Chapter 4: Overfitting

Section 4.1: When does overfitting occur?

Section 4.2: Regularization

Definition 1. Optimization with a *soft order constraint* is defined to be

$$g = \underset{h \in \mathcal{H}}{\arg\min}\, E_{\text{in}}(h) \qquad \text{subject to} \qquad \Omega(h) \leq C \tag{1}$$

where $\Omega : \mathcal{H} \to \mathbb{R}$ is a *regularization function* that penalizes "complex" hypotheses, and $C : \mathbb{R}$ is a hyperparameter that determines how complex a function is allowed to be.

Example 1. The *L1 regularizer* is defined to be

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_1. \tag{2}$$

Example 2. The *L2 regularizer* is defined to be

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2. \tag{3}$$

Example 3. The *elastic net* is defined to be

$$\Omega(\mathbf{w}) = \alpha\|\mathbf{w}\|_2^2 + (1 - \alpha)\|\mathbf{w}\|_1 \tag{4}$$

where $\alpha \in [0, 1]$.

Definition 2. Define the *augmented error* to be

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \lambda\Omega(h). \tag{5}$$

Then the *augmented error minimization* problem is

$$g = \operatorname*{arg\,min}_{h \in \mathcal{H}} E_{\text{aug}}(h). \tag{6}$$

Theorem 1. If $\lambda = \Theta(\frac{1}{C})$, then under reasonable conditions, optimizing the augmented error in Eq (6) is equivalent to optimizing the soft order constraint in Eq (1).

Fact 1. If $\Omega(h) \approx |E_{\text{in}}(h) - E_{\text{out}}(h)|$, then $E_{\text{aug}} \approx E_{\text{out}}$, and $g \approx f$.

	Effect of $\lambda \uparrow$ or $C \downarrow$						
Regularizer	E_{in}	$\lvert E_{\text{out}} - E_{\text{in}} \rvert$	E_{out}	VCdim	nnz($\mathbf{w}$)	time/iter	num iter
L1							
L2							
Elastic Net							

Fact 2. Consider the hypothesis class of k-sparse halfspaces

$$\mathcal{H}_k = \left\{ \mathbf{x} \mapsto \text{sign}(\mathbf{w}^T\mathbf{x}) \;:\; \mathbf{w} \in \mathbb{R}^d,\; \text{nnz}(\mathbf{w}) = k \right\}. \tag{7}$$

The VC-dimension of $\mathcal{H}_k$ is $\Theta(k \log(d/k))$. This hypothesis class cannot be optimized directly using any of the gradient descent algorithms because it is non-convex.

Problem 1. You are training a logistic regression model with N training data points of dimension d. Do you expect the optimal value of C and λ to be large or small in the following circumstances?

1. $N > d$.

2. $N < d$.

Problem 2. You are training a logistic regression model with N training data points of dimension d using stochastic gradient descent. How will the optimal values of C and λ change (increase/decrease/stay constant) in the following circumstances?

1. You double the number of data points in the training dataset.

2. You double the number of data points in the test dataset.

3. You apply the 3rd degree polynomial kernel.

4. You apply the PCA kernel.

5. You change the optimization algorithm from SGD to 2nd order gradient descent.

6. You double the number of dimensions of the training data.

Problem 3. In this problem you will derive a closed for solution to the ridge regression problem, which is closely related to the OLS regression problem. Recall that OLS uses the linear hypothesis class

$$\mathcal{H} = \left\{ \mathbf{x} \mapsto \mathbf{w}^T\mathbf{x} : \mathbf{w} \in \mathbb{R}^d \right\} \tag{8}$$

and the squared loss

$$\ell(\hat{y}, y) = (\hat{y} - y)^2 \tag{9}$$

so that the in-sample error is defined as

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2 = \|X\mathbf{w} - \mathbf{y}\|_2^2, \tag{10}$$

where X is a $N \times d$ matrix with ith row equal to $\mathbf{x}_i$ and $\mathbf{y}$ is the d dimensional vector with ith position equal to y_i. Finally, we computed the parameters for the OLS model by solving the equation

$$\hat{\mathbf{w}}^{\text{OLS}} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \|X\mathbf{w} - \mathbf{y}\|_2^2. \tag{11}$$

The ridge regression model modifies Equation (11) above by adding L2 regularization to get

$$\hat{\mathbf{w}}^{\text{ridge}} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2. \tag{12}$$

Your task is to derive a closed form solution for $\mathbf{w}^{\text{ridge}}$.

NOTE: This is a common interview question, and something that you should be able to do "without thinking". I happen to think it's a bad interview question in the sense that it doesn't directly measure what you'll be doing on the job (you're job isn't calculus). That said, the ability to solve this problem correlates pretty highly with having a detailed mathematical understanding of machine learning concepts that are practical on the job, and so "lazy" interviewers will ask this to get a sense of your math abilities.

Problem 4 (optional)**.** Problems 4.8 and 4.17 in the textbook provide useful insights that commonly show up in technical interviews. These problems are “optional” in the sense that they will not show up on your midterm exam, but I still strongly recommend you at least look at them.