IML 2020 - ex 5 - Validation, Feature Selection and Regularization

June 2020

Theoretical part

Validation

1. In this question we will see when the model selection paradigm has a benefit over the standard method when choosing between k possible hypothesis classes: $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots \subseteq \mathcal{H}_k$, where $\mathcal{H}_k$ is finite.

   Suppose we are given m examples $S_{all} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ and, as usual, we would like to learn a hypothesis with small generalization error. In class we discussed the polynomial fitting problem where we had k hypothesis classes to choose from $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots \subseteq \mathcal{H}_k$. In this question we compare two methods for choosing a hypothesis:

   - *Standard Method:* Find the best hypothesis in $\mathcal{H}_k$ using the ERM rule and using all the m training examples.
   - *Model Selection:* Do the following steps:
     - Divide the m examples into a training set S with size $(1-\alpha)m$ and a validation set V of size αm for some $\alpha \in (0, 1)$ (assume that αm is an integer).
     - For each hypothesis class $\mathcal{H}_i$, $i \in [k]$, find $h_i \in ERM_{\mathcal{H}_i}(S)$
     - Return $h^* \in ERM_{\mathcal{H}}(V)$, where $\mathcal{H} = \{h_1, \ldots, h_k\}$

   Assume $\mathcal{H}_k$ is finite and the loss function is bounded by 1.

   (a) Bound the generalization error using the standard method. Namely, prove that agnostically PAC learning $\mathcal{H}_k$ provides the following bound: for $h^* \in ERM_{\mathcal{H}_k}(S_{all})$, with probability at least $1 - \delta$,

   $$L_{\mathcal{D}}(h^*) \leqslant \min_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) + \sqrt{\frac{2\ln(2|\mathcal{H}_k|/\delta)}{m}}.$$

   Hint: Use Hoeffding and the union bound.

   (b) Bound the generalization error using model selection. Namely, suppose that $\operatorname{argmin}_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h)$ comes from $\mathcal{H}_j$ for some $j \in [k]$ (this implies that $\operatorname{argmin}_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) \in \mathcal{H}_{j+1}, \mathcal{H}_{j+2}, \ldots, \mathcal{H}_k$). Prove that with probability at least $1 - \delta$,

   $$L_{\mathcal{D}}(h^*) \leqslant \min_{h \in \mathcal{H}_k} L_D(h) + \sqrt{\frac{2}{\alpha m} \ln \frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m} \ln \frac{4|\mathcal{H}_j|}{\delta}}\ .$$

Hint: Use the previous item on the training step and on the validation step (switch in both cases $\delta \to \delta/2$), and recall that the probability of two independent events equals the product of their individual probabilities $P(A \cap B) = P(A)P(B)$.

(c) Show that the two bounds are incomparable: describe a case where the standard method is better and a case where model selection is better, in terms of the generalization error.

Orthogonal design

2. In the special case of an orthogonal design matrix (i.e $XX^T = I_d$), we can derive closed forms for best-subset, lasso and ridge regression solutions, as functions of the LS (least squares) solution. For $\lambda > 0$, let $\widehat{w}^{LS}$, $\widehat{w}_\lambda^{ridge}$, $\widehat{w}_\lambda^{lasso}$, $\widehat{w}_\lambda^{subset}$ denote the least squares (standard linear regression), Ridge, Lasso and Best Subset solutions, respectively. moreover, η_λ^{soft} and η_λ^{hard} are functions $\mathbb{R} \to \mathbb{R}$ as you seen in class and recitation. Prove the next two equality.

(a) $\widehat{w}_\lambda^{ridge} = \frac{\widehat{w}^{LS}}{1+\lambda}$

(b) $\hat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{hard}(\hat{w}^{LS})$

Regularization

3. In this section we will show that although Ridge adds bias to the estimation of $\mathbf{w}$ it still decreases the MSE.

Let X be a **constant** d-by-m regression matrix and assume that XX^T is invertible. Let $\hat{\mathbf{w}}(\lambda) = \text{argmin}_{\mathbf{w}} \left(||\mathbf{y} - X^T\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2\right)$ be the ridge estimator and let $\hat{\mathbf{w}}(\lambda = 0) \equiv \hat{\mathbf{w}}$ be the least squares estimator.

- Assume the linear model is correct, namely $\mathbf{y} = X^T\mathbf{w} + \epsilon$ where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
- Recall that in this case: $\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}$.

(a) Show that $\hat{\mathbf{w}}(\lambda) = A_\lambda \hat{\mathbf{w}}$ where

$$A_\lambda = (XX^T + \lambda I_d)^{-1}(XX^T)$$

(b) Conclude from the previous item that the Ridge estimator is biased for any $\lambda > 0$, i.e. show that $\lambda > 0 \Rightarrow \mathbb{E}[\hat{\mathbf{w}}(\lambda)] \neq \mathbf{w}$.

(c) Show that: $\text{Var}(\hat{\mathbf{w}}(\lambda)) = \sigma^2 A_\lambda (XX^T)^{-1} A_\lambda^T$.
Hint: Use the fact that for a non random matrix B and a random vector $\mathbf{z}$ we have $\text{Var}(B\mathbf{z}) = B\text{Var}(\mathbf{z})B^T$ and that $\text{Var}(\hat{\mathbf{w}}) = \sigma^2(XX^T)^{-1}$.

(d) Derive explicit expressions for the (squared) bias and variance of $\hat{\mathbf{w}}(\lambda)$ as a function of λ, i.e. write a bias-variance decomposition for the mean square error of $\hat{\mathbf{w}}(\lambda)$. Show by differentiation that

$$\frac{d}{d\lambda}\text{MSE}(\lambda)|_{\lambda=0} = \frac{d}{d\lambda}\text{bias}^2(\lambda)|_{\lambda=0} + \frac{d}{d\lambda}\text{Var}(\lambda)|_{\lambda=0} < 0$$

(e) Conclude that, if the linear model is correct, a little Ridge regularization helps to reduce the MSE.

Practical part

k-Fold Cross Validation on Polynomial Fitting

4. In this exercise we will perform polynomial fitting. Although we have seen this several times in the course, here the focus will be on the details of cross-validation.

   Both the domain and the label set are real scalars: $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$.

   The prior knowledge is that the relation between the instances and their labels can be approximately explained by a polynomial of degree $d \in [15] := \{1, \ldots, 15\}$. For each $d \in [15]$, let $\mathcal{H}_d$ be the class of polynomials of degree d. Your task is to train each of the classes over the training set and perform validation over the 15 resulting hypotheses in order to choose the final output. Finally, you will test the performance of the resulting predictor over the test set. Here are the exact details.

   (a) Generate a data set according to the following:

      i. $\mathcal{X} = [-3.2, 2.2]$ and x is uniformly sampled from this domain.

      ii. The relation between y and x is $y(x) = f(x) + \epsilon$ where

      $$f(x) = (x+3)(x+2)(x+1)(x-1)(x-2)$$

      and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

      iii. Take $\sigma = 1$ and generate $m = 1500$ instances.

      iv. Divide the $m = 1500$ instances into 2 sets: a 1000 for (training + validation) (call it D), and 500 for a test set (call it T) (you should not touch T until you reach item (g)).

   (b) Split D into two sets of 500 instances each - one for training (S) and one for validation (V): $D = S \cup V$. Train each of the classes using S to obtain a single hypothesis h_d for each $d \in [15]$, which minimizes the loss over S.

   (c) Perform validation over the set $\{h_1, \ldots, h_{15}\}$ to obtain a single h^* which minimizes the loss over the validation set V.

   (d) Items (b)-(c) are k-fold cross validation with $k = 2$ (well, almost - each data point was only used once, either for training *or* for validation). Perform k-fold cross validation but now with $k = 5$ on the set D, which contains 1000 examples.

   (e) Plot the training and validation errors (averaged over the k folds) of the polynomials of degree $d \in [15]$ as a function of d. Which has the lowest validation error? Denote this by d^*.

   (f) Perform $\text{ERM}_{\mathcal{H}_{d^*}}$ on the set D, namely find the polynomial of degree d^* which has a minimal error on these examples, and denote this polynomial by h^*.

   (g) Test the performance of h^* over the test sequence T. That is, calculate the test error of h^*. Is it very different from the error you found in the previous item?

   (h) Repeat all of the above steps for $\sigma = 5$. What has changed?

k-Fold and Regularization

5. In this exercise we will compare the Lasso and Ridge regularization. We will use the data of estimating the glucose levels of diabetes patients. You may read more about this data in this link.

   (a) Load the data using the following command:

   ```
   from sklearn import datasets
   X, y = datasets.load_diabetes(return_X_y=True)
   ```

   (b) Regularization is useful when our training set is small. Therefore, create the training set to be the first $m = 50$ samples, while the rest is left for the test set.

   (c) Perform the following twice: (a) for ridge regularization and (b) for lasso regularization:

      i. Run k-fold cross-validation over the training set for $k = 5$, using different values for the regularization parameter λ (explore the range of possible values for λ).

      ii. Explain in your handouts which values of λ you have chosen to examined by the validation and why.

   (d) Plot the training and validation errors (averaged over the k folds) as a function of λ over the range you've chosen for λ.

   (e) Find the best λ (that gives the lowest average error over the validation) for ridge regression and the best λ for lasso regression.

   (f) Calculate the test error of:

      i. the best ridge hypothesis

      ii. the best lasso hypothesis

      iii. linear regression (without regularization)

   (g) Which of the three has the smallest error? Did the regularization manage to improve the test results, comparing to the non-regulated case for this data?