Homework Set 2, CPSC 8420, Spring 2022

Your Name

Due 03/17/2022, Thursday, 11:59PM EST

Problem 1

For PCA, from the perspective of maximizing variance, please show that the solution of $\boldsymbol{\phi}$ to maximize $\|\mathbf{X}\boldsymbol{\phi}\|_2^2$, $s.t.$ $\|\boldsymbol{\phi}\|_2 = 1$ is exactly the first column of $\mathbf{U}$, where $[\mathbf{U}, \mathbf{S}] = svd(\mathbf{X}^T\mathbf{X})$. (Note: you need prove why it is optimal than any other reasonable combinations of $\mathbf{U}_i$, say $\hat{\boldsymbol{\phi}} = 0.8 * \mathbf{U}(:, 1) + 0.6 * \mathbf{U}(:, 2)$ which also satisfies $\|\hat{\boldsymbol{\phi}}\|_2 = 1$.)

How to Calculate the Residual Sum of Squares

RSS = $\sum^n{}_{i=1} (y^i - f(x_i))^2$

Where:

y_i = the i^{th} value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

Formula for Residuals

The formula for residuals is straightforward:

Residual = observed y – predicted y

Problem 2

Why might we prefer to minimize the sum of absolute residuals instead of the residual sum of squares for some data sets? Recall clustering method K-means when calculating the centroid, it is to take the mean value of the data-points belonging to the same cluster, so what about K-medians? What is its advantage over of K-means? Please use a synthetic (toy) experiment to illustrate your conclusion.

Problem 3

Let's revisit Least Squares Problem: $\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{n \times p}$.

1. Please show that if $p > n$, then vanilla solution $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$ is not applicable any more.
2. Let's assume $\mathbf{A} = [1,2,4;1,3,5;1,7,7;1,8,9], \mathbf{y} = [1;2;3;4]$. Please show via experiment results that Gradient Descent method will obtain the optimal solution with Linear Convergence rate if the learning rate is fixed to be $\frac{1}{\sigma_{max}(\mathbf{A}^T\mathbf{A})}$, and $\boldsymbol{\beta}_0 = [0;0;0]$.
3. Now let's consider ridge regression: $\underset{\boldsymbol{\beta}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$, where $\mathbf{A}, \mathbf{y}, \boldsymbol{\beta}_0$ remains the same as above while learning rate is fixed to be $\frac{1}{\lambda+\sigma_{max}(\mathbf{A}^T\mathbf{A})}$ where λ varies from $0.1, 1, 10, 100, 200$, please show that Gradient Descent method with larger λ converges faster.

Problem 4

Please download the image from `https://en.wikipedia.org/wiki/Lenna#/media/File:Lenna_(test_image).png` with dimension $512 \times 512 \times 3$. Assume for each RGB channel data X, we have $[U, \Sigma, V] = svd(X)$. Please show each compression ratio and reconstruction image if we choose first $2, 5, 20, 50, 80, 100$ components respectively. Also please determine the best component number to obtain a good trade-off between data compression ratio and reconstruction image quality. (Open question, that is your solution will be accepted as long as it's reasonable.)