CSE353 Assignment 6 (Nonlinear Transform, Overfitting/Underfitting, Regularization, and Cross-Validation)

Due Dec 07 2021 5:00PM

Main TA for this assignment:
Lihan Huang
Email: lihan.huang@stonybrook.edu
Office hours: Thursday 10AM-noon
Zoom: https://us02web.zoom.us/j/3165392097?pwd=WVJpeFVnQUhXSGFRSWROMEJ5S29FUT09

One **training data set** is provided:

'TrainingData_x.txt', where $\boldsymbol{x}$ ($N \times 1$) contains N training samples of feature dimension 1, i.e., $\boldsymbol{x} = [x_1; x_2; \dots; x_N]$;

"TrainingData_y.txt', where $\boldsymbol{y}$ ($N \times 1$) is a column vector denoting the output, i.e., $\boldsymbol{y} = [y_1, y_2, \dots, y_N]^T$;

Similarly, a **testing dataset is** provided in 'TestingData_x.txt' and 'TestingData_y.txt'.

1. Nonlinear transform and overfitting/underfitting

Q-th order polynomial transform: $\boldsymbol{z}_n = \phi_Q(x_n) = \left[1, x_n, x_n^2, x_n^3, \dots x_n^Q\right]^T$

Linear regression on a transformed dataset $\{(\boldsymbol{z}_n, y_n)\}_{n=1\dots N}$: $\boldsymbol{w}_{Poly} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{y}$, where $\boldsymbol{Z} = [\boldsymbol{z}_1^T; \boldsymbol{z}_2^T; \dots; \boldsymbol{z}_N^T]$; and the estimated $\boldsymbol{w}_{poly} = \left[w_0, w_1, \dots, w_Q\right]^T$.

Choose a Q (e.g., Q= 2,3, or 19) to perform the feature transform, then estimate $\boldsymbol{w}_{poly}$ from the transformed training dataset and calculate the training error (squared distance error, norm-2). Apply the estimated $\boldsymbol{w}_{poly}$ on the testing dataset with its transformed features and calculate the testing error.

$$err_{sqr} = \sum_{n=1}^{N}\left(\boldsymbol{w}_{poly}^T\phi(x_n) - y_n\right)^2$$

Plot the training & testing sample points, and the estimated curve (i.e., final hypothesis $g(x) = \boldsymbol{w}_{poly}^T\phi(x)$).

Report the training and testing errors.

Compare the results using different Q-values (Q = 2, 3, or 19), what do you observe? E.g., which is overfitting and which is underfitting?

2. Regularization

Apply the Q-th order polynomial transform to the feature of each sample, (i.e., $\boldsymbol{z}_n = \phi(x_n) = [1, x_n, x_n^2, x_n^3, \dots x_n^Q]^T$, then apply the <u>regularized</u> linear regression to dataset $\{(\boldsymbol{z}_n, y_n)\}_{n=1 \dots N}$ to estimate $\boldsymbol{w}_{regpoly} = [w_0, w_1, \dots, w_Q]^T$, i.e., obtain $\boldsymbol{w}_{regpoly}$ by solving $\min_{\boldsymbol{w}_{regpoly}} \left|\left|\boldsymbol{Z}\boldsymbol{w}_{regpoly} - \boldsymbol{y}\right|\right|^2 + \lambda \left|\left|\boldsymbol{w}_{regpoly}\right|\right|^2$, where $\boldsymbol{Z} = [\boldsymbol{z}_1^T; \boldsymbol{z}_2^T; \dots; \boldsymbol{z}_N^T]$ and $\boldsymbol{y} = [y_1, y_2, \dots y_N]^T$.

(1) Choose Q = 6 and $\lambda = 0.1$, estimate $\boldsymbol{w}_{regpoly}$ by the regularized linear regression and report the squared error cost on the training and testing datasets:

$$err_{sqr} = \sum_{n=1}^{N} \left(\boldsymbol{w}_{regpoly}^T \phi(x_n) - y_n\right)^2$$

Plot the training & testing sample points and the estimated curve (i.e., final hypothesis $g(x) = w_{regpoly}^T \phi(x)$).

(2) Choose Q = 6 and $\lambda = 100$, repeat the above steps.

Compare the two results, what do you observe?

3. Cross validation

The parameter λ is critical in the regularized linear regression. Apply the leave-one-out and v-fold (e.g., 5-fold) cross validation strategies to select the best λ from the pool of {0.01, 0.1, 1,10,100,1000,10^6}.

Upload your codes with enough comments and a brief report to Blackboard by the due date & time, including

a) Introduction. Brief summary of what you think the assignment is about,

b) Method. Brief outline of your (algorithmic) approach,

c) Experiments. Tables and/or pictures of intermediate and final results that convince us that the program does what you think it does.

d) Discussions and Conclusions. Any design decisions you had to make and your experimental observations. What do you observe about the behavior of your program when you run it? Does it seem to work the way you think it should? Play around a little with different setting to see what happens. Note, your open-ended exploration is highly valued.