Machine Learning from Data Assignment 8

Greg Stewart

October 29, 2018

Exercise 4.3

Deterministic noise depends on H, as some models approximate f better than others.

(a) *Assume H fixed and the complexity of f increased. Will deterministic noise in general go up or down? Is there a higher or lower tendency to overfit?*

Deterministic noise will go up as the final hypothesis from H is able to model less of the target function f. Likewise, the tendency to overfit increases.

(b) *Assume f fixed and complexity of H increased. Will deterministic noise go up or down? Is there a higher or lower tendency to overfit?*

Decreasing the complexity of H will also increase deterministic noise as the simpler hypothesis cannot model f as well. However, overfitting will decrease.

Exercise 4.5

A more general soft constraint is the Tikhonov regularization constraint

$$\mathbf{w}^T\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{w} \leq C$$

which captures the relationship between the w_i—the matrix Γ is the Tikhonov regularizer.

(a) *What should Γ be to obtain the constraint $\sum_{q=0}^{Q} w_q^2 \leq C$?*

$\mathbf{\Gamma} = \mathbf{I}$, the identity matrix, to get this constraint.

(b) *What should Γ be to obtain the constraint $(\sum_{q=0}^{Q} w_q)^2 \leq C$.*

Γ can simply be a row of ones in this case.

Exercise 4.6

We've seen both hard-order constraint and soft-order constraint. Which do you expect to be more useful for binary classification using the perceptron model?

The soft order constraint holds the potential to obtain a good fit for classification with the additional benefit of lower out-of-sample error, so I would expect it to be more useful.

Exercise 4.7

Fix g^-, learned from D_{train}, and define $\sigma_{val}^2 = Var_{D_{val}}[E_{val}(g^-)]$. We consider how σ_{val}^2 depends on K. Let

$$\sigma^2(g^-) = Var_{\mathbf{x}}[e(g^-(\mathbf{x}), y)]$$

be the pointwise variance in the out-of- sample error of g^-.

(a) *Show* $\sigma_{val}^2 = \frac{1}{K}\sigma^2(g^-)$.

$E_{val}(g^-)$ is defined as the sum over D_{val} of $e(g^-(\mathbf{x}), y)$. There are K points in the validation data set, and for each $\mathbf{x}$ we have the variance given in the problem description. Thus for the validation set we have

$$\begin{aligned}\sigma_{val} &= \frac{1}{K} Var_{\mathbf{x}}[e(g^-(\mathbf{x}), y)] \qquad \text{for } x \in D_{val} \\ &= \frac{1}{K}\sigma^2(g^-)\end{aligned}$$

(b) *In classification problem, where* $e(g^-(\mathbf{x}), y) = [\![g^-(\mathbf{x}) \neq y]\!]$, *express* σ_{val}^2 *in terms of* $\mathbb{P}[g^-(\mathbf{x}) \neq y]$.

Let $\mathbb{P}[g^-(\mathbf{x}) \neq y] = p$ From the definition shown in (a), we have in this case that

$$\sigma_{val}^2 = \frac{1}{K} Var_{\mathbf{x}}[\![g^-(\mathbf{x} \neq y]\!].$$

To calculate this we need $\mathbb{E}[E_{val}]$ and $\mathbb{E}[E_{val}^2]$.

$$\begin{aligned}\mathbb{E}[E_{val}] &= \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K}[\![g^-(\mathbf{x}_k) \neq y_k]\!]\right] \\ &= \mathbb{P}[g^-(\mathbf{x}) \neq y] \\ &= p\end{aligned}$$

$$\begin{aligned}\mathbb{E}[E_{val}^2] &= \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K}[\![g^-(\mathbf{x}_k) \neq y_k]\!]^2\right] \\ &= p\end{aligned}$$

because the pointwise error is either 0 or 1, both of which are unchanged when squared. So, for variance, we get

$$\begin{aligned}\sigma_{val}^2 &= \frac{1}{K}\left(\mathbb{E}[E_{val}^2] - \mathbb{E}[E_{val}]^2\right) \\ &= \frac{1}{K}\left(p - p^2\right) \\ &= \frac{1}{K}(\mathbb{P}[g^-(\mathbf{x}) \neq y] - \mathbb{P}[g^-(\mathbf{x}) \neq y]^2)\end{aligned}$$

(c) *Show that for any* g^- *in a classification problem,* $\sigma_{val}^2 \leq \frac{1}{4K}$.

The maximum possible value for $\mathbb{P}[g^-(\mathbf{x}) \neq y]$ is $\frac{1}{2}$. Plugging in this value to the result in (b) gets us

$$\sigma_{val}^2 = \frac{1}{K}\left[\frac{1}{2} - \left(\frac{1}{2}\right)^2\right] = \frac{1}{4K}$$

As this is an upper bound, it means we must have that

$$\sigma_{val}^2 \leq \frac{1}{4K}.$$

(d) *Is there a uniform upper bound for $Var[E_{val}(g^-)]$ similar to (c) in the case of regression with squared error $e(g^-(\mathbf{x}), y) = (g^-(\mathbf{x}) - y)^2$?*

No, no upper bound exists for squared error.

(e) *For regression with squared error, if we train using fewer points (smaller $N - K$) to get g^-, do you expect $\sigma^2(g^-)$ to be higher or lower?*

Training with fewer points, I'd expect $\sigma^2(g-)$ to be **higher**.

(f) *Conclude that increasding the size of the validation set can result in a better or worse estimate of E_{out}.*

For the most part, increasing the size of the validation set only makes the estimate for E_{out} worse—there is no upper bound for squared error, which is by and large a more useful metric. Thus error will likely increase as the validation set is increased in size and the training set decreases in size.

Exercise 4.8

Is E_m an unbiased estimate for the out of sample error $E_{out}(g_m^-)$?

Yes, it's unbiased because no g_m^- has been picked yet.