TUM EI 70360: Machine Learning and Optimization
Fall 2023
Lecturer: Reinhard Heckel
Teaching Assistant: Tobit Klug

Problem Set 3
Issued: Tuesday, Oct. 31, 2023
Due: Thursday, Nov. 9, 2023

Problem 1 (Model selection failure)**.** Suppose we have randomly shuffled a dataset and split it into train, validation, and test set. We have trained a few methods on the training set, chosen the model that performs best on the validation set, and tested that model on the test set. The train and validation errors of the best model are small, but the test error is not. What is a likely reason for that?

Problem 2 (Number of models in model selection)**.** Suppose we have two validation sets, a small one and a large one that has 10 times as many examples. About how many models can we evaluate on the large validation set, without running into problems (i.e., without the confidence intervals becoming to small?)

Problem 3 (Ridge regression and cross validation)**.** In this problem you will use ridge regression to estimate the salary of baseball players based on measured features. This data set is taken from the ISLR package, an R package that accompanies the Introduction to Statistical Learning textbook. See `https://cran.r-project.org/web/packages/ISLR/ISLR.pdf` for more information. We have converted the data to CSV format to make it easier to use in languages other than R. Download the file `hitters.zip` from the course website. Inside this archive are `hitters.x.csv` and `hitters.y.csv`. The former is a 263×19 matrix of feature variables, and the latter a 263-dimensional vector of salaries; each row corresponds to one of the players. In Python, you can import the data as follows:

```
import csv
import numpy as np

# load feature variables and their names
X = np.loadtxt('hitters.x.csv', delimiter=',', skiprows=1)
with open('hitters.x.csv', 'r') as f:
    X_colnames = next(csv.reader(f))

# load salaries
y = np.loadtxt('hitters.y.csv', delimiter=',', skiprows=1)
```

1. A common practice in machine learning is to scale the feature variables so that they have standard deviation 1, before solving regression problems. Explain why such scaling is appropriate in our particular application. Also, apply this scaling to the data before going on to the subsequent parts of the problem.

2. We would like to include a bias term in our ridge regression setup, which we can accomplish by adding an additional feature with a constant value of 1 across all instances, but we do not want the bias weight to be included in the ℓ_2 norm penalty. Give an expression for the closed-form ridge regression using the augmented data matrix $\tilde{\mathbf{X}} = [\mathbf{1}\ \mathbf{X}]$ that does not include the bias weight in the penalty.

3. For 100 values of λ, evenly spaced in the interval $[10^{-3}, 10^{7}]$ on a logarithmic scale, compute the ridge regression solution using the closed-form expression with bias as previously discussed, and verify that, as λ decreases, the value of the penalty term $\|\theta\|_2$ increases. You should do this by plotting the ℓ_2 norm of the regression coefficients versus λ on a log-log plot.

4. Verify that, for a very small value of λ, the ridge regression estimate is very close to the least squares estimate. Also verify that, for a very large value of λ, the ridge regression estimate approaches 0 in all components (except the intercept, which is not penalized).

5. Perform 5-fold cross-validation to determine the best value of λ. You should implement this cross-validation procedure yourself. Produce a plot of the cross-validation error curve as a function of λ.

6. Report the coefficient estimates at the best value of λ as determined by cross-validation. Suppose that you were coaching a young baseball player who wanted to strike it rich in the major leagues. What handful of attributes would you tell this player to focus on?