop(self, output_grad):
        return output_grad * self.act_d(self.last_input)

# define a base class for layers
class Layer(object):

    def fprop(self, input):
        """ Calculate layer output for given input
            (forward propagation).
        """
        raise NotImplementedError('This is an interface class, please use a derived in

    def bprop(self, output_grad):
        """ Calculate input gradient and gradient
            with respect to weights and bias (backpropagation).
        """
        raise NotImplementedError('This is an interface class, please use a derived in

    def output_size(self):
        """ Calculate size of this layer's output.
        input_shape[0] is the number of samples in the input.
        input_shape[1:] is the shape of the feature.
        """
        raise NotImplementedError('This is an interface class, please use a derived in

# define a base class for loss outputs
# an output layer can then simply be derived
# from both Layer and Loss
class Loss(object):

    def loss(self, output, output_net):
        """ Calculate mean loss given real output and network output. """
        raise NotImplementedError('This is an interface class, please use a derived in

    def input_grad(self, output, output_net):
        """ Calculate input gradient real output and network output. """
        raise NotImplementedError('This is an interface class, please use a derived in

# define a base class for parameterized things
class Parameterized(object):

    def params(self):
        """ Return parameters (by reference) """
```

```
        raise NotImplementedError('This is an interface class, please use a derived in

    def grad_params(self):
        """ Return accumulated gradient with respect to params. """
        raise NotImplementedError('This is an interface class, please use a derived in

# define a container for providing input to the network
class InputLayer(Layer):

    def __init__(self, input_shape):
        if not isinstance(input_shape, tuple):
            raise ValueError("InputLayer requires input_shape as a tuple")
        self.input_shape = input_shape

    def output_size(self):
        return self.input_shape

    def fprop(self, input):
        return input

    def bprop(self, output_grad):
        return output_grad

class FullyConnectedLayer(Layer, Parameterized):
    """ A standard fully connected hidden layer, as discussed in the lecture.
    """

    def __init__(self, input_layer, num_units,
                 init_stddev, activation_fun=Activation('relu')):
        self.num_units = num_units
        self.activation_fun = activation_fun
        # the input shape will be of size (batch_size, num_units_prev)
        # where num_units_prev is the number of units in the input
        # (previous) layer
        self.input_shape = input_layer.output_size()
        # this is the weight matrix it should have shape: (num_units_prev, num_units)
        self.W = np.random.normal(0, init_stddev, (self.input_shape[1], self.num_units
        # and this is the bias vector of shape: (num_units)
        self.b = np.zeros(shape = (1, num_units)) #FIXME
        # create dummy variables for parameter gradients
        # no need to change these here!
        self.dW = None
        self.db = None

    def output_size(self):
        return (self.input_shape[0], self.num_units)

    def fprop(self, input):
```

```
        # you again want to cache the last_input for the bprop
        # implementation below!
        self.last_input = input
        a = np.dot(self.last_input, self.W) + self.b
        if self.activation_fun == None:
            h=a
        else:
            h = self.activation_fun.fprop(a)
        # NOTE: Use numpy dot product
        # h = t(a), a = W * x + b

        #FIXME
        #raise NotImplementedError("You should implement this!")
        return h

    def bprop(self, output_grad):
        """ Calculate input gradient (backpropagation). """

        # HINT: you may have to divide dW and db by n
        #       to make gradient checking work
        #       OR you divide the gradient in the output layer by n
        n = output_grad.shape[0]
        # accumulate gradient wrt. the parameters first
        # we will need to store these to later update
        # the network after a few forward backward passes
        # NOTE: you should also handle the case were
        #       activation_fun is None (meaning no activation)
        # the gradient wrt. W should be stored as self.dW
        # the gradient wrt. b should be stored as selfdb
        #raise NotImplementedError("you should implement this")
        # NOTE: self.dW is also a numpy dot product
        grad = output_grad

        if self.activation_fun is not None:
            grad = self.activation_fun.bprop(output_grad)

        self.dW = 1/n * np.dot(self.last_input.T, grad)
        self.db= 1/n  * np.sum(grad,  axis=0,  keepdims=True)

        # the gradient wrt. the input should be calculated here
        grad_input = np.dot(grad, self.W.T)
        return grad_input

    def params(self):
        return self.W, self.b

    def grad_params(self):
        return self.dW, self.db
```

```
# finally we specify the interface for output layers
# which are layers that also have a loss function
# we will implement two output layers:
#  a Linear, and Softmax (Logistic Regression) layer
# The difference between output layers and and normal
# layers is that they will be called to compute the gradient
# of the loss through input_grad(). bprop will never
# be called on them!
class LinearOutput(Layer, Loss):
    """ A simple linear output layer that
        uses a squared loss (e.g. should be used for regression)
    """
    def __init__(self, input_layer):
        self.input_size = input_layer.output_size()

    def output_size(self):
        return (1,)

    def fprop(self, input):
        return input

    def bprop(self, output_grad):
        raise NotImplementedError(
            'LinearOutput should only be used as the last layer of a Network'
            + ' bprop() should thus never be called on it!'
        )

    def input_grad(self, Y, Y_pred):
        return  (Y_pred - Y)

    def loss(self, Y, Y_pred):
        loss = 0.5 * np.square(Y_pred - Y)
        return np.mean(np.sum(loss, axis=1))

class SoftmaxOutput(Layer, Loss):
    """ A softmax output layer that calculates
        the negative log likelihood as loss
        and should be used for classification.
    """

    def __init__(self, input_layer):
        self.input_size = input_layer.output_size()

    def output_size(self):
        return (1,)

    def fprop(self, input):
```

```
        return softmax(input)

    def bprop(self, output_grad):
        raise NotImplementedError(
            'SoftmaxOutput should only be used as the last layer of a Network'
            + ' bprop() should thus never be called on it!'
        )

    def input_grad(self, Y, Y_pred):
        # TODO ######################################################
        # TODO: implement gradient of the negative log likelihood loss
        # TODO ######################################################
        # HINT: since this would involve taking the log
        #       of the softmax (which is np.exp(x)/np.sum(x, axis=1))
        #       this gradient computation can be simplified a lot,
        #       you may find a connection to the LinearOutput layer!
        return (Y_pred - Y)

    def loss(self, Y, Y_pred):
        # Assume one-hot encoding of Y
        out = Y_pred
        # to make the loss numerically stable
        # you should add an epsilon in the log ;)
        eps = 1e-10
        # TODO ####################################
        # calculate negative log likelihood
        # TODO ####################################
        loss =  np.sum(-np.log(out+eps)*Y, axis=1)#FIXME
        return np.mean(loss)
```

4 Neural Network class

With all layers in place (and properly implemented by you) we can finally define a neural network. For our purposes a neural network is simply a collection of layers which we will cycle through and on which we will call fprop and bprop to compute partial derivatives with respect to the input and the parameters.

Pay special attention to the *check_gradients()* function in which you should implement automatic differentiation. This function will become your best friend when checking the correctness of your implementation.

```
In [380]: class NeuralNetwork:
              """ Our Neural Network container class.
              """
              def __init__(self, layers):
                  self.layers = layers
```

```
def _loss(self, X, Y):
    Y_pred = self.predict(X)
    return self.layers[-1].loss(Y, Y_pred)

def predict(self, X):

    Yi_pred = X
    """ Calculate an output Y for the given input X. """
    for i in range(len(self.layers)):
        Yi_pred = self.layers[i].fprop(Yi_pred)

    Y_pred= Yi_pred
    return Y_pred

def backpropagate(self, Y, Y_pred, upto=0):
    """ Backpropagation of partial derivatives through
        the complete network up to layer 'upto'
    """
    next_grad = self.layers[-1].input_grad(Y, Y_pred)

    for i in reversed(range(upto, len(self.layers)-1)):
        next_grad = self.layers[i].bprop(next_grad)

    return next_grad

def classification_error(self, X, Y):
    """ Calculate error on the given data
        assuming they are classes that should be predicted.
    """
    Y_pred = unhot(self.predict(X))
    error = Y_pred != Y
    return np.mean(error)

def sgd_epoch(self, X, Y, learning_rate, batch_size):
    n_samples = X.shape[0]
    n_batches = n_samples // batch_size
    for b in range(n_batches):
        # start by extracting a batch from X and Y
        # (you can assume the inputs are already shuffled)
        x_batch=X[b*batch_size:(b+1)*batch_size]
        y_batch=Y[b*batch_size:(b+1)*batch_size]
        y_pred=self.predict(x_batch)
        final_grad=self.backpropagate(y_batch,  y_pred)
        for i in range(1,len(self.layers)-1):
            W,  b=self.layers[i].params()
            dW,  db=self.layers[i].grad_params()
            W-=learning_rate*dW
            b-=learning_rate*db
```

```
            # HINT: layer.params() returns parameters *by reference*
            #       so you can easily update in-place

    def gd_epoch(self, X, Y):
        # A few hints:
        #   There are two strategies you can follow:
        #   Either shove the whole dataset throught the network
        #   at once (which can be problematic for large datasets)
        #   or run through it batch wise as in the sgd approach
        #   and accumulate the gradients for all parameters as
        #   you go through the data. Either way you should then
        #   do one gradient step after you went through the
        #   complete dataset!
        y_pred=self.predict(X)
        final_grad=self.backpropagate(y_batch,  y_pred,  upto=1)
        for i in range(1,len(self.layers)-1):
            W,  b=self.layers[i].params()
            dW,  db=self.layers[i].grad_params()
            W-=learning_rate*dW
            b-=learning_rate*db

    def train(self, X, Y, X_valid, Y_valid, learning_rate=0.1, max_epochs=100, batch_s
              descent_type="sgd", y_one_hot=True):

        """ Train network on the given data. """
        n_samples = X.shape[0]
        n_batches = n_samples // batch_size
        if y_one_hot:
            Y_train = one_hot(Y)
            Y_valid_hot = one_hot(Y_valid)
        else:
            Y_train = Y
        print("... starting training")
        train_loss = np.zeros(max_epochs+1)
        train_error = np.zeros(max_epochs+1)
        valid_loss = np.zeros(max_epochs+1)
        valid_error = np.zeros(max_epochs+1)
        for e in range(max_epochs+1):
            if descent_type == "sgd":
                self.sgd_epoch(X, Y_train, learning_rate, batch_size)
            elif descent_type == "gd":
                self.gd_epoch(X, Y_train, learning_rate)
            else:
                raise NotImplementedError("Unknown gradient descent type {}".format(de

            # Output error on the training data
            train_loss[e] = self._loss(X, Y_train)
            train_error[e] = self.classification_error(X, Y)
```

```
            valid_loss[e]=self._loss(X_valid,  Y_valid_hot)
            valid_error[e]=self.classification_error(X_valid,  Y_valid)
            print('epoch {:.4f}, loss {:.4f}, train error {:.4f}'.format(e, train_loss
            print('epoch{:.4f},  loss{:.4f},  valid  error{:.4f}'.format(e,  valid_los
        return train_loss, valid_loss

    def check_gradients(self, X, Y):
        """ Helper function to test the parameter gradients for
        correctness. """
        for l, layer in enumerate(self.layers):
            if isinstance(layer, Parameterized):
                print('checking gradient for layer {}'.format(l))
                for p, param in enumerate(layer.params()):
                    # we iterate through all parameters
                    param_shape = param.shape
                    # define functions for conveniently swapping
                    # out parameters of this specific layer and
                    # computing loss and gradient with these
                    # changed parametrs
                    def output_given_params(param_new):
                        """ A function that will compute the output
                            of the network given a set of parameters
                        """
                        # copy provided parameters
                        param[:] = np.reshape(param_new, param_shape)
                        # return computed loss
                        return self._loss(X, Y)

                    def grad_given_params(param_new):
                        """A function that will compute the gradient
                           of the network given a set of parameters
                        """
                        # copy provided parameters
                        param[:] = np.reshape(param_new, param_shape)
                        # Forward propagation through the net
                        Y_pred = self.predict(X)
                        # Backpropagation of partial derivatives
                        self.backpropagate(Y, Y_pred, upto=l)
                        # return the computed gradient
                        return np.ravel(self.layers[l].grad_params()[p])

                    # let the initial parameters be the ones that
                    # are currently placed in the network and flatten them
                    # to a vector for convenient comparisons, printing etc.
                    param_init = np.ravel(np.copy(param))

                    # To debug you network's gradients use scipys
                    # gradient checking!
```

```
epsilon = 1e-5
import scipy.optimize
err = scipy.optimize.check_grad(output_given_params,
      grad_given_params, param_init)
print('diff scipy {:.2e}'.format(err))
assert(err < epsilon)

# reset the parameters to their initial values
param[:] = np.reshape(param_init, param_shape)
```

5 Gradient Checking

After implementing everything it is always a good idea to setup some layers and perform gradient checking on random data. **Note** that this is only an example! It is not a useful network architecture ;). We also expect you to play around with this to test all your implemented components.

```
In [381]: input_shape = (5, 10)
          n_labels = 6
          layers = [InputLayer(input_shape)]

          layers.append(FullyConnectedLayer(
                          layers[-1],
                          num_units=15,
                          init_stddev=0.1,
                          activation_fun=Activation('relu')
          ))
          layers.append(FullyConnectedLayer(
                          layers[-1],
                          num_units=6,
                          init_stddev=0.1,
                          activation_fun=Activation('tanh')
          ))
          layers.append(FullyConnectedLayer(
                          layers[-1],
                          num_units=n_labels,
                          init_stddev=0.1,
                          activation_fun=Activation('relu')
          ))
          layers.append(SoftmaxOutput(layers[-1]))
          nn = NeuralNetwork(layers)
```

```
In [382]: # create random data
          X = np.random.normal(size=input_shape)
          # and random labels
          Y = np.zeros((input_shape[0], n_labels))
          for i in range(Y.shape[0]):
              idx = np.random.randint(n_labels)
              Y[i, idx] = 1.
```

```
In [383]: nn.check_gradients(X, Y)
```

```
checking gradient for layer 1
diff scipy 1.21e-07
diff scipy 4.46e-08
checking gradient for layer 2
diff scipy 1.04e-07
diff scipy 2.98e-08
checking gradient for layer 3
diff scipy 5.76e-08
diff scipy 1.57e-08
```

6 Training on MNIST

Finally we can let our network run on the MNIST dataset!
First load the data and reshape it.

```
In [384]: # load
          Dtrain, Dval, Dtest = mnist()
          X_train, y_train = Dtrain
          X_valid, y_valid = Dval
          X_test, y_test = Dtest
```

```
... loading data
... done loading data
```

Dtrain contains 50k images which are of size 28 x 28 pixels. Hence:

```
In [385]: print("X_train shape: {}".format(np.shape(X_train)))
          print("y_train shape: {}".format(np.shape(y_train)))
```

```
X_train shape: (50000, 1, 28, 28)
y_train shape: (50000,)
```

y_train will automatically be converted in the *train()* function to one_hot encoding.
But we need to reshape X_train, as our Network expects flat vectors of size 28*28 as input!

```
In [386]: X_train = X_train.reshape(X_train.shape[0], -1)
          print("Reshaped X_train size: {}".format(X_train.shape))
          X_valid = X_valid.reshape((X_valid.shape[0], -1))
          print("Reshaped X_valid size: {}".format(X_valid.shape))
          X_test = X_test.reshape((X_test.shape[0], -1))
          print("Reshaped X_test size: {}".format(X_test.shape))
```

```
Reshaped X_train size: (50000, 784)
Reshaped X_valid size: (10000, 784)
Reshaped X_test size: (10000, 784)
```

Ah, much better ;-)!

Now we can finally really start training a Network!

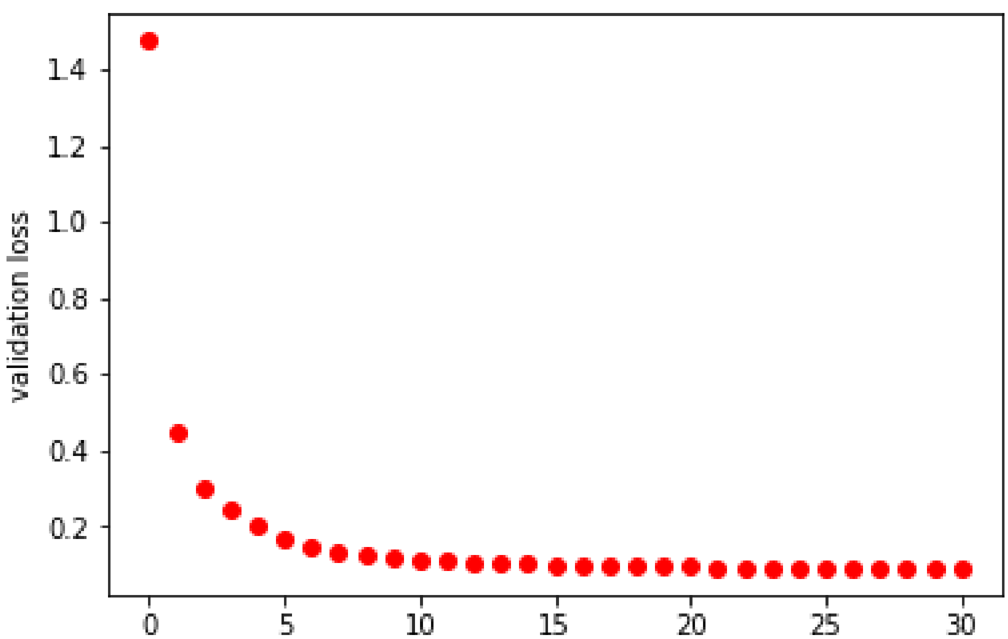
I pre-defined a small Network for you below. Again This is not really a good default and will not produce state of the art results. Please play around with this a bit. See how different activation functions and training procedures (gd / sgd) affect the result.

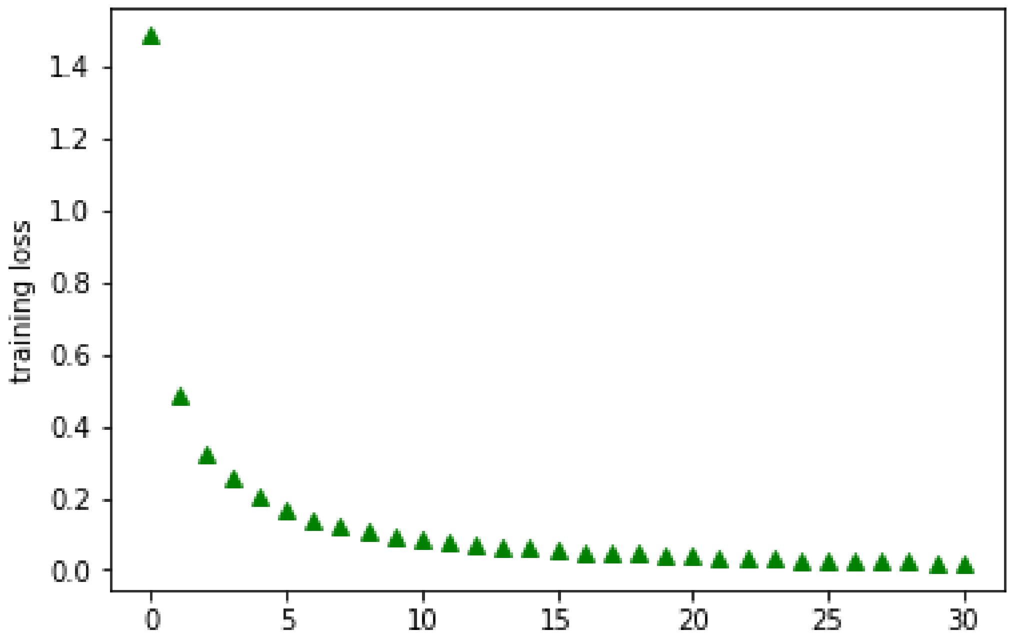
```
In [387]: import time

          # Setup a small MLP / Neural Network
          # we can set the first shape to None here to indicate that
          # we will input a variable number inputs to the network
          input_shape = (None, 28*28)
          layers = [InputLayer(input_shape)]
          layers.append(FullyConnectedLayer(
                          layers[-1],
                          num_units=128,
                          init_stddev=0.01,
                          activation_fun=Activation('relu')
          ))
          layers.append(FullyConnectedLayer(
                          layers[-1],
                          num_units=96,
                          init_stddev=0.01,
                          activation_fun=Activation('sigmoid')
          ))
          layers.append(FullyConnectedLayer(
                          layers[-1],
                          num_units=10,
                          init_stddev=0.01,
                          # last layer has no nonlinearity
                          # (softmax will be applied in the output layer)
                          activation_fun=None
          ))
          layers.append(SoftmaxOutput(layers[-1]))

          nn = NeuralNetwork(layers)
          # Train neural network
          t0 = time.time()

          train_loss, valid_loss = nn.train(X_train, y_train, X_valid, y_valid, learning_rate=0.
                   max_epochs=30, batch_size=64, y_one_hot=True)

          t1 = time.time()
          print('Duration: {:.1f}s'.format(t1-t0))

          error = nn.classification_error(X_test, y_test)*100
          print("Error = {:.3f}%".format(error))
```

```python
import matplotlib.pyplot as plt
plt.plot(range(31), valid_loss , 'ro')
plt.ylabel('validation loss')
plt.show()
plt.plot(range(31), train_loss , 'g^')
plt.ylabel('training loss')
plt.show()
```

```
... starting training
epoch 0.0000, loss 1.4880, train error 0.5340
epoch0.0000,  loss1.4752,  valid  error0.5250
epoch 1.0000, loss 0.4856, train error 0.1413
epoch1.0000,  loss0.4496,  valid  error0.1312
epoch 2.0000, loss 0.3230, train error 0.0976
epoch2.0000,  loss0.2999,  valid  error0.0884
epoch 3.0000, loss 0.2544, train error 0.0766
epoch3.0000,  loss0.2404,  valid  error0.0697
epoch 4.0000, loss 0.2046, train error 0.0614
epoch4.0000,  loss0.1981,  valid  error0.0568
epoch 5.0000, loss 0.1670, train error 0.0488
epoch5.0000,  loss0.1685,  valid  error0.0457
epoch 6.0000, loss 0.1402, train error 0.0410
epoch6.0000,  loss0.1488,  valid  error0.0413
epoch 7.0000, loss 0.1195, train error 0.0353
epoch7.0000,  loss0.1342,  valid  error0.0373
epoch 8.0000, loss 0.1042, train error 0.0299
epoch8.0000,  loss0.1243,  valid  error0.0356
epoch 9.0000, loss 0.0927, train error 0.0269
epoch9.0000,  loss0.1174,  valid  error0.0331
epoch 10.0000, loss 0.0835, train error 0.0242
epoch10.0000,  loss0.1122,  valid  error0.0319
epoch 11.0000, loss 0.0772, train error 0.0231
epoch11.0000,  loss0.1098,  valid  error0.0317
epoch 12.0000, loss 0.0703, train error 0.0209
epoch12.0000,  loss0.1060,  valid  error0.0311
epoch 13.0000, loss 0.0644, train error 0.0192
epoch13.0000,  loss0.1029,  valid  error0.0303
epoch 14.0000, loss 0.0597, train error 0.0179
epoch14.0000,  loss0.1008,  valid  error0.0294
epoch 15.0000, loss 0.0553, train error 0.0164
epoch15.0000,  loss0.0990,  valid  error0.0293
epoch 16.0000, loss 0.0504, train error 0.0147
epoch16.0000,  loss0.0965,  valid  error0.0281
epoch 17.0000, loss 0.0474, train error 0.0137
epoch17.0000,  loss0.0958,  valid  error0.0280
epoch 18.0000, loss 0.0443, train error 0.0127
epoch18.0000,  loss0.0949,  valid  error0.0275
epoch 19.0000, loss 0.0413, train error 0.0117
```

```
epoch19.0000,  loss0.0938,  valid  error0.0277
epoch 20.0000, loss 0.0383, train error 0.0110
epoch20.0000,  loss0.0932,  valid  error0.0278
epoch 21.0000, loss 0.0358, train error 0.0100
epoch21.0000,  loss0.0926,  valid  error0.0269
epoch 22.0000, loss 0.0331, train error 0.0091
epoch22.0000,  loss0.0919,  valid  error0.0264
epoch 23.0000, loss 0.0311, train error 0.0088
epoch23.0000,  loss0.0917,  valid  error0.0267
epoch 24.0000, loss 0.0289, train error 0.0082
epoch24.0000,  loss0.0913,  valid  error0.0257
epoch 25.0000, loss 0.0271, train error 0.0075
epoch25.0000,  loss0.0910,  valid  error0.0255
epoch 26.0000, loss 0.0253, train error 0.0067
epoch26.0000,  loss0.0910,  valid  error0.0254
epoch 27.0000, loss 0.0232, train error 0.0061
epoch27.0000,  loss0.0903,  valid  error0.0254
epoch 28.0000, loss 0.0218, train error 0.0056
epoch28.0000,  loss0.0903,  valid  error0.0252
epoch 29.0000, loss 0.0200, train error 0.0049
epoch29.0000,  loss0.0898,  valid  error0.0247
epoch 30.0000, loss 0.0186, train error 0.0045
epoch30.0000,  loss0.0900,  valid  error0.0244
Duration: 176.0s
Error = 2.680%
```





7 Figure out a reasonable Network that achieves good performance

As the last part of this task, setup a network that works well and gets reasonable accuracy, say ~ 1-3 percent error on the **validation set**. Train this network on the complete data and compute the **test error**.

Visualize the validation loss and training loss for each iteration in a plot, e.g. using matplotlib