CSCI 6971/4971: Homework 1

Assigned Thursday January 24 2019. Due at beginning of class Thursday January 31 2019

Remember to typeset your submission, and label it with your name. Coding should be done in Python+NumPy+Sklearn. Attach a printout of your code. Labels all plots appropriately and structure and comment your code neatly and appropriately.

2. [50 pts. Regularized Linear Regression Models with moderate training data.] Recall the two forms of regularization introduced in the class: ℓ_2 and ℓ_1. We will compare the effects of the choice of regularizer on a linear regression problem.

   Obtain the year prediction dataset from the UCI Machine Learning repository (`https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD`). Caveat: this is a >200MB download that unzips to an approximate 500MB text file.

   Write a script that does the following (throughout, feel free to use the functionalities present in Sklearn for e.g., cross-validation, rather than rolling your own):

   (a) Following the dataset description in the repository, load the first 100K training examples from the text file into the rows of a matrix $\mathbf{X}_{\text{train}}$, and the corresponding years into a vector $\mathbf{y}_{\text{train}}$.

   (b) Following the dataset description in the repository, load the first 20K testing examples from the text file into the rows of a matrix $\mathbf{X}_{\text{test}}$, and the corresponding years into a vector $\mathbf{y}_{\text{test}}$.

   (c) Compute the OLS estimate

   $$\boldsymbol{\beta}_{\text{OLS}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{X}_{\text{train}}\boldsymbol{\beta} - \mathbf{y}_{\text{train}}\|_2^2,$$

   and report the test MSE.

   (d) Use 5-fold cross-validation to choose the best ridge regression estimator

   $$\boldsymbol{\beta}_{\text{RR}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n_{\text{train}}}\|\mathbf{X}_{\text{train}}\boldsymbol{\beta} - \mathbf{y}_{\text{train}}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

   for $\lambda \in \{1\times10^{-4}, 1\times10^{-3}, ..., 1\times10^{2}\}$. Report the training objective values *and* training MSEs for all choices of λ. Indicate the best choice of λ and report the corresponding test MSE.

   (e) Similarly, use 5-fold cross-validation to choose the best Lasso estimator

   $$\boldsymbol{\beta}_{\text{lasso}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n_{\text{train}}}\|\mathbf{X}_{\text{train}}\boldsymbol{\beta} - \mathbf{y}_{\text{train}}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

   over the same set of regularization parameters as the previous task. Use `sklearn.linear_model.lasso` to fit the Lasso model: note that their formulation of the problem is different, so you should pass in 2λ as the regularization value in order to ensure you are minimizing the Lasso objective as we define it. Report the training objective values *and* training MSEs for all choices of λ. Indicate the best choice of λ and report the corresponding test MSE.

   (f) Use the support of $\boldsymbol{\beta}_{\text{lasso}}$ to reduce the number of columns in $\mathbf{X}_{\text{train}}$ to $\mathbf{X}_{\text{train,reduced}}$, then fit an OLS estimator $\boldsymbol{\beta}_{\text{OLS,reduced}}$ to this reduced dataset. Report the test MSE of this estimator.

   (g) Explain the relative performance on the test set of these different models: OLS, ridge regression, Lasso, and refitting a sparse model using OLS after Lasso.

3. [50 pts. Regularized Linear Regression Models with less training data.] Repeat the previous exercise, but using only the first 20K training examples to form $\mathbf{X}_{\text{train}}$.

(a) How do the optimal regularization parameters compare to the previous optimal regularization parameters?

(b) Is regularization more or less beneficial than in the previous exercise? Explain your answer.