

ASSIGNMENT 5, due Thurs, April 12

Discussions are encouraged but no written records can be taken away from a discussion. Books and notes can be consulted but not copied from. Homeworks are due in class or in my mail box by 4pm on the due date. For experimental problems, debugging help may be obtained but all code should be your own and all results reported should be from running your own code. In order to verify “curious” results, we may request you submit your code so DO NOT delete your code after you hand in a problem set.

The purpose of this assignment is to investigate experimentally some of the techniques described in class. Since there is randomness in the choice of data sets and weights, your results may vary considerably from your neighbor's. We will use the same two data sets (train and test) that were used in Assignment 2. The neural network architecture will be the same as before (one hidden layer), with 10 hidden units ($d^1 = 10$) with *tanh* for the output activation function. You are strongly urged to start on this problem set EARLY! Note that the learning rate and number of training iterations varies from problem to problem, so pay careful attention. In all cases, the empirical risk will be the same as in Assignment 2.

1 (300pts) Early Stopping with a Validation Set

- (a) Sample 40 data points randomly from the training set. Select a random initial set of weights as in Assignment 2. From these 40 data points, sample 5 data points (the validation set) randomly. Train on the remaining 35 data points for 4000 iterations using $\eta = 0.2$. As training proceeds, keep track of the error on training set (size 35), the validation set (size 5) and test set (size 1000). Provide a plot of the training, validation and test error versus iteration number.
- (b) Perform the following experiment. Sample 40 data points randomly from the training set. Select a random initial set of weights $\mathbf{w}(0)$ and train for 4000 iterations with $\eta = 0.2$ and obtain the error on the test set (the set of size 1000) after all the training is done. Using the **same** set of 40 points, split it into a validation set and training set as in (a). Now train using the **same** initial weight $\mathbf{w}(0)$ for 4000 iterations with $\eta = 0.2$. When training is done, select the hypothesis (set of weights) that had lowest error on the validation set and obtain the test error for this hypothesis.
 - i. Why do we use the same set of 40 points and the same initial weights?
 - ii. Articulate exactly what the two test errors you obtained signify.
- (c) Repeat the entire experiment in (b) 20 times to obtain 20 test errors with and without early stopping. Provide a plot of these 20 errors for both cases, and, also report the average of the 20 errors for both cases.
 - i. Articulate exactly what these two averages signify?
 - ii. Why do we rather report the average instead of the numbers obtained in (b)?
- (d) Repeat (b) and (c), changing only the validation set size. Suggested validation set sizes are 2, 4, 5, 6, 8, 10, 20, 30, but you may try additional sizes as well. For each validation set size, you have 20 test performances with and without validation. One can thus compute the performance gain due to validation, for each validation set size. Provide a plot of this performance gain versus validation set size. Give an explanation for the behavior of your curve.
- (e) Describe a method for obtaining the best validation set size that could be used in practice. (Remember in practice you do not have access to a test set, all you have is a training set)

2 (300pts) Weight Decay

Using weight decay, one minimizes the error function

$$E(\mathbf{w}_{ij}^l) = R_{emp}(\mathbf{w}_{ij}^l) + \lambda \sum_{l,i,j} (w_{ij}^l)^2$$

Remember that practically speaking, this just changes the weight update step for each weight w_{ij}^l by adding a term $-2\eta\lambda w_{ij}^l$.

- (a) Sample 30 data points randomly from the training set. Select a random set of weights, $\mathbf{w}(0)$. Starting from this same set of initial weights each time, train using weight decay parameters (λ) of $10^{-2} \times [0, 0.04, 0.08, 0.12, 0.16, 0.20, 0.24, 0.28, 0.32, 0.36, 0.4]$
Train for 4000 iterations using $\eta = 0.1$. Obtain the training and test errors for each λ at the *end* of training.
 - i. Why do we use the same set of initial weights for each lambda?
 - ii. Articulate exactly what the test errors you have obtained signify.
- (b) Repeat (a) 20 times to obtain 20 training and test errors for each λ . Plot the average of the training errors and the average of the test errors (for each λ) as function of λ .
 - i. Why do we rather use the average instead of the numbers obtained in (a)?
 - ii. Articulate exactly what each point on each curve signifies.
 - iii. Give an explanation for the observed behavior.
- (c) Describe a method for obtaining the best weight decay parameter that could be used in practice.

3 (400pts) Variable Learning Rate, Momentum, Steepest Descent

For this problem use all the 100 training data points.

- (a) Select an initial set of weights randomly. Start with a learning rate $\eta = 0.01$ and implement the variable learning rate gradient descent with increment parameter 1.03 and decrement parameter 0.8 for 4000 iterations. Compare with the case of static learning rate $\eta = 0.01$ starting from the same set of initial weights. – i.e. show a plot of training error vs. iteration number for both cases on the same plot.
- (b) Implement gradient descent with momentum parameters of $[0, 0.3, 0.6, 0.9, 0.95]$ and plot the training error as a function of iteration number for these four cases on the same plot. Each time start at the same set of initial weights (selected randomly) and train for 2000 iterations with $\eta = 0.1$.
- (c) Implement steepest descent using 3 quadratic interpolation searches for the line search minimum. Compare this with simple gradient descent. Each time start at the same set of initial weights (initially selected at random) and train for 4000 iterations with $\eta = 0.1$.
- (d) When comparing methods, why do we always start at the same set of initial weights?

4 (150pts Extra Credit) Bagging Using Bootstrap

- (a) Describe an experiment that you could do to determine whether Bagging using 10 Bootstrapped sets, as discussed in class, would be effective for this learning problem, if data set size were 15.
- (b) Implement your experiment for an architecture with 2 hidden units training for 500 iterations of steepest descent with 3 quadratic interpolations, and show the results.