DS-GA 3001: Homework Problem Set 1

Special Topics in Data Science: (Spring 2018)

MATHEMATICS OF DATA SCIENCE: Graphs and Networks

Afonso S. Bandeira
bandeira@cims.nyu.edu
http://www.cims.nyu.edu/~bandeira

Shuyang Ling
sling@cims.nyu.edu
http://www.cims.nyu.edu/~sling

Due on February 1, 2018

This homework problem set is due on February 1, before class at the homework dropbox in the CDS reception.

If you have questions about the homework feel free to contact me or Shuyang, or stop by our office hours.

Try not to look up the answers, you'll learn much more if you try to think about the problems without looking up the solutions. If you need hints, feel free to email me of Shuyang.

You can work in groups but each student must write his/her own solution based on his/her own understanding of the problem. Please list, on your submission, the students you work with for the homework (this will not affect your grade).

Late submissions will be graded with a penalty of 10% per day late.

If you need to impose extra conditions on a problem to make it easier (or consider specific cases of the question, like taking n to be 2, e.g.), state explicitly that you have done so. Solutions where extra conditions were assume, or where only special cases where treated, will also be graded (probably scored as a partial answer).

Linear Algebra

Problem 1.1 *Show the following: If $M \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $d \leq n$ then*

$$\max_{\substack{U\in\mathbb{R}^{n\times d}\\ U^TU=\mathrm{I}_{d\times d}}} \mathrm{Tr}\left(U^TMU\right) = \sum_{k=1}^{d} \lambda_k^{(+)}(M),$$

where $\lambda_k^{(+)}$ is the largest k-th eigenvalue of M.

Estimators

Problem 1.2 *Given $x_1, \cdots, x_n$ i.i.d. samples from a distribution X with mean μ and covariance Σ, show that*

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \textit{and} \quad \Sigma_n = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - \mu_n)(x_k - \mu_n)^T,$$

are unbiased estimators for μ and Σ, i.e., show that $\mathbb{E}[\mu_n] = \mu$ and $\mathbb{E}[\Sigma_n] = \Sigma$.

Connectivity of the Erdős-Rényi random graph

The Erdős-Rényi random graph $G(n,p)$ is a graph with n nodes, where each edge (i,j) appears (independently) with probability p. In this problem set, you will show a remarkable phase transition: if $\lambda < 1$, then $G(n, \frac{\lambda \ln n}{n})$ has, with high probability, isolated nodes while, if $\lambda > 1$, the graph is connected (with high probability).

Problem 1.3 *Let I_i be a random variable indicating whether node i is isolated: $I_i = 1$ if node i is isolated, and $I_i = 0$ otherwise. Let $X = \sum_{i=1}^{n} I_i$ be the number of isolated nodes.*

The goal is to show that $\Pr\{X = 0\}$ is small when $\lambda < 1$ (meaning that there are isolated notes, with high probability). In the proof you can use the approximation

$$(1 - \lambda/n)^n \approx e^{-\lambda} \quad (\textit{for large } n)$$

1. *Show that $\mathbb{E}[X] \approx n^{-\lambda+1}$.* ***Note:*** *The fact that $\mathbb{E}[X] \to \infty$ is not sufficient to show $\Pr\{X = 0\} \to 0$ (**why? Can you give a counter-example?**). We need to ensure that X concentrates around its mean.*

2. *Use (a simple) concentration inequality showed in the first lab to finish the proof. (The technique you have just derived is known as the second moment method)*

Problem 1.4 *Prove that, if $\lambda \geq 1$, $G(n, \frac{\lambda \ln n}{n})$ is connected with high probability:*

1. *Derive the probability for a set of k nodes ($k \leq n/2$) being disconnected from the rest of the graph.*
2. *Prove the probability of graph G having a disconnected component goes to zero as n grows (hint: use union bound).*

Little Grothendieck problem

Problem 1.5 *Let $C \succeq 0$ (C is positive semidefinite). In this homework you'll show an approximation ratio of $\frac{2}{\pi}$ to the problem*

$$\max_{x_i = \pm 1} \sum_{i,j=1}^{n} C_{ij} x_i x_j.$$

Similarly to `Max-Cut`*, we consider*

$$\max_{\substack{v_i \in \mathbb{R}^n \\ \|v_i\|^2 = 1}} \sum_{i,j=1}^{n} C_{ij} v_i^T v_j.$$

The goal is to show that, for $r \sim \mathcal{N}(0, I_{n \times n})$, taking $x_i^\natural = \operatorname{sign}(v_i^T r)$ a randomized rounding,

$$\mathbb{E}\left[\sum_{i,j=1}^{n} C_{ij} x_i^\natural x_j^\natural\right] \geq \frac{2}{\pi} \sum_{i,j=1}^{n} C_{ij} v_i^T v_j$$

Hints:

1. *The main difficulty is that $\mathbb{E}\left[\operatorname{sign}(v_i^T r) \operatorname{sign}(v_j^T r)\right]$ is not linear in $v_i^T v_j$ and C_{ij} might be negative for some (i, j)'s.*
2. *Show that that $\mathbb{E}\left[\operatorname{sign}(v_i^T r) v_j^T r\right]$ is linear in $v_i^T v_j$. What is it equal to?*
3. *Construct S with entries $S_{ij} = \left(v_i^T r - \sqrt{\frac{2}{\pi}} \operatorname{sign}(v_i^T r)\right)\left(v_j^T r - \sqrt{\frac{2}{\pi}} \operatorname{sign}(v_j^T r)\right)$*
4. *Show that $\operatorname{Tr}(CS) \geq 0$.*