Machine Learning: Homework #6

Due on Monday, November 6, 2017

Abu-Mostafa 3:00pm

Andrew Watson

Contents

Overfitting and Deterministic Noise 3

Problem 1 3

Regularization with Weight Decay 3

Problem 2 3

Problem 3 4

Problem 4 4

Problem 5 4

Problem 6 4

Regularization for Polynomials 5

Problem 7 5

Neural Networks 5

Problem 8 5

Problem 9 5

Problem 10 6

Overfitting and Deterministic Noise

Problem 1

Deterministic noise depends on $\mathcal{H}$, as some models approximate f better than others. Assume that $\mathcal{H}' \subset \mathcal{H}$ 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 $\mathcal{H}'$.

Let $(h')^*(x) \in \mathcal{H}'$ be the best possible approximation of $f(x)$ in $\mathcal{H}'$, that is $\|(f - (h')^*)(x)\|$ is as small as possible. However because $\mathcal{H}' \subset \mathcal{H}$, then $(h')^* \in \mathcal{H}$ and if we define $h'(x)$ to be the best possible approximation of f in $\mathcal{H}$, then $\|(h' - f)(x)\| \leq \|((h')^* - f)(x)\|$. Therefore, in general deterministic noise will monotonically increase if we use $\mathcal{H}'$.

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.

Problem 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

For details, see the Jupyter notebook containing the code for this section. Below are the in-sample and out-of-sample errors for several different choices of the weight decay parameter $\lambda = 10^k$.

```
Without regularization
E_in = 0.029 E_out = 0.084
With regularization parameter k = -3
E_in = 0.029 E_out = 0.080
With regularization parameter k = -2
E_in = 0.029 E_out = 0.084
With regularization parameter k = -1
E_in = 0.029 E_out = 0.056
With regularization parameter k = 0
E_in = 0.000 E_out = 0.092
With regularization parameter k = 1
E_in = 0.057 E_out = 0.124
With regularization parameter k = 2
E_in = 0.200 E_out = 0.228
With regularization parameter k = 3
E_in = 0.371 E_out = 0.436
```

Problem 3

Now add weight decay to Linear Reression ,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.

See above

Problem 4

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

See above

Problem 5

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

See above

Problem 6

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

Problem 6 continued on next page. . .

See above

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

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

where $L_0(x) = 1$.

Problem 7

Consider the following hypothesis set defined by the constraint:

$$\mathcal{H}(Q, C, Q_0) = \{h \mid h(x) = \mathbf{w}^T\mathbf{z} \in \mathcal{H}_Q; w_q = C \text{ for } q \geq Q_0\},$$

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

$\mathcal{H}(10,0,3)$ is the set of all polynomials up to degree 2 ($\mathcal{H}_2$), and similarly $\mathcal{H}(10,0,4)$ is the set of all polynomials up to degree 3 ($\mathcal{H}_3$). Because $\mathcal{H}_2 \subset \mathcal{H}_3$, $\mathcal{H}_2 \cap \mathcal{H}_3 = \mathcal{H}_2$. Therefore the answer is (c).

Neural Networks

Problem 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)?

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 $l = 1, \dots, L-1$, and each layer is fully connected to the layer above it.

Problem 9

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

Problem 10

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