ELEG/FSAN 817
Large Scale Machine Learning
Homework 1–High Dimensional Data

Due: 11:59PM, Wednesday, October 5, 2022

Please turn in a single PDF consisting of your discussion, figures, numeric values, and code. For the second problem you don't need to typeset: you can attach images of handwritten analysis.

I. Problem 1 (Simulation)

For this problem you will write code in the language of your choice to produce figures and numerical results. Numeric scripting languages such as python, MATLAB, Julia, or R are suggested. You will need access to functions for random number generation, linear algebra, and plotting. You will also have to answer questions and interpret the results.

The overall goal is to generate data based on the description in Bach's "Bolasso: Model Consistent Lasso Estimation through the Bootstrap" §4.1. **Steps:**

1) Start by generating a random matrix $\tilde{\mathbf{G}} \in \mathbb{R}^{p \times p}$, with $p = 32$. Each entry is independent and distributed as a standard normal $\mathcal{N}(0, 1)$ (Gaussian distribution).
2) Normalize each row of the matrix to have a ℓ_2-norm (Euclidean) of 1. Call the resulting matrix $\mathbf{G}$.
3) Then generate a sample of sample size $n = 1000$ for feature vectors of length $p = 32$ as independent draws of the random vector $X \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, with $X \in \mathbb{R}^p$, $\mathbf{Q} = \mathbf{G}\mathbf{G}^\top$. The resulting sample $\{x_i\}_{i=1}^n$ with $x_i \in \mathbb{R}^p$ should have a mean close to a vector of 0s and a sample covariance $\hat{\mathbf{Q}}$ similar to $\mathbf{Q}$.
4) Now consider $Y = \mathbf{w}^\top X + \epsilon$, where $\mathbf{w}$ is a random vector (that is fixed across the sample) and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
5) Given $r = 8$, $w_i = \begin{cases} s_i \tilde{w}_i & 1 \leq i \leq r \\ 0 & r < i \leq p \end{cases}$ for $i \in \{1, \ldots, p\}$, where $s_i \sim \text{Rademacher}$ and $\tilde{w}_i \sim \text{Uniform}([\frac{1}{3}, 1])$ for $i \in \{1, \ldots, r\}$.
6) The noise variance is dependent on $\mathbf{w}^\top \mathbf{Q} \mathbf{w}$, specifically $\sigma = 0.1\sqrt{\mathbf{w}^\top \mathbf{Q} \mathbf{w}}$.
7) The end result are tuples consisting of a vector in $\mathbb{R}^p$ and regression target in $\mathbb{R}$, $\{(x_i, y_i)\}_{i=1}^n$.
8) Consider a simple classification scheme. The class label $c_i \in \{-1, 1\}, i \in \{1, \ldots, n\}$ is defined by thresholding $\mathbf{w}^\top x_i$ at $\eta = -1.5$, where $c_i = \begin{cases} 1 & \mathbf{w}^\top x_i \geq \eta \\ -1 & \mathbf{w}^\top x_i < \eta \end{cases}$.
9) For class 1, find the instance $i^\star$ with the maximum value of regression value $y_{i^\star} = \max_{i \in \{1, \ldots, n\}} y_i$; it will serve as the class 1 prototype. You will consider the distance from this prototype point $x_{i^\star}$ to the remaining $n - 1$ instances.

Deliverables:

1) Human readable code to correctly generate a sample and identify the class-1 prototype [4 points].
2) Based on the generation process for $\mathbf{G}$ state the value of the diagonal entries of $\mathbf{Q}$. [2 points]
3) Calculate and report the ℓ_2 norm of the empirical mean vector, and compute the normalized Frobenius norm between the true covariance matrix and the sample's covariance matrix. $\frac{\|\mathbf{Q} - \hat{\mathbf{Q}}\|_F}{\|\mathbf{Q}\|_F} = \frac{\sqrt{\sum_{i=1,j=1}^{p,p} (Q_{ij} - \hat{Q}_{ij})^2}}{\sqrt{\sum_{i=1,j=1}^{p,p} Q_{ij}^2}}$ [4 points].
4) Based on the generation process for $\mathbf{Q}$ and $\mathbf{w}$, are the features in the subset of relevant features $\text{supp}(\mathbf{w})$ correlated to the irrelevant features? Please explain. How would $\mathbf{Q}$ have to be structured

to ensure the two feature sets (relevant and irrelevant) were uncorrelated? (Hint: write a condition on a subset of the entries of $\mathbf{Q}$) [5 points]

5) Make a scatter plot with two sets of markers, for each data point the horizontal coordinate is the Euclidean distance from the class-1 prototype and the vertical coordinate is the observed value of $y_i = \mathbf{w}^\top x_i + \epsilon$ for $i \in \{1, \dots, n\}$. Make the markers different shapes and colors to indicate the class of each point. [5 points]
6) Make a similar plot but use the ℓ_∞ norm to compute the distance (Chebychev distance) [5 points]
7) For each of the distances, find the smallest k such that the k-th nearest neighbor from the class-1 prototype is from the other class. Report the median of k across 500 random realizations (generating a random $\tilde{\mathbf{Q}}$, $\tilde{\mathbf{w}}$, and $\tilde{\mathbf{s}}$ on each realization and then generating the sample of size 1000). Plot a histogram of k across these 500 runs. Which distance gives the smallest median value of k? [10 points]
8) Which of the following metrics would be a poor choice of a distance/dissimilarity measure for a k-nearest neighbor classifier in the presence of a large number of irrelevant features? Justify your answer. [5 points]
   - $d_A(x, x') = \max_{I \subseteq \{1,\dots,d\}} \sum_{i \in I} |x_i - x'_i|$
   - $d_B(x, x') = \max_{i \in \{1,\dots,d\}} |x_i - x'_i|$
   - $d_C(x, x') = \left(\sum_{i=1}^d |x_i - x'_i| \right)^{10}$
   - $d_D(x, x') = \sum_{i=1}^d |x_i - x'_i|^{\frac{1}{10}}$
9) Write a short 5–10 sentence discussion of the results and insights from these exercises. [5 points]

II. Problem 2 (Theory)

1) Let $\mathbf{u} \sim \mathcal{N}(\mathbf{c}, \mathbf{I}_q)$ denote a random vector with a multivariate normal distribution with mean $\mathbf{c} \in \mathbb{R}^q$ and identity covariance. Let $\mathbf{A} \in \mathbb{R}^{d \times q}$ be an independent random matrix with entries

$$A_{ij} = \begin{cases} +1 & \text{with probability } p \\ 0 & \text{with probability } 1 - 2p \\ -1 & \text{with probability } p \end{cases}.$$

   Simplify the expressions of $\mathbb{E}[\mathbf{A}\mathbf{u}]$ and $\mathbb{E}[\mathbf{A}(\mathbf{u} - \mathbf{c})]$ as much as possible. [5 points]

2) Simplify the expression $\mathbb{E}[\mathbf{A}\mathbf{A}^\top]$. (Hint: you can define the result element-wise.) [10 points]