Chapter 4.3: Validation / Model Selection

Model selection is the processes of selecting the best hyperparameters for a learning problem. It is the most important step in real-world data mining tasks.

One of the weaknesses of our textbook is that it does not discuss in detail very many models to select hyper-parameters for. So in order to have more interesting models to work with, we spent the last few lectures covering the "Model Zoo" and "Transfer Learning" topics. Now we return to Chapter 4.3 of the textbook. Previously we discussed Chapters 4.1-4.2 of the textbook on how regularization relates to overfitting.

Section 4.3: Validation (+Review)

The motivating equation of Chapter 4 is

$$E_{\text{out}}(h) = E_{\text{in}}(h) + \text{overfit penalty}. \tag{1}$$

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

$$g = \arg\min_{h \in \mathcal{H}} E_{\text{in}}(h) \qquad \text{subject to} \qquad \Omega(h) \leq C \tag{2}$$

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.

Definition 2. Define the *augmented error* to be

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \lambda \Omega(h). \tag{3}$$

Then the *augmented error minimization* problem is

$$g = \arg\min_{h \in \mathcal{H}} E_{\text{aug}}(h). \tag{4}$$

Theorem 1. If $\lambda = \Theta(\frac{1}{C})$, then under reasonable conditions, optimizing the augmented error in Eq (4) is equivalent to optimizing the soft order constraint in Eq (2).

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

Section 4.3.1: The Validation Set

Illustrate the validation set notation below.

Equation 4.10 of the textbook states

$$E_{\text{out}}(g) \leq E_{\text{out}}(g^-) \leq E_{\text{val}}(g^-) + O\left(\sqrt{\frac{1}{K}}\right).$$

Describe the difference between a test set and validation set.

Section 4.3.2: Model Selection

Illustrate the model selection notation below.

Equation 4.12 from the textbook states

$$E_{\text{out}}(g_{m^*}) \leq E_{\text{out}}(g_{m^*}^-) \leq E_{\text{val}}(g_{m^*}^-) + O\left(\sqrt{\frac{\log M}{K}}\right).$$

In 2015, Baidu research (led by Andrew Ng) was accused of cheating on the ImageNet LSVRC challenge. The lead scientist on the project Dr. Wen Ru was fired from his position as "distinguished scientist at Baidu's Institute of Deep Learning."

Reference: `https://dswalter.github.io/machine-learnings-first-cheating-scandal.html`

Section 4.3.3: Cross Validation

The textbook introduces the following un-numbered equation describing the role of K:

$$E_{\text{out}}(g) \approx E_{\text{out}}(g^-) \approx E_{\text{val}}(g^-).$$

Describe cross validation below.

Consider the following two popular datasets:

1. The MNIST dataset has 60k data points. 10k of those data points are reserved for testing and 50k data points for training. It is traditional to randomly split the training set into a train/validation split with 45k and 5k data points.

2. The ImageNet dataset has 1.2 million data points. It is traditional to use 1 million of those data points for training, 50k for validation, and the remaining 150k for testing.

Cross validation is not traditionally used on these data sets. Why?

Problems

Problem 1. You are training a logistic regression model to determine whether an image contains a bee or an ant. You have about 400 labeled images that have been randomly split into a training set of size 250 and a validation set of size 150.

1. You decide to train two models, one where the inputs to your model are the features generated by the ResNet18 feature map and another where the inputs to your model are the ResNet50 feature map. Both feature maps generate the same number of features, but the ResNet50 feature map internally has 50 hidden layers.

   For both models, you train for M epochs with a step size of 10^{-3} and weight decay of 10^{-4}. You observe that E_{train} for the ResNet18 feature map is 0.18 and E_{train} for the ResNet50 feature map is 0.11. Based on this observation, you conclude that the model based on ResNet50 will likely have lower E_{out} and you decide to use this as your final model. You estimate E_{out} by calculating E_{val} for the ResNet50 model. What can you say about the relationship between E_{out} and E_{val}?

2. You decide to only use the ResNet18 feature map to train your model. You train the model for M epochs using stochastic gradient descent with a step size of 10^{-3} and weight decay of 10^{-4}. After each epoch, you evaluate the model on the validation set. Finally, you select the model that had the best validation error. What can you say about the relationship between E_{out} and E_{val}?