EE 660 **Homework Week 10** Posted 10/31/2019

Due: Fri. 11/8/2019, 2:00 PM

1. For a regression problem with 2 features, consider the effect of different regularizers and different amounts of regularization, graphically as described below. You may do this by hand, or you may use a computer to assist you if you prefer.

   Assume the unconstrained objective function is $f_{obj}(\underline{w}) = \frac{1}{N}\mathrm{RSS}(\underline{w}, \mathcal{D}_i)$. For simplicity, in this problem we assume $w_0 = 0$, consistent with a dataset that has been standardized in both x and y. Consider 10 different datasets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_{10}$, each resulting in an unconstrained (unregularized) minimum at $\hat{\underline{w}}_{\text{lin}}^{(i)}$ given by:

$$\mathcal{D}_1 : \hat{\underline{w}}_{\text{lin}}^{(1)} = (10,0), \quad \mathcal{D}_2 : \hat{\underline{w}}_{\text{lin}}^{(2)} = (10,2), \quad \mathcal{D}_3 : \hat{\underline{w}}_{\text{lin}}^{(3)} = (10,4),$$
$$\mathcal{D}_4 : \hat{\underline{w}}_{\text{lin}}^{(4)} = (10,6), \quad \mathcal{D}_5 : \hat{\underline{w}}_{\text{lin}}^{(5)} = (8,6), \quad \mathcal{D}_6 : \hat{\underline{w}}_{\text{lin}}^{(6)} = (8,8),$$
$$\mathcal{D}_7 : \hat{\underline{w}}_{\text{lin}}^{(7)} = (6,8), \quad \mathcal{D}_8 : \hat{\underline{w}}_{\text{lin}}^{(8)} = (6,10), \quad \mathcal{D}_9 : \hat{\underline{w}}_{\text{lin}}^{(9)} = (4,10), \quad \mathcal{D}_{10} : \hat{\underline{w}}_{\text{lin}}^{(10)} = (2,10)$$

   Assume the shape of $\mathrm{RSS}(\underline{w}, \mathcal{D}_i) = \text{constant}$ curves in 2D weight space are circles (special case of ellipses), for simplicity.

   In each regularizer case given below, make a plot in 2D weight space, showing:

   (i) the 10 unregularized-minimum points $\hat{\underline{w}}_{\text{lin}}^{(i)}$ given above,

   (ii) the region that satisfies the given regularizer constraint, and

   (iii) the resulting 10 regularized minimum points, i.e., solution of

$$\hat{\underline{w}}_{\text{reg}}^{(i)} = \underset{\underline{w}}{\arg\min}\, f_{obj}(\underline{w}, \mathcal{D}_i) \quad \text{s.t.} \quad \Omega(\underline{w}) \le C \ .$$

   for each i. Also show or justify how you found the resulting $\hat{\underline{w}}_{\text{reg}}^{(i)}$. (Showing your method for one or two points in each regularizer case, should be sufficient.)

   (iv) Also, answer: how many of the resulting $\hat{\underline{w}}_{\text{reg}}^{(i)}$, $i = 1,2,\cdots,10$, are more sparse than the corresponding $\hat{\underline{w}}_{\text{lin}}^{(i)}$? For the purpose of this problem, define sparsity as the number of components $\hat{w}_j^{(i)}$ that have value 0, for a given i.

   Tip: For cases in which there are more than one possible $\hat{\underline{w}}_{\text{reg}}^{(i)}$ for a given dataset and a given constraint, pick any one.

(a) L2 regularization: $\Omega(\underline{w}) = \|\underline{w}\|_2^2$, $C = 2^2$.

(b) L1 regularization: $\Omega(\underline{w}) = \|\underline{w}\|_1$, $C = 2$.

(c) Lp regularization (based on p-norm): $\Omega(\underline{w}) = \|\underline{w}\|_p^p$, as $p \to \infty$, $C = 1$.

Hint: if you're not sure of the shape of $\|\underline{w}\|_p^p = 1$, try plotting it numerically for increasing p, e.g. $p = 4, 10, 100$.

(d) Repeat (a), except with $C = 5^2$.

(e) Repeat (b), except with $C = 5$.

2. Suppose you develop and optimize a machine learning system, starting with setting aside a test dataset $\mathcal{D}_{Test}$, and using the remaining data points as the set $\mathcal{D}'$. Your hypothesis set is $\mathcal{H}_1$, and you use $\mathcal{D}'$ as a training set to find its best hypothesis h_{g1}. Let $d_{VC}(\mathcal{H}_1) = d_{VC}^{(1)}$, $N' = |\mathcal{D}'|$, and $N_{Test} = |\mathcal{D}_{Test}|$. When you are finished, you pull out the test set and calculate $E_{Test}(h_{g1})$. In this problem, all generalization bounds are with tolerance δ (with probability $\geq 1-\delta$).

   (a) Draw a flow chart (like we did in Lecture 16, p. 6, and like AML Fig. 4.11), that shows the dataset usage, hypothesis set, and procedure.

   (b) Give an inequality for the generalization bound based on the training error $E_{\mathcal{D}'}(h_{g1})$, and the generalization bound based on the test-set error $E_{Test}(h_{g1})$.

Afterwards, independently of the results you got above, you think of a different approach that you also want to try. So you start the process all over again, setting aside the same test set $\mathcal{D}_{Test}$. You define a hypothesis set $\mathcal{H}_2$ for your model. Let $d_{VC}(\mathcal{H}_2) = d_{VC}^{(2)}$.

In this case, however, you also use some model selection to choose the optimum number of features in a feature selection process. So you split $\mathcal{D}'$ into a training set $\mathcal{D}_{Tr}$ and a validation set $\mathcal{D}_{Val}$, that are disjoint. You use $\mathcal{D}_{Tr}$ to train each model (based on a given number of features d), and use model selection to compare different values of d, with $d = 1,2,3,\cdots,d_{\max}$, in which $d_{\max}$ is the maximum number of features you try. You choose the best number of features by comparing $E_{Val}\left(h_{g2}^{(d)}\right)$ for each value of d. Let $N_{Tr} = |\mathcal{D}_{Tr}|$, and $N_{Val} = |\mathcal{D}_{Val}|$.

(c) Draw a flow chart (like we did in Lecture 16, p. 6, and like AML Fig. 4.11), that shows the dataset usage, hypothesis sets, parameter values d, and procedure, for this second approach only.

(d) Give:

(i) An inequality for the generalization bound based on the training-set error $E_{Tr}\left(h_{g2}^{(d)}\right)$ for a given number of features d;

(ii) An inequality for the generalization bound based on the validation-set error $E_{Val}\left(h_{g2}^{(d^*)}\right)$ for the optimal number of features d^*;

(iii) An inequality for the generalization bound based on the test-set error $E_{\mathcal{D}_{Test}}\left(h_{g2}^{(d^*)}\right)$ for the best hypothesis $h_{g2}^{(d^*)}$.

Finally, you compare the best results from the 2 systems you developed, and pick the one with the lower test-set error.

(e) Give an inequality for the generalization bound based on the test-set error $E_{Test}\left(h_g^*\right)$ for the best hypothesis h_g^*.

Hint: what is the effective hypothesis set used by $\mathcal{D}_{Test}$ to pick between the two machine-learning systems you developed?