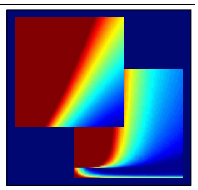
Learning From Data
Yaser Abu-Mostafa, *Caltech*
http://work.caltech.edu/telecourse
Self-paced version



Homework # 6

All questions have multiple-choice answers ([a], [b], [c], ...*). You can collaborate with others, but do not discuss the selected or excluded choices in the answers. You can consult books and notes, but not other people's solutions. Your solutions should be based on your own work. Definitions and notation follow the lectures.*

Note about the homework

- The goal of the homework is to facilitate a deeper understanding of the course material. The questions are not designed to be puzzles with catchy answers. They are meant to make you roll up your sleeves, face uncertainties, and approach the problem from different angles.
- The problems range from easy to difficult, and from practical to theoretical. Some problems require running a full experiment to arrive at the answer.
- The answer may not be obvious or numerically close to one of the choices, but one (and only one) choice will be correct if you follow the instructions precisely in each problem. You are encouraged to explore the problem further by experimenting with variations on these instructions, for the learning benefit.
- You are also encouraged to take part in the forum

  http://book.caltech.edu/bookforum

  where there are many threads about each homework set. We hope that you will contribute to the discussion as well. Please follow the forum guidelines for posting answers (see the "BEFORE posting answers" announcement at the top there).

• Overfitting and Deterministic Noise

1. Deterministic noise depends on $\mathcal{H}$, as some models approximate f better than others. Assume $\mathcal{H}' \subset \mathcal{H}$ and that f is fixed. **In general** (but not necessarily in all cases), if we use $\mathcal{H}'$ instead of $\mathcal{H}$, how does deterministic noise behave?

[**a**] In general, deterministic noise will decrease.

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

[**c**] In general, deterministic noise will be the same.

[**d**] There is deterministic noise for only one of $\mathcal{H}$ and $\mathcal{H}'$.

• 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 $\mathbf{x} = (x_1, x_2)$, so that $\mathcal{X} = \mathbb{R}^2$, followed by the corresponding label from $\mathcal{Y} = \{-1, 1\}$. We are going to apply Linear Regression with a non-linear transformation for classification. The nonlinear transformation is given by

$$\phi(x_1, x_2) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, |x_1 - x_2|, |x_1 + x_2|).$$

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

2. 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?

[**a**] 0.03, 0.08

[**b**] 0.03, 0.10

[**c**] 0.04, 0.09

[**d**] 0.04, 0.11

[**e**] 0.05, 0.10

3. Now add weight decay to Linear Regression, that is, add the term $\frac{\lambda}{N} \sum_{i=0}^{7} w_i^2$ to the squared in-sample error, using $\lambda = 10^k$. What are the closest values to the in-sample and out-of-sample classification errors, respectively, for $k = -3$? Recall that the solution for Linear Regression with Weight Decay was derived in class.

[a] 0.01, 0.02

[b] 0.02, 0.04

[c] 0.02, 0.06

[d] 0.03, 0.08

[e] 0.03, 0.10

4. Now, use $k = 3$. What are the closest values to the new in-sample and out-of-sample classification errors, respectively?

[a] 0.2, 0.2

[b] 0.2, 0.3

[c] 0.3, 0.3

[d] 0.3, 0.4

[e] 0.4, 0.4

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

[a] 2

[b] 1

[c] 0

[d] -1

[e] -2

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

[a] 0.04

[b] 0.06

[c] 0.08

[d] 0.10

[e] 0.12

• Regularization for Polynomials

Polynomial models can be viewed as linear models in a space $\mathcal{Z}$, under a nonlinear transform $\Phi : \mathcal{X} \to \mathcal{Z}$. Here, Φ transforms the scalar x into a vector $\mathbf{z}$ of Legendre

polynomials, $\mathbf{z} = (1, L_1(x), L_2(x), ..., L_Q(x))$. Our hypothesis set will be expressed as a linear combination of these polynomials,

$$\mathcal{H}_Q = \left\{ h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} = \sum_{q=0}^{Q} w_q L_q(x) \right\},$$

where $L_0(x) = 1$.

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

$$\mathcal{H}(Q, C, Q_o) = \{h \mid h(x) = \mathbf{w}^{\mathrm{T}}\mathbf{z} \in \mathcal{H}_Q; w_q = C \text{ for } q \geq Q_o\},$$

which of the following statements is correct:

[a] $\mathcal{H}(10, 0, 3) \cup \mathcal{H}(10, 0, 4) = \mathcal{H}_4$

[b] $\mathcal{H}(10, 1, 3) \cup \mathcal{H}(10, 1, 4) = \mathcal{H}_3$

[c] $\mathcal{H}(10, 0, 3) \cap \mathcal{H}(10, 0, 4) = \mathcal{H}_2$

[d] $\mathcal{H}(10, 1, 3) \cap \mathcal{H}(10, 1, 4) = \mathcal{H}_1$

[e] None of the above

• Neural Networks

8. A fully connected Neural Network has $L = 2$; $d^{(0)} = 5$, $d^{(1)} = 3$, $d^{(2)} = 1$. If only products of the form $w_{ij}^{(l)} x_i^{(l-1)}$, $w_{ij}^{(l)} \delta_j^{(l)}$, and $x_i^{(l-1)} \delta_j^{(l)}$ count as operations (even for $x_0^{(l-1)} = 1$), 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)?

[a] 30

[b] 35

[c] 40

[d] 45

[e] 50

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 $x_0^{(0)}$ is counted here as a unit), one output unit, and 36 hidden units (each $x_0^{(l)}$ is also counted as a unit). The hidden units can be arranged in any number of layers $l = 1, \cdots, L-1$, and each layer is fully connected to the layer above it.

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

[**a**] 46

[**b**] 47

[**c**] 56

[**d**] 57

[**e**] 58

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

[**a**] 386

[**b**] 493

[**c**] 494

[**d**] 509

[**e**] 510