Benjamin Ye

CS/CNS/EE 156a: Learning Systems (Fall 2023)

November 6, 2023

**Homework 6**

|  |  |
| --- | --- |
| **Problem** | **Answer** |
| 1 | [b] |
| 2 | [a] |
| 3 | [d] |
| 4 | [e] |
| 5 | [d] |
| 6 | [b] |
| 7 | [c] |
| 8 | [d] |
| 9 | [a] |
| 10 | [e] |

**Overfitting and Deterministic Noise**

1. Deterministic noise depends on , as some models approximate better than others. Assume that and that is fixed. In general (but not necessarily in all cases), if we use instead of , how does deterministic noise behave?

**Answer: [b] In general, deterministic noise will increase.**

In general, there are fewer hypotheses in than because is a subset of . As such, we expect that will model less of the target function correctly in the input space . Therefore, the deterministic noise (or the bias) is should increase since it is higher for smaller models.

**Regularization with Weight Decay**

In the following problems, use the data provided in the files

<http://work.caltech.edu/data/in.dta>

<http://work.caltech.edu/data/out.dta>

as a training and test set, respectively. Each line of the files corresponds to a two-dimensional input , so that , followed by the corresponding label from . We are going to apply linear regression with a non-linear transformation for classification. The nonlinear transformation is given by

Recall that the classification error is defined as the fraction of misclassified points.

1. Run linear regression on the training set after performing the non-linear transformation. What values are closest (in Euclidean distance) to the in-sample and out-of-sample classification errors, respectively?

**Answer: [a] 0.03, 0.08**

1. Now add weight decay to linear regression, that is, add the term to the squared in-sample error, using . What are the closest values to the in-sample and out-of-sample classification errors, respectively, for ? Recall that the solution for linear regression with weight decay was derived in class.

**Answer: [d] 0.03, 0.08**

1. Now, use . What are the closest values to the new in-sample and out-of-sample classification errors, respectively?

**Answer: [e] 0.4, 0.4**

1. What value of , among the following choices, achieves the smallest out-of-sample classification error?

**Answer: [d]**

1. What value is closest to the minimum out-of-sample classification error achieved by varying (limiting to integer values)?

**Answer: [b] 0.06**

See the next page for the linear regression results and a brief discussion.

The output used to answer problems 2–6 is

[HW6 P2–6]

Linear regression (without regularization) statistics:

**E\_in=0.029**, **E\_out=0.084**

Linear regression (with weight decay regularization using lambda=10^k) statistics:

k=-5: E\_in=0.029, E\_out=0.084

k=-4: E\_in=0.029, E\_out=0.084

k=-3: **E\_in=0.029**, E\_out=**0.080**

k=-2: E\_in=0.029, E\_out=0.084

**k=-1**: E\_in=0.029, **E\_out=0.056**

k=0: E\_in=0.000, E\_out=0.092

k=1: E\_in=0.057, E\_out=0.124

k=2: E\_in=0.200, E\_out=0.228

k=3: **E\_in=0.371**, **E\_out=0.436**

k=4: E\_in=0.429, E\_out=0.452

k=5: E\_in=0.429, E\_out=0.456

k=6: E\_in=0.429, E\_out=0.456

For problem 6, the out-of-sample classification error stops changing for and .

For , we get overfitting as , while for , we get underfitting as . The expected for the weight decay regularization should have a minimum in the range , which is exactly what we see for .

(The Python 3 source code is available on the following pages.)

import numpy as np

import requests

def validate\_binary(w, x, y):

    return np.count\_nonzero(np.sign(x @ w) != y, axis=0) / x.shape[0]

def linear\_regression(

        N=None, f=None, vf=None, \*, x=None, y=None, transform=None, noise=None,

        regularization=None, N\_test=1\_000, x\_test=None, y\_test=None, rng=None,

        seed=None, hyp=False, \*\*kwargs):

    if rng is None:

        rng = np.random.default\_rng(seed)

    if y is None:

        if x is None:

            x, y = generate\_data(N, f, bias=True, rng=rng)

        else:

            N = x.shape[0]

            y = f(x)

    if transform:

        x = transform(x)

    if noise:

        i = rng.choice(N, round(noise[0] \* N), False)

        y[i] = noise[1](y[i])

    if regularization is None:

        w = np.linalg.pinv(x) @ y

    elif regularization == "weight\_decay":

        w = np.linalg.inv(

            x.T @ x + kwargs["wd\_lambda"] \* np.eye(x.shape[1], dtype=float)

        ) @ x.T @ y

    if y\_test is None:

        if x\_test is None:

            x\_test, y\_test = generate\_data(N\_test, f, bias=True, rng=rng)

        else:

            N\_test = x\_test.shape[0]

            y\_test = f(x\_test)

    if transform:

        x\_test = transform(x\_test)

    if noise:

        i = rng.choice(N\_test, round(noise[0] \* N\_test), False)

        y\_test[i] = noise[1](y\_test[i])

    return (w, vf(w, x, y), vf(w, x\_test, y\_test))[1 - hyp:]

if \_\_name\_\_ == "\_\_main\_\_":

    (CWD / "data").mkdir(exist\_ok=True)

    data = {"train": "in.dta", "test": "out.dta"}

    for dataset, file in data.items():

        if not (CWD / "data" / file).exists():

            r = requests.get(f"http://work.caltech.edu/data/{file}")

            with open(CWD / "data" / file, "wb") as f:

                f.write(r.content)

        data[dataset] = np.loadtxt(CWD / "data" / file)

    transform = lambda x: np.hstack((

        np.ones((len(x), 1), dtype=float),

        x,

        x[:, :1] \*\* 2,

        x[:, 1:] \*\* 2,

        x[:, :1] \* x[:, 1:],

        np.abs(x[:, :1] - x[:, 1:]),

        np.abs(x[:, :1] + x[:, 1:])

    ))

    print("\n[HW6 P2–6]\nLinear regression (without regularization) "

        "statistics:")

    E\_in, E\_out = linear\_regression(

        vf=validate\_binary, x=data["train"][:, :-1], y=data["train"][:, -1],

        transform=transform, x\_test=data["test"][:, :-1],

y\_test=data["test"][:, -1]

    )

    print(f"  {E\_in=:.3f}, {E\_out=:.3f}")

    print("Linear regression (with weight decay regularization using "

          "lambda=10^k) statistics:")

    for k in (ks := np.arange(-5, 7)):

        E\_in, E\_out = linear\_regression(

            vf=validate\_binary, x=data["train"][:, :-1], y=data["train"][:, -1],

            transform=transform, regularization="weight\_decay",

wd\_lambda=10.0 \*\* k, x\_test=data["test"][:, :-1],

y\_test=data["test"][:, -1]

        )

        print(f"  {k=}: {E\_in=:.3f}, {E\_out=:.3f}")

**Regularization for Polynomials**

Polynomial models can be viewed as linear models in a space , under a nonlinear transform . Here, transforms the scalar into a vector of Legendre polynomials, . Our hypothesis set will be expressed as a linear combination of these polynomials,

where .

1. Consider the following hypothesis set defined by the constraint:

Which of the following statements is correct?

**Answer: [c]**

From the definition, the hypothesis set contains all Legendre polynomials of order and below. With the constraint, the set can lose higher-order polynomials if and .

For [a], has polynomials of order 2 and lower and is equivalent to , and has polynomials of order 3 and lower and is equivalent to . Therefore, their union is , not . This statement is false.

For [b], is plus higher-order polynomials with powers , while is plus higher-order polynomials with powers . Therefore, their union is some combination of and higher-order terms, not just . This statement is false.

For [c], is and is , as established earlier. Their intersection is simply . This statement is true (and the answer).

For [d], the intersection of and is plus higher-order polynomials with powers , not just . This statement is false.

**Neural Networks**

1. A fully connected neural network has ; , , . If only products of the form , , and count as operations (even for ), without counting anything else, which of the following is the closest to the total number of operations in a single iteration of backpropagation (using SGD on one data point)?

**Answer: [d] 45**

The first step is the forward propagation algorithm, in which the inputs and outputs of a hidden or output layer are related via

where is a transformation function and is a vector of incoming signals given by

By including the bias nodes, the number of times is evaluated across the layers is

Then, in the backpropagation algorithm, the SGD weights are updated using

where is the sensitivity of the error with respect to the input signal . As the gradient is evaluated with respect to each of the weights (of which there are 22 since each connection between nodes is a weight), there are multiplications of to be performed.

(Explanation continues on the following page…)

The sensitivity can be obtained by running a modified version of the neural network backwards, and is given by

With , which does not count as an operation for this problem, there are

operations to be carried out.

Putting everything together, we have a total of

multiplicative operations in a single iteration of backpropagation.

Let us call every “node” in a neural network a unit, whether that unit is an input variable or a neuron in one of the layers. Consider a neural network that has 10 input units (the constant is counted here as a unit), one output unit, and 36 hidden units (each is also counted as a unit). The hidden units can be arranged in any number of layers , and each layer is fully connected to the layer above it.

1. What is the minimum possible number of weights that such a network can have?

**Answer: [a] 46**

To minimize the number of weights, the network should have as few connections as possible. As such, the neural network should have 18 hidden layers with two nodes (including the bias node) each.

The number of dimensions in each layer is

where the input layer is (first value) and the output layer is (last value).

The total number of weights is

1. What is the maximum possible number of weights that such a network can have?

**Answer: [e] 510**

To maximize the number of weights, the network should have as many connections as possible. Based on the logic in problem 9, the neural network should have few hidden layers with many nodes.

For a single hidden layer with dimension (or 36 nodes), the total number of weights is

For a neural network with two hidden layers, the total number of weights is

The only constraint is that such that there are a total of 36 nodes. Substituting into ,

To maximize , we set its derivative with respect to to 0 and solve for :

Therefore, the first hidden layer should have dimension (or 22 nodes) and the second hidden layer should have dimension (or 14 nodes). In this scenario, the total number of weights is

As [e] 510 is the largest choice, it must be the answer. However, to be sure, we also check a neural network with three hidden layer:

With the domains and , Wolfram Mathematica reports for , , and . Indeed, the neural network with two hidden layers with 22 and 14 nodes, respectively, has the maximum number of weights.