CIS4930 - Math for Machine Learning

Homework 4

Fernando Sckaff

March 6th, 2023

Question 1

Analytically evaluate the following derivatives (with the answers given in the uploaded Graham fragment):

1. $\frac{\partial \operatorname{trace}(A^T W)}{\partial W}$
2. $\frac{\partial \operatorname{trace}(W^T A W B)}{\partial W}$
3. $\frac{\partial \operatorname{trace}[(Y-XW)^T(Y-XW)]}{\partial W}$
4. $\frac{\partial W^{-1}}{\partial w_{ij}}$ where each entry of the matrix $W_{ij} = w_{ij}$ (Hint: Consider identity $W^{-1}W = I$.

Solution: Using Alexander Graham Kronecker's *Products Matrix Calculus Formulae.*

1.

$$\frac{\partial \operatorname{trace}(A^T W)}{\partial W} = A$$

2.

$$\frac{\partial \operatorname{trace}(W^T A W B)}{\partial W} = AWB + A^T W B^T$$

3.

$$\begin{aligned}
\frac{\partial \operatorname{trace}[(Y-XW)^T(Y-XW)]}{\partial W} &= \frac{\partial}{\partial W} \operatorname{trace}[Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW] \\
&= \frac{\partial}{\partial W} \operatorname{trace}[Y^T Y - 2W^T X^T Y + W^T X^T XW] \\
&= \frac{\partial}{\partial W} \Big(\operatorname{trace}[Y^T Y] - 2 \operatorname{trace}[W^T X^T Y] + \operatorname{trace}[W^T X^T XW] \Big) \\
&= -2X^T Y + 2X^T XW
\end{aligned}$$

4.

$$\frac{\partial W^{-1}}{\partial w_{ij}}$$

To start,

$$\begin{aligned}
WW^{-1} &= I \\
\frac{\partial}{\partial w_{ij}}(WW^{-1}) &= \frac{\partial}{\partial w_{ij}} I \\
\frac{\partial}{\partial w_{ij}}(WW^{-1}) &= 0 \\
\frac{\partial W}{\partial w_{ij}}(W^{-1}) + W\frac{\partial}{\partial w_{ij}}(W^{-1}) &= 0 \qquad \text{[Chain Rule]}
\end{aligned}$$

Solving for $\frac{\partial}{\partial w_{ij}}(W^{-1})$, we get:

$$\frac{\partial}{\partial w_{ij}}(W^{-1}) = -W^{-1}\frac{\partial W}{\partial w_{ij}}W^{-1}$$

$\frac{\partial W}{\partial w_{ij}}$ is a matrix with all entries equal to zero except for the (i, j) entry, which is equal to 1.

Question 2

Show that the objective function for W (a $K_1 \times K_2$ matrix of weights),

$$E(W) = \sum_{i=1}^{N} \|y_i - W^T x_i\|_2^2 + \lambda \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} w_{kl}^2$$

where $\{x_i, y_i\}_{i=1}^N$ is the set of training patterns (instances) with $x_i \in \mathbb{R}^{\mathbb{K}_1}$ and $y_i \in \mathbb{R}^{\mathbb{K}_2}$ can be written as

$$E(W) = \text{trace}\left[(Y - XW)^T (Y - XW)\right] + \lambda \, \text{trace}(W^T W)$$

where X is $N \times K_1$ and Y is $N \times K_2$. Show all steps.

Solution: I must show two things.
(1) $\sum_{i=1}^{N} \|y_i - W^T x_i\|_2^2 = \text{trace}\left[(Y - XW)^T (Y - XW)\right]$
(2) $\sum_{k=1}^{K_1} \sum_{l=1}^{K_2} w_{kl}^2 = \text{trace}(W^T W)$

We can expand the first term as follows:

$$\begin{aligned}
\sum_{i=1}^{N} \|y_i - W^T x_i\|_2^2 &= \sum_{i=1}^{N} (y_i - W^T x_i)^T (y_i - W^T x_i) \\
&= \sum_{i=1}^{N} (y_i^T - x_i^T W)(y_i - W^T x_i) \\
&= \sum_{i=1}^{N} (y_i^T y_i - y_i^T W^T x_i - x_i^T W y_i + x_i^T W^T W x_i) \qquad [\text{OBS - } \sum_{i=1}^{n} a_{ii} = \text{trace}(A)] \\
&= \text{trace}(Y^T Y) - \text{trace}(Y^T XW) - \text{trace}(W^T X^T Y) + \text{trace}(W^T X^T XW) \\
&= \text{trace}[Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW] \\
&= \text{trace}[Y^T Y - 2Y^T XW + W^T X^T XW] \\
&= \text{trace}\left[(Y - XW)^T (Y - XW)\right]
\end{aligned}$$

We can rewrite the second term as follows:

$$\begin{aligned}
\lambda \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} w_{kl}^2 &= \lambda \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} (w_{kl}^T w_{kl}) \\
&= \lambda \, \text{trace}(W^T W)
\end{aligned}$$

Combining the two terms, we get:

$$E(W) = \text{trace}\left[(Y - XW)^T (Y - XW)\right] + \lambda \, \text{trace}(W^T W) \quad \text{😄}$$

Question 3

SVD-based image reconstruction:

1. Load the hendrix_final.png image and extract the R, G and B channels. (Convert each channel image to float64 or equivalent.)
2. Execute the SVD separately on the R, G and B channels of the image. Plot (using a log-log plot) the non-zero singular values for the R channel. Comment on the nature of the plot, specifically the drop in singular values for each channel.
3. Plot the Frobenius norm of the reconstruction error matrix for each channel w.r.t. the dimension (increasing from 1 to the rank).
4. Give your own criterion as to how many dimensions you would pick (the same number for all three channels) to get the best trade off between reconstruction error and image fidelity (to the original).
5. Display the original and final reconstructed images (combined from R, G and B reconstructions and using your criterion) side by side. You may reduce the size of the original image (at the very outset) in order to ease the computational burden. Comment on your criterion for the choice of reduced size (if any).

Solution: on Google Colab, Link:

https://colab.research.google.com/drive/1Eftdd0YjZvsGEnPrSeKmkqz3b5ohFMZn?usp=sharing

Question 4

Run the least-squares linear discriminant code shown in class on the Iris dataset (http://en.wikipedia.org/wiki/Iris_flower_data_set/) It has three classes and four features. You may use all of the data for training. Show the three 2D scatter plots $\left[w_1^T x, w_2^T x\right]$, $\left[w_1^T x, w_3^T x\right]$ and $\left[w_2^T x, w_3^T x\right]$ where you plot all the training set data in each scatter plot. There is no test data set. Comment on the nature of the plot.

Solution: on Google Colab, Link:

https://colab.research.google.com/drive/15w0XrtB8Kc7t3ZCPPqSw1j9aJhhUk8G7?usp=sharing

F.S.

Note:-

- Question 1
  - I. Correct
  - II. Correct
  - II. Correct
  - IV. Correct
- Question 2: Correct
- Question 3: Correct
- Question 4: 15/20 Couldn't make out the virginica/versicolor divide because you used the same color.