STAT 535 Homework 5
Out November 12, 2020
Due November 19, 2020
©Marina Meilă
mmp@stat.washington.edu

Problem 1 – Double descent
This problem reproduces the demo shown in class and the setup from Mei and Montanari. The data in `hw5-train-dd500.dat` contains $N = 500$ labeled data points in $d = 100$ dimensions. Estimate the following predictor

$$f(x; W, \beta) \;=\; \sum_{j=1}^{p} \beta_j \sigma(W_{j:} x), \tag{1}$$

where $\beta_{1:p}$ are real parameters to be determined, and the matrix $W \in \mathbb{R}^{p \times d}$ has random i.i.d. rows, each row sampled uniformly on the sphere with radius $\sqrt{d}$ in d dimensions; σ represents the ReLU activation $\sigma(z) = \max(z, 0)$.

a. Write the expression for estimating β by ridge regression, given data and W.

b. Denote $\gamma = \frac{p}{N}$. For γ in the range $(0, 5]$, and $\lambda = 10^{-8}$ repeat the following. Sample values for the matrix W. Then, with this W and the training data, estimate β. Calculate the error $\hat{L}_{LS}(\gamma)$ of this predictor on the training set, and on the test data `h25-test-dd5000.dat`; denote the latter by $L_{LS}(\gamma)$ (it's an estimate of the expected loss). Don't forget to normalize the losses by the sample size. *Basic programming hint: when you run a for loop over a sample, always pre-allocate a vector where to store the results. Otherwise you may have very slow running code.*

Plot $\hat{L}_{LS}$ and L_{LS} vs γ on the same graph. Does your graph resemble the graphs in Mei and Montanari? From your graph, identify the ranges of γ for the three regimes underfitting, overfitting, and interpolation. What is the lowest value of the test set error you obtain in your experiment?

c. Now plot L_{LS} with a logarithmic scale for the y-axis. Trim the axis limits appropriately to have a legible plot.

Problem 2 – Model selection
In this problem, you apply AIC, BIC and Cross-validation (CV) to linear regression. Your models are:

$$(\mathcal{M}_1) \quad y \;=\; \beta^T x + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2) \tag{2}$$

$$(\mathcal{M}_2) \quad y \;=\; \beta_0 \mathbf{1}^T x + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2) \tag{3}$$

with $x \in \mathbb{R}^d, y \in \mathbb{R}$ and sample size N. Denote by Y, respectively X the N output vector and $N \times d$ design matrix obtained from your i.i.d. sample. Note that in the second model all regressioh coefficients have the same value β_0.

a. Write the formulas for the Maximum Likelihood estimation of β and β_0. Write the formulas for the ML estimation of σ^2 in each of the two models.

b. Let $f_{1,2}$ denote the ML predictors obtained for $\mathcal{M}_{1,2}$. Use the training data from Problem 1 to estimate the parameters of f_1 and f_2. What are your numerical values for $\beta_{[1:10\,21\,74]}$, β_0, $\sigma^2_{1,2}$? What are the log-likelihoods of each model?

c. Write the expressions of AIC and BIC for $\mathcal{M}_{1,2}$, and calculate their numerical values. Which model is selected by each of them?

d. Use the test set from Problem 1 as validation set for your estimated models. Which model is chosen by CV?

Problem 3 – Test set sample size

This problem explores the accuracy of comparing predictors on a test set, when the predictors make few errors. You are running a prediction competition, where the task is binary classification, and the loss L denotes the 0/1 loss. You have a test set $\mathcal{D}$ of $N = 500$ samples. There are two entries in the competition, predictor f_1 which makes 1 error on $\mathcal{D}$ and predictor f_2 which makes 2 errors. Your problem is whether to declare f_1 the winner (the naive approach) or to declare both f_1 and f_2 winners. To declare a single winner, you would like to have confidence at least $1 - \delta$ that $L_1 < L_2$, where $L_{1,2}$ denote the *expected* losses of $f_{1,2}$. Use the bounds in Lecture 12 and in the tutorial by Langford to answer the following.

a. Let $\delta = e^{-10}$. Can you conclude w.p. $1 - \delta$ from your current test set results (and the results available in Lecture 12) that $L_1 < L_2$?

b. Suppose you sample a larger test set, and on it $\hat{L}_1 = 1/500$ again. Now you conclude that $L_1 < 2/500$ w.p. $1 - e^{-10}$. What is a sufficiently large value of N to allow you this conclusion?

c. Suppose now that you sample yet another test, on it $\hat{L}_{1,2}$ have the same values as on the original $\mathcal{D}$, and your confidence is $1 - 2e^{-10}$. What is a sufficiently large value of N to allow you to conclude that $L_1 < L_2$? *(Note the distinction between this question and 3.b)*