Mathematics of Data Science Prelim Exam Sampler

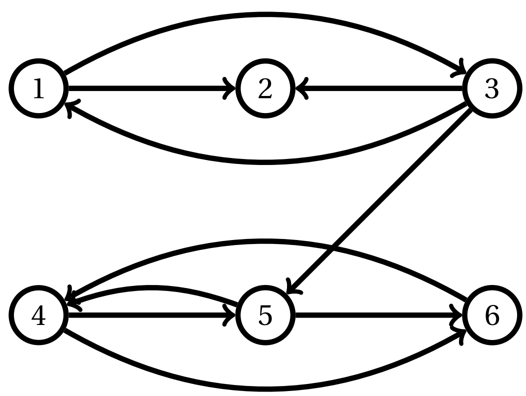
These collection of 18 problems is intended to help you review and prepare for the prelim on *Mathematics of Data Science*. The exam tests for knowledge of MAT 167 (Applied Linear Algebra) MAT 168 (Mathematical Optimization) and MAT 170 (Mathematics of Data and Decisions). Some problems are harder and longer than what you will find in the exam. Try to write the solutions as you would in the exam: Write all details when solving every problem. Be organized and use the notation appropriately. Initially try to solve the problems without any assistance.

A prelim exam will be 3 hours long and have 6 problems. In what follows we present in consecutive order six problems from MAT 167, six problems from MAT 168, and 6 problems from MAT 170, plus some computer projects At the end of the collection we have the syllabus of the test.

PART I: Problems from Applied Linear Algebra

Problem 1

Consider the following web graph.



(a) (5 pts) Write the *adjacency matrix* L of the above web graph.

(b) (10 pts) Describe the HITS algorithm.

(c) (5 pts) Now, write the "Google" matrix (i.e., *probability transition matrix*) G of the above graph assuming that a transition (i.e., random walk) from each node to its outlinked nodes is equally likely.

(d) (5 pts) Explain that the "Google" matrix in Part (b) is not *row stochastic*. Then describe how to turn it into a row stochastic matrix $\widetilde{G}$, and explicitly write $\widetilde{G}$.

(e) (5 pts) Let $\widetilde{G}$ be row stochastic after what you did in Part (d). Is the underlying Markov chain is *reducible* or *irreducible*? State your reasoning too. If it is reducible, then describe how to make it irreducible.

(f) (10 pts) Assuming the modified "Google" matrix $\widetilde{G}$ is row stochastic, describe the PageRank algorithm.

Score of this page:_______________

Problem 2 Let $X = [x_1 \cdots x_n] \in R^{d\times n}$ be a data matrix where each column x_j is a vector in R^d, $j = 1:n$ and assume $d > n$.

(a) (5 pts) Let $\overline{x} \in R^d$ be the *mean* of these n column vectors. Let $\widetilde{X} \in R^{d\times n}$ be a *centered* data matrix where each column vector is $x_j - \overline{x}$, $j = 1:n$. Write explicitly the *centering matrix* $C \in R^{n\times n}$ such that $\widetilde{X} = XC$.

[Hint: First, try to express $\overline{x}$ using the vector $\mathbf{1}_n = [1,1,\ldots,1]^\top \in R^n$.]

(b) Let S be the *sample covariance matrix* of the dataset $\{x_1,\ldots,x_n\}$. Express S using the centered matrix $\widetilde{X}$.

(c) Explain S has rank $n-1$ if $\{x_1,\ldots,x_n\}$ are linearly independent.

(d) Explain why the largest $n-1$ eigenvalues and the corresponding eigenvectors of S and the square of the first $n-1$ singular values divided by n and the corresponding *left* singular vectors of $\widetilde{X}$ coincide.

Score of this page:________________

Problem 3 (35 pts) Consider the following matrix: A = $\begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$.

(a) (10 pts) Compute the SVD of A.

(b) (5 pts) Compute the orthonormal basis of range(A).

(c) (5 pts) Compute the pseudoinverse $A^\dagger$ of A.

(d) (5 pts) Compute the orthogonal projector onto range(A).

(e) (5 pts) Compute the orthonormal basis of null(A).

(f) (5 pts) Compute the orthogonal projector onto range$(A^\top)$.

Score of this page:________________

Problem 4 (20 pts) Let us consider the following *diagonal* matrix:

$$D = (d_1, \ldots, d_n) = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{bmatrix},$$

where $d_j \in R$.

(a) (10 pts) Prove that $\|D\|_p = \max_{1 \le j \le n} |d_j|$ for any p with $1 \le p \le \infty$.

[Hint: Recall the definition of p-norm of a matrix and that of a vector. Also, it may be easier to separate the case of $1 \le p < \infty$ and that of $p = \infty$ in your proof.]

(b) Suppose $d_j \ne 0$ for $j = 1 : n$. What is the *condition number* $\kappa(D)$ of this matrix D?

Score of this page:________________

Problem 5 (25 pts)

Consider a *term-document matrix* $A = [{}_1, \cdots, {}_n] \in^{m\times n}$ where $m > n$. Let $\in^m$ be a query vector.

(a) (10 pts) Describe how to find relevant documents for this query vector. More precisely define the *cosine similarity* between the query vector and the document vector ${}_j$, $j = 1 : n$, and describe how to use this cosine similarity for retrieving relevant documents.

[Hint: The document ${}_j$ is considered to be relevant if its cosine similarity with is larger than the tolerance value `tol`.]

(b) (15 pts) Describe how to use the best rank k approximation to A via SVD, and describe the advantage of using such an approximation compared to the strategy in Part (a), which does not involve any approximation.

[Hint: Use the notation $A_k = U_k \Sigma_k V_k^{\mathsf{T}} = U_k H_k$ with $H_k = [h_1, \cdots, h_n] := \Sigma_k V_k^{\mathsf{T}} \in R^{k\times n}$ for the best rank k approximation of A using the SVD.]

Score of this page:________________

Problem 6 The following are questions about eigenvalues and singular values.

(a) What are the singular values of a $1 \times n$ matrix? Write down its singular value decomposition.

(b) Prove that if A is a square non-singular matrix then the singular values of the A^{-1} are the reciprocals of the singular values of A.

(c) Suppose A is an $m \times m$ matrix with SVD $A = U\Sigma V^T$. Use this to find the **singular value** decomposition of the $2m \times 2m$ matrix:

$$\begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}$$

d) If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of a square non-singular matrix, what are the eigenvalues of its inverse matrix?

e) Suppose now A is an $m \times n$ matrix. Prove that the non-zero eigenvalues of $A^T A$ and AA^T are the same. Are the eigenvectors the same too? Why?

Score of this page:________________

These collection of 18 problems is intended to help you review and prepare for the prelim on *Mathematics of Data Science*. The exam tests for knowledge of MAT 167 (Applied Linear Algebra) MAT 168 (Mathematical Optimization) and MAT 170 (Mathematics of Data and Decisions). Some problems are harder and longer than what you will find in the exam. Try to write the solutions as you would in the exam: Write all details when solving every problem. Be organized and use the notation appropriately. Initially try to solve the problems without any assistance.

A prelim exam will be 3 hours long and have 6 problems. In what follows we present in consecutive order six problems from MAT 167, six problems from MAT 168, and 6 problems from MAT 170, plus some computer projects At the end we have provided the syllabus and suggested reading to prepare for the test.

PART II: Problems from Optimization

1. Given $n = 1, 2, \ldots$, consider linear program (P_n) given by

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} x_i & \\ \text{s.t.} \quad & x_i \leq 1, & i = 1, \ldots, n \\ & x_i \geq 0, & i = 1, \ldots, n. \end{aligned}$$

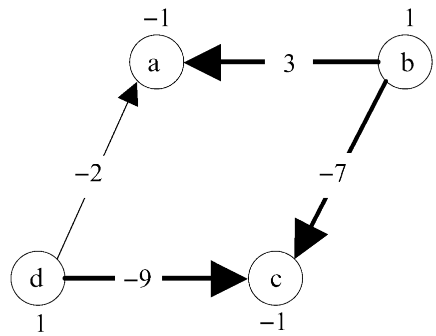
(a) For $n = 2$, sketch the feasible region.

(b) For $n = 3$, write its dual linear program and state primal and dual solutions that verify the strong duality theorem (that is, solutions such that $c^T x^* = b^T y^*$ where c is the primal objective, b is the dual objective and x^*, y^* are primal and dual solutions, respectively).

(c) For general n, what is the optimal solution?

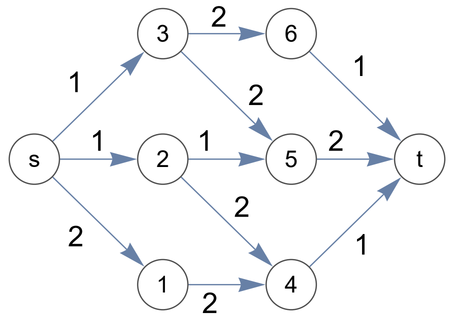
(d) Let $T(n)$ be the number of pivots of the simplex method on (P_n) with any valid pivot rule. Determine $T(n)$.

2. Consider the following network flow problem, where the numbers at the arcs are the costs per unit of flow and the numbers at the nodes are the supplies at each node (demands if negative).



State a linear programming formulation for the associated minimum cost network flow problem.

3. Consider the following directed graph where the labels at the edges are capacities.



(a) Determine (by any means) the maximum flow from node s to node t. Specify the flow across each edge.

(b) State a cut separating s from t that, with the maximum flow, satisfies the max-flow min-cut theorem.

4. Answer the following questions. For true-or-false questions give a justification. A notation introduced in one part of the problem may be used in later parts of the problem. Throughout this problem, we consider the linear optimization problem

$$\max c^T x \quad \text{s.t.} \quad Ax = b,\ x \geq 0,$$

where the matrix $A \in R^{m \times n}$ has full row rank. The feasible region is denoted by F.

a) Define "basis" of A.

b) Write the dictionary corresponding to basis $\mathcal{B}$ in matrix notation, using the given data A, b, c.

c) True or false: If basis $\mathcal{B}$ is primal feasible and dual feasible, then the primal basic solution corresponding to $\mathcal{B}$ is optimal.

d) True or false: Every optimal solution x^* is the basic solution of some basis $\mathcal{B}$.

e) True or false: If x^* is an optimal solution that is the basic solution corresponding to basis $\mathcal{B}$, then basis $\mathcal{B}$ is primal and dual feasible.

f) True or false: Every vertex of F corresponds to a unique basis.

i) True or false: The problem is unbounded if its feasible region is unbounded.

j) True or false: The feasible region is unbounded if the problem is unbounded.

m) State the Carathéodory theorem, explain what it implies on the support of a feasible solution of the linear program.

5. **Duality of Linear Programs**

a) Dualize the following problem. (Call the dual variables y_i.)

$$\begin{aligned} \max \zeta = 4x_1 + 8x_2 + 15x_3 & \\ 16x_1 + 23x_2 + 42x_3 \geq 108 & \\ 4x_1 - 8x_2 + 15x_3 \leq 0 & \\ 16x_1 - 23x_2 = 42 & \\ x_1 \geq 0, \quad x_2 \leq 0, \quad x_3 \in . & \end{aligned}$$

If you introduced slack variables for some reason, get rid of them for your final answer.

b) Introduce slack variables where necessary to turn this problem into equation form.

c) State the complementary slackness theorem (the complete theorem, not just the equations of the complementarity conditions!) for this problem.

d) State the weak duality theorem.

e) State the strong duality theorem.

f) True or false: If the dual problem is unbounded, the primal is infeasible.

g) True or false: If the dual problem is infeasible, the primal is unbounded.

6. a) Convex function models are important in applications. State reasons why a convex model is preferable to a non-convex model.

   b) Suppose that $g_1, g_2, \dots, g_m$ are convex functions from R^n into R, then prove

$$\bar{g}(x) = \max\{g_1(x), g_2(x), \dots, g_m(x)\}$$

   is also a convex function.

   Now given $1 \leq k \leq n$ we define the function

$$s_k(x) = \sum_{i=1}^{k} x_{[i]},$$

   where $x_{[i]}$ is the i-th largest component of x. Use what you know to show that the function s_k is convex.

   c) Consider the mathematical model below.

$$\begin{array}{ll} \max & x_1 + 2x_2 \\ & x_1 + x_2 \leq 1 \\ & x_1 - x_2 \leq 0 \\ & 2x_1^2 - x_2 \geq 0 \\ & x_1, x_2 \geq 0 \\ & x_1, x_2 \end{array}$$

   i Sketch the feasible set of this problem. Is the problem convex?

   ii What is the optimal cost?

   iii Write the Lagrangian function of this problem.

   iv State the Karush Kuhn Tucker conditions for optimality.

These collection of 18 problems is intended to help you review and prepare for the prelim on *Mathematics of Data Science*. The exam tests for knowledge of MAT 167 (Applied Linear Algebra) MAT 168 (Mathematical Optimization) and MAT 170 (Mathematics of Data and Decisions). Some problems are harder and longer than what you will find in the exam. Try to write the solutions as you would in the exam: Write all details when solving every problem. Be organized and use the notation appropriately. Initially try to solve the problems without any assistance.

A prelim exam will be 3 hours long and have 6 problems. In what follows we present in consecutive order six problems from MAT 167, six problems from MAT 168, and 6 problems from MAT 170, plus some computer projects . At the end we provide the syllabus and references that can help you prepare for the test.

PART III: Problems from Modeling with Data and Machine Learning

1. **Data Proximity and Linear Algebra:** We consider a collection of m news articles that are represented as m points x_i, $i = 1, \ldots, m$ in R^n, with a bag-of-words method. Here n is the total number of words in a given dictionary, and the j-th element in x_i contains the number of times word j appears in article i. We are given two additional articles, called the origin and destination articles, represented by two vectors x_0 and x_d in R^n. Your goal is to find a small number $k << m$ of articles that are in some way connecting the origin and destination articles (think of trying to find a sequence of articles that connect the beginning and the end of the coverage of a certain topic).

   To this end, find the k points that are closest to the line L (not the segment!) passing through x_0 and x_d.

   (a) Show that to simplify, and without loss of generality, we can reduce the problem to the case when $x_0 = 0$ and x_d is normalized ($||x_d||_2 = 1$).

   (b) For a given point x, find the point on the line L that is closest, in Euclidean norm, to x.

   (c) Show that the distance D from x to its projection on the line satisfies the equation $D^2 = x^T Px$, where the matrix $P = I - x_d(x_d)^T$.

   (d) Finally, explain how you can find a set of k articles that connect x_0 and x_d, in the sense defined above.

2. **Math models for the optimal order of enrolling in courses.**

A directed graph $G = (V, A)$ can be used to represent the order of actions to be take in any project (task a needs to be done *before* b, if there is an arc from a to b). A *topological ordering* of the vertices is an assignment of the value y_i to each vertex such that for every arc ij then $y_i \geq y_j + 1$. The assignment says the best order to do a project.

(a) Show that a directed graph has a topological ordering, then there is no directed cycle. Write a simple discrete model to detect whether a directed graph has a cycle. **Hint:** Linear algebra is a plausible way to detect cycles, why?

(b) A UC Davis student in major X (say applied math major, economic major, etc.) has to take certain required courses that have pre-requisites. Create a directed graph whose nodes are all possible Math courses in each of the majors and there is an arc from course a to course b if course a is a *pre-requisite* to b, thus a has to be taken before b! How big is this graph? Let us call this the *Math-major-course* graph, MMC Make some comments about the structure of the MMC graph (number of nodes? number of arcs?). How does the graph differs from a stats major's graph?

(c) Write a math model that, given one of the 3 majors of the UC Davis mathematics department, tells the students at least one good order, one that does not skip pre-requisites, in which to take their math courses and graduate in minimum amount of time. Assume that there is a limit of two math courses to take each quarter. Can you find at least two good orders in each of the majors?

(d) Write pseudo-code (in any language of your preference) that can solve this problem for any of the majors of the mathematics department.

3. **A Geometric Optimization model** Your problem is packing m of spheres in a box of minimal area. The spheres have a given radius r_i, and the problem is to determine the precise location of the centers x_i. The constraints in this problem are that the spheres should not overlap, and should be contained in a square of center 0 and half-size R. The objective is to minimize the area of the containing square.

   - Show that two spheres of radius r_1, r_2 and centers x_1, x_2 respectively do not intersect if and only if $||x_1 - x_2||_2$ exceeds a certain number, which you can determine.
   - Use this to formulate the sphere packing problem as an optimization model. Is the formulation you have found a convex optimization problem?
   - Write pseudocode (in any language of your preference) to solve the packing problem of five and six circular disks of the same radius inside a square of half-size R.

4. **How to spread misinformation and lies efficiently**

FACEBOOK and X (formerly known as TWITTER) are now commonly used to spread falsehoods and lies. In this problem, you will consider a simplified model of influence in social networks: the social network is represented by an undirected graph $G = (V, E)$ and for a set of nodes $S \subseteq V$, we denote by $N(S)$ the set of their neighbors: $N(S) = \{v \in V \mid \exists u \in S,\ (u, v) \in E\}$. The *influence* $I(S)$ of a set of nodes is measured by $I(S) = |N(S)|$.

(a) As the evil designer of a marketing campaign to falsely influence the opinion of people and destroy the world your goal is to find a subset $S \subseteq V$ of at most K nodes whose influence is maximal. Write a mathematical model to solve this optimization problem. How does the answer depend on K?

(b) A *vertex-cover* of a graph G is a set S of vertices of G such that each edge of G is incident with at least one vertex of S. The vertex cover number $\tau(G)$ is the size of a minimum vertex cover in G. A *dominating set* for a graph is a subset D of vertices such that every vertex not in D is adjacent to at least one member of D. The domination number $\gamma(G)$ is the smallest dominating set. Are there any relationships among the influence function $I(S)$, $\tau(G)$, and $\gamma(G)$? Explain.

Formulate discrete models that given a graph finds a vertex cover with smallest number of vertices and a smallest dominating set. Explain the reasoning on your variables and constraints.

(c) Show that if M is a matching and S is some vertex cover $|S| \geq |M|$. In particular show that

$$\max\{|M| : M \text{ is a matching of } G\} \leq \min\{|S| : S \text{ is a vertex-cover of } G\}.$$

HINT: How many vertices do you need to cover just the edges of M?

(d) Write pseudo-code (in any language of your preference such as MATLAB, Python, SCIP, Gurobi, etc) so given a real data set (a graph) you solve the influence problem for the graph

5. **SVM Models for automatic cancer diagnostic (15 pts)**

In this project, we will use a simple support vector machine (SVM) for *breast cancer diagnosis.* The project will use real data from the *Wisconsin Diagnosis Breast Cancer Database* (WDBC). You will find a discriminating function (a separating plane in this case) to determine if an unknown tumor sample is benign or malignant. In order to do this, you will use part of the data in the above database as a "training set" to generate your separating hyperplane and the remaining part as a "testing" set to test your separating plane. Attributes 3 to 32 form a 30-dimensional vector representing each case as a point in 30-dimensional real space R^{30}. To generate the separating plane, a training set, consisting of two disjoint point sets B and M in R^{30} representing confirmed benign and malignant cases.

(a) Formulate the problem as a linear program that finds the separating hyperplane. Solve the problem using the M and B as a training set from the first 500 cases of the wdbc.data file available from

`https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`

The last 69 points should be used as a testing set. Solve the linear program and print out the separating hyperplane, and the minimum value of the LP.

(b) Test the separating plane on the 69 cases of the testing set. Report the number of misclassified points on the testing set. It is probably a good idea if you write code to do this.

(c) There is also a *quadratic programming* formulation of the separation problem (to maximize margin distance). Formulate the quadratic program in and try to solve it that way. Is the separating hyperplane better?

(d) Suppose that the oncologist wants to use only 2 of the 30 attributes in her diagnosis. Give a way to find a pair of attributes most effective in determining a correct diagnosis

6. **A data ranking model** An exam with m questions is given to n students. The instructor collects all the grades in a $n \times m$ matrix G, with G_{ij} the grade obtained by student i on question j. We would like to assign a difficulty score to each question, based on the available data.

   (a) Assume that the matrix G is well approximated by a rank-one matrix sq^T , with $s \in R^n$ and $q \in R^m$. (You may assume that both s, q have non-negative components.) Explain how to use the approximation to assign a difficulty score to each question.

   (b) What is the interpretation of vector s?

   (c) How would you compute a rank-one approximation to G?

Data Science Syllabus

Sources:

[1] Lars Eldén. *Matrix Methods in Data Mining and Pattern Recognition*, SIAM, 2007. https://doi.org/10.1137/1.9780898718867

[2] Calafiore & El Ghaoui. *Optimization Models*, Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781107279667

[3] Guenin, Könemann & Tunçel. *A Gentle Introduction to Optimization*, Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781107282094

[4] Higham & Higham. *Deep Learning: An Introduction for Applied Mathematicians*, SIAM Review, 2019. https://doi.org/10.1137/18M1165748

[5] Gilbert Strang. *Linear Algebra and Learning from Data*, SIAM, 2019. https://epubs.siam.org/doi/book/10.1137/1.9780692196380

[6] Robert J. Vanderbei. *Linear Programming: Foundations and Extensions*, Springer, 2020. https://link.springer.com/book/10.1007/978-3-030-39415-8

Books contain lots of good questions that can be used to prepare and to create exams. Moreover they have electronic versions available for free through UC Davis (and most universities).

1. Applied Linear Algebra (at the level of MAT 167)
1.1 Vectors, Matrices, Eigenvalues [1, Ch 2,15] or [2, Ch 1,2,3,7] or [5, Part I]
1.2 Singular Value Decomposition [1, Ch 6] or [2, Ch 5] or [5, Part I]
1.3 Linear Equations, Least Squares, and LU Decomposition [1, Ch 3] or [2, Ch 6,7] or [5, Part I]
1.4 Orthogonality, QR Decomposition, and Gram-Schmidt Projections [1, Ch 4,5] or [5, Part I,II]
1.5 Symmetric Matrices, PSD [1, Ch 3.2] or [2, Ch 4] or [5, Part I,II]

2. Optimization (At the level of MAT 168/170)
2.1 Convex Analysis, Gradient Descent [2, Ch 8] or [5, Part VI] or [3, Ch 7.3,7.4] or [6, Ch 10]
2.2 Linear Programming and its Duality [3, Ch 2,3] or [2, Ch 9] or [6, Ch 2,3]
2.3 Network Flows, Max-Flow Min-Cut Theorem [3, Ch 1.4,1.5,5.3] or [6, Ch 15]
2.4 General Optimality Conditions (Karush-Kuhn-Tucker) in Nonlinear Optimization [3, Sec 7.5,7.7]

3. Data/Machine Learning (at the level of MAT 170)
3.1 Basic Models for Supervised Learning and Regression [2, Sec 13.1,13.2] or [6, Ch 12]
3.2 Binary Classification Models (Support Vector Machines, Logistic Regression) [2, Sec 13.3]
3.3 Unsupervised Learning [2, Sec 13.5]
3.4 *k*-Means, Spectral Clustering [1, Ch 9] and [5, Part IV.7]
3.5 Basics of Neural Networks [4] or [5, Part VII]

Students are also expected to have some “hands-on” experience with programming and feel comfortable working with computers and data. (However, the prelim exam will not include programming.)