

Final Exam, CPSC 8420, Spring 2022

Last Name, First Name

Due 05/06/2022, Friday, 11:59PM EST

Problem 1 [15 pts]

Consider the elastic-net optimization problem:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda[\alpha\|\beta\|_2^2 + (1 - \alpha)\|\beta\|_1]. \quad (1)$$

1. Show the objective can be reformulated into a lasso problem, with a slightly different \mathbf{X}, \mathbf{y} .
2. If we fix $\alpha = .5$, please derive the closed solution by making use of alternating minimization that each time we fix the rest by optimizing one single element in β . You need randomly generate \mathbf{X}, \mathbf{y} and initialize β_0 , and show the objective decreases monotonically with updates.

You may input your answers here. \LaTeX version submission is encouraged.



Problem 2 [15 pts]

- For PCA, the loading vectors can be directly computed from the q columns of \mathbf{U} where $[\mathbf{U}, \mathbf{S}, \mathbf{U}] = \text{svd}(\mathbf{X}^T \mathbf{X})$, please show that any $[\pm \mathbf{u}_1, \pm \mathbf{u}_2, \dots, \pm \mathbf{u}_q]$ will be equivalent to $[\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_q]$ in terms of the same variance while satisfying the orthonormality constraint.
- We consider the case when original dimensionality of the data is much larger than the number of samples $d \gg m$ ($\mathbf{X} \in \mathbb{R}^{d \times m}$). What's the complexity of obtaining the optimal solution of PCA via Singular Value Decomposition? Please consider a more efficient solution by considering the relationships of eigenvalues/eigenvectors between $\mathbf{X}^T \mathbf{X}$ and $\mathbf{X} \mathbf{X}^T$.



As a part of an exercise I have to prove the following:

35

Let A be an $(n \times m)$ matrix. Let A^T be the transposed matrix of A . Then AA^T is an $(n \times n)$ matrix and $A^T A$ is an $(m \times m)$ matrix. AA^T then has a total of n eigenvalues and $A^T A$ has a total of m eigenvalues.



20

What I need to prove is the following:



AA^T has an eigenvalue $\mu \neq 0 \iff A^T A$ has an eigenvalue $\mu \neq 0$

In other words, they have the same non-zero eigenvalues, and if one has more eigenvalues than the other, then these are all equal to 0.

How can I prove this?

Thanks and regards.

Problem 3 [10 pts]

Assume that in a community, there are 10% people suffer from COVID. Assume 80% of the patients come to breathing difficulty while 25% of those free from COVID also have symptoms of shortness of breath. Now please determine that if one has breathing difficulty, what's his/her probability to get COVID? (*hint*: you may consider Naive Bayes)

Problem 4 [20 pts]

Recall the objective for RatioCut: $RatioCut(A_1, A_2, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{|A_i|}$. If we introduce indicator

vector: $h_j \in \{h_1, h_2, \dots, h_k\}, j \in [1, k]$, for any vector $h_j \in R^n$, we define: $h_{ij} = \begin{cases} 0 & v_i \notin A_j \\ \frac{1}{\sqrt{|A_j|}} & v_i \in A_j \end{cases}$,

we can prove: $h_i^T L h_i = \frac{cut(A_i, \bar{A}_i)}{|A_i|}$, and therefore:


$$RatioCut(A_1, A_2, \dots, A_k) = \sum_{i=1}^k h_i^T L h_i = \sum_{i=1}^k (H^T L H)_{ii} = tr(H^T L H), \quad (2)$$


thus we relax it as an optimization problem:


$$\underbrace{arg \min}_H tr(H^T L H) \quad s.t. \quad H^T H = I. \quad (3)$$


Now let's explore Ncut, with objective: $NCut(A_1, A_2, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \frac{W(A_i, \bar{A}_i)}{vol(A_i)}$, where $vol(A) :=$

$\sum_{i \in A} d_i, d_i := \sum_{j=1}^n w_{ij}$. Similar to RatioCut, we define: $h_{ij} = \begin{cases} 0 & v_i \notin A_j \\ \frac{1}{\sqrt{vol(A_j)}} & v_i \in A_j \end{cases}$. Now

1. Please show that $h_i^T L h_i = \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$. 

2. Show that $NCut(A_1, A_2, \dots, A_k) = tr(H^T L H)$. 

3. The constraint now is: $H^T D H = I$. 

4. Find the solution to $\underbrace{arg \min}_H tr(H^T L H) \quad s.t. \quad H^T D H = I$. 

Problem 5 [10 pts]

We consider the following optimization problem (\mathbf{Y} is given and generated randomly):

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \|\mathbf{X}\|_* \quad (4)$$

where $\mathbf{Y}, \mathbf{X} \in \mathbb{R}^{100 \times 100}$ and $\|\cdot\|_*$ denotes the nuclear norm (sum of singular values). Now please use gradient descent method to update \mathbf{X} . ($\frac{\partial \|\mathbf{X}\|_*}{\partial \mathbf{X}} = \mathbf{U}\mathbf{V}^T$, where \mathbf{U}, \mathbf{V} is obtained from reduced SVD, namely $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{X}, 0)$). Plot the objective changes with 1000 iteration.



Problem 6 [20 pts]

We turn to Logistic Regression:

$$\min_{\beta} \sum_{i=1}^m \ln(1 + e^{\langle \beta, \hat{x}_i \rangle}) - y_i \langle \beta, \hat{x}_i \rangle, \quad (5)$$

where $\beta = (w; b)$, $\hat{x} = (x; 1)$. Assume $m = 100$, $x \in \mathbb{R}^{99}$. Please randomly generate x, y and find the optimal β via 1) gradient descent; 2) Newton's method and 3) stochastic gradient descent (SGD) where the batch-size is 1. (need consider choosing appropriate step-size if necessary). Change $m = 1000$, $x \in \mathbb{R}^{999}$, observe which algorithm will decrease the objective faster in terms of iteration (X -axis denotes number of iteration) and CPU time. [You will receive another 5 bonus points if you implement backtracking line search]



Problem 7 [10 pts]

Please design an (either toy or real-world) experiment to demonstrate that PCA can be helpful for denoising.



Bonus Problem 8 [10 pts]

$$\text{Solve: } \min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{y}. \quad (6)$$

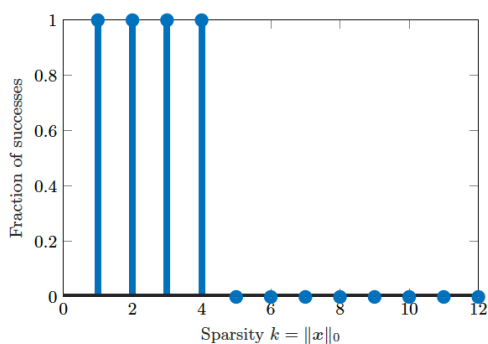
We have proved that if $\mathbf{y} = \mathbf{Ax}_o$ with

$$\|\mathbf{x}_o\|_0 \leq \frac{1}{2} \text{krank}(\mathbf{A}). \quad (7)$$

Then \mathbf{x}_o is the unique optimal solution to the ℓ^0 minimization problem

$$\min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{y}. \quad (8)$$

However, when \mathbf{A} is of size 5×12 , the following figure illustrates the fraction of success across 100 trials. Apparently $\text{krank}(\mathbf{A}) \leq \text{rank}(\mathbf{A}) \leq 5$, therefore, when sparsity $k = 1, 2$ satisfying Eq. (7)



it has 100% recovery success rate is not surprising. However, the above experiment also shows even $k = 3, 4$ which violates Eq. (7), still it can be recovered at 100%. Please explain this phenomenon.