CSCE636 Homework 2

Solution Keys

October 14, 2019

1

Let $\mathbf{H} = \mathbf{U}\Sigma\mathbf{U}^T$, where $\boldsymbol{UU^T = U^TU = I}$ and $\mathbf{U} = [\mathbf{u_1}, ..., \mathbf{u_n}]$, $\boldsymbol{\Sigma = diag(\lambda_1, \lambda_2, ..., \lambda_n)}$. Then we know that $\boldsymbol{u_1, ..., u_n}$ is a group of orthogonal vectors and $\boldsymbol{u_t^T u_t = 1, 1 \leq t \leq n}$.
Let $\boldsymbol{A = [a_1, ..., a_k]}$, and $\boldsymbol{A^T U = [b_1, ..., b_n], b_t = A^T u_t, 1 \leq t \leq n}$, we have

$$\begin{aligned}
trace(A^T HA) &= trace(A^T U\Sigma U^T A) \\
&= trace(\sum_{t=1}^{n} \lambda_t b_t b_t^T) \\
&= \sum_{t=1}^{n} trace(\lambda_t b_t b_t^T) \\
&= \sum_{t=1}^{n} \lambda_t trace(b_t b_t^T) \\
&= \sum_{t=1}^{n} \lambda_t trace(b_t^T b_t) \\
&= \sum_{t=1}^{n} \lambda_t b_t^T b_t \\
&= \sum_{t=1}^{n} \lambda_t u_t^T AA^T u_t \\
&= \sum_{t=1}^{n} \lambda_t ||A^T u_t||_2^2
\end{aligned}$$

8 points up to here.

Since $\boldsymbol{A^T A = I}$ and $\boldsymbol{A \in R^{n \times k}, k < n}$, we know that $\boldsymbol{a_1, ..., a_k}$ must be a group of orthogonal vectors. Let $\boldsymbol{A' = [a_{k+1}, ..., a_n]}$ and $\boldsymbol{a_1, ..., a_k, a_{k+1}, ..., a_n}$ can form a group of standard orthogonal basis of R^n. Then $\boldsymbol{[A, A'][A, A']^T = [A, A']^T [A, A'] = I}$. So we have

$$\boldsymbol{||A^T u_t||_2^2 \leq ||A^T u_t||_2^2 + ||(A')^T u_t||_2^2 = ||[A, A']^T u_t||_2^2 = u_t^T [A, A'][A, A']^T u_t = u_t^T u_t = 1}$$

The above inequality becomes equality only when $\boldsymbol{(A')^T u_t = 0}$, thus $\boldsymbol{u_t}$ is in the orthogonal complement space of the linear space formed by $\boldsymbol{a_{k+1}, ..., a_n}$, and this orthogonal complement space is exactly the column space of $\boldsymbol{A}$. So in this case there must exist a vector $\boldsymbol{q_t}$ such that $\boldsymbol{u_t = Aq_t, q_t \in R^k}$. In addition, $\boldsymbol{u_t}$ is a unitary vector so we have $\boldsymbol{1 = u_t^T u_t = q_t^T A^T A q_t = q_t^T q_t}$. In conclusion, we always have $\boldsymbol{||A^T u_t||_2^2 \leq 1}$ and only when $\boldsymbol{u_t = Aq_t}$, $\boldsymbol{q_t}$ is a unit vector, then we will have $\boldsymbol{||A^T u_t||_2^2 = 1}$. This is the complete proof of the first sub-question, 4 points.

Besides, we know that $\boldsymbol{A^TUU^TA = A^TA = I, I \in R^{k\times k}}$ so we know that

$$\begin{aligned} k &= trace(A^TUU^TA) \\ &= trace(U^TAA^TU) \\ &= \sum_{t=1}^{n} u_t^T AA^T u_t \\ &= \sum_{t=1}^{n} ||A^T u_t||_2^2 \end{aligned}$$

This is the complete proof of the second sub-question, 4 points.

Thus we have

$$\begin{aligned} trace(A^THA) &= \sum_{t=1}^{n} \lambda_t ||A^T u_t||_2^2 \\ &\leq \sum_{t=1}^{k} \lambda_t ||A^T u_t||_2^2 + \lambda_{k+1} \sum_{t=k+1}^{n} ||A^T u_t||_2^2 \\ &= \sum_{t=1}^{k} \lambda_t ||A^T u_t||_2^2 + \lambda_{k+1}(k - \sum_{t=1}^{k} ||A^T u_t||_2^2) \\ &= \sum_{t=1}^{k} (\lambda_t - \lambda_{k+1}) ||A^T u_t||_2^2 + k\lambda_{k+1} \\ &\leq \sum_{t=1}^{k} (\lambda_t - \lambda_{k+1}) + k\lambda_{k+1} \\ &= \sum_{t=1}^{k} \lambda_t \end{aligned}$$

4 points.

All inequalities become equalities when $\boldsymbol{||A^Tu_t||_2 = 1}$ for $1 \leq t \leq k$ and $\boldsymbol{||A^Tu_t||_2 = 0}$ for $t > k$. So we can conclude that $\boldsymbol{u_t = Aq_t}$ for $1 \leq t \leq k$ and apparently $\boldsymbol{q_1, ..., q_k}$ are an arbitrary group of unit orthogonal vectors. Thus $\boldsymbol{A = [u_1, ..., u_k]Q, Q = [q_1, ..., q_t]^T \in R^{k\times k}}$. So far we can claim that $\boldsymbol{\max(trace(A^THA)) = \sum_{t=1}^{k} \lambda_t}$, and $\boldsymbol{A^* = [u_1, ..., u_k]Q}$, where $\boldsymbol{Q}$ is an arbitrary orthogonal matrix.

2

Let $h(z) = \tanh z = \frac{e^z - e^{-z}}{e^z + e^{-z}}$, we have

$$\frac{dh}{dz} = \frac{(e^z + e^{-z})^2 - (e^z - e^{-z})^2}{(e^z + e^{-z})^2} = 1 - h(z)^2$$

Thus we can get

$$\begin{aligned}
\bigtriangledown E_{in}(\mathbf{w}) &= \bigtriangledown \frac{1}{N}\sum_{n=1}^{N}(h(\mathbf{w^T x_n}) - \mathbf{y_n})^{\mathbf{2}} \\
&= \frac{2}{N}\sum_{n=1}^{N}(h(\mathbf{w^T x_n}) - \mathbf{y_n}) \bigtriangledown (\mathbf{h}(\mathbf{w^T x_n})) \\
&= \frac{2}{N}\sum_{n=1}^{N}(h(\mathbf{w^T x_n}) - \mathbf{y_n})(\mathbf{1} - \mathbf{h}(\mathbf{w^T x_n})^{\mathbf{2}}) \bigtriangledown (\mathbf{w^T x_n}) \\
&= \frac{2}{N}\sum_{n=1}^{N}(h(\mathbf{w^T x_n}) - \mathbf{y_n})(\mathbf{1} - \mathbf{h}(\mathbf{w^T x_n})^{\mathbf{2}})\mathbf{x_n} \\
&= \frac{2}{N}\sum_{n=1}^{N}(\tanh(\mathbf{w^T x_n}) - \mathbf{y_n})(\mathbf{1} - \tanh^{\mathbf{2}}(\mathbf{w^T x_n}))\mathbf{x_n}
\end{aligned}$$

If $\mathbf{w} \to \infty$, then $\tanh^2(\mathbf{w^T x_n})$ will be very close to 1 and $\bigtriangledown E_in(\mathbf{w})$ will approach to 0. When this happened, the back propagation will change a very little amount in $\mathbf{w}$, which essentially causes gradient vanishing.

3

$$s^{(1)} = W^{(1)}x^{(0)} \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 1 \end{bmatrix}$$

$$x(1) = \begin{bmatrix} 1 \\ \tanh(s^{(1)}) \end{bmatrix} = \begin{bmatrix} 1 \\ \tanh(0.7) \\ \tanh(1) \end{bmatrix} = \begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix}$$

$$s^{(2)} = W^{(2)T}x^{(1)} = [-1.48]$$

$$x^{(2)} = \begin{bmatrix} 1 \\ \tanh(s^{(2)}) \end{bmatrix} = \begin{bmatrix} 1 \\ \tan(-1.48) \end{bmatrix} = \begin{bmatrix} 1 \\ -0.90 \end{bmatrix}$$

$$x^{(3)} = s^{(3)} = W^{(3)T}x^{(2)} = [-0.8]$$

$$\delta^{(3)} = 2(x^{(3)} - y) = [-3.6]$$

$$\frac{\partial e}{\partial W^{(3)}} = x^{(2)}\delta^{(3)} = \begin{bmatrix} 1 \\ -0.9 \end{bmatrix}[-3.6] = \begin{bmatrix} -3.6 \\ 3.24 \end{bmatrix}$$

$$\delta^{(2)} = 2 \times (1 - 0.9^2) \times \delta^{(3)} = -1.368$$

$$\frac{\partial e}{\partial W^{(2)}} = x^{(1)}\delta^{(2)} = \begin{bmatrix} 1 \\ 0.6 \\ 0.76 \end{bmatrix}[-1.368] = \begin{bmatrix} -1.368 \\ -0.82 \\ -1.04 \end{bmatrix}$$

$$\delta^{(1)} = \begin{bmatrix} 1 - 0.6^2 \\ -3 \times (1 - 0.76^2) \end{bmatrix} \times \delta^{(2)} = \begin{bmatrix} -0.88 \\ 1.73 \end{bmatrix}$$

$$\frac{\partial e}{\partial W^{(1)}} = x^{(0)}\delta^{(1)T} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}[-0.881.73] = \begin{bmatrix} -0.88 & 1.73 \\ -1.76 & 3.46 \end{bmatrix}$$

4

(e) For $d = 32$, the reconstruction error of PCA and AE are 129.4 and 129.9; for $d = 64$, the error are 85.8 and 86.2; for $d = 128$, the error are 45.6 and 46.5, respectively.
The reconstruction error of PCA is always very close, but a little bit smaller than that of AE. This is because in this settings, AE is a linear model, and PCA always gives optimal solutions (minimal reconstruction errors here) among all linear models. In addition, AE is a convex problem when weights are shared, so it can achieve minimal reconstruction error in theory. In practice, however, results obtained by gradient-descent usually reach a value near the minimal unless training infinite epochs.

(f) For $p = d = 32, 64, 128$, the Frobenius norm of $\mathbf{W} - \mathbf{G}$ are $||\mathbf{W} - \mathbf{G}||_F = 7.9, 11.3, 16.1$ respectively. Therefore, $\mathbf{W}$ and $\mathbf{G}$ are not the same (5 points).
Exploring the relations between $\mathbf{W}$ and $\mathbf{G}$ is an open-ended question. Your solutions will be considered as correct if you give either of the below 2:
Sol1: According to note, $\mathbf{G}$ is a column orthogonal matrix because G is component of orthogonal eigenvectors of centered data matrix. Therefore, $\mathbf{G}^T\mathbf{G} = \mathbf{I}$. Also, suppose $\mathbf{W}$ obtained by AE is as optimal as PCA , then $\mathbf{W}$ can be denotes as $\mathbf{W} = \mathbf{GQ}$, where $\mathbf{Q}$ is an orthogonal matrix. Therefor, $\mathbf{W}$ is column orthogonal and we have $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. Frobenius norm of $\mathbf{G}^T\mathbf{G} - \mathbf{W}^T\mathbf{W}$ could be computed experimentally and $||\mathbf{G}^T\mathbf{G} - \mathbf{W}^T\mathbf{W}||_F = 0.007, 0.01, 0.02$ for $p = d = 32, 64, 128$ respectively, which is in line with our analysis the matrix $\mathbf{G}$ and $\mathbf{W}$ all have relation that $\mathbf{W^TW} \approx \mathbf{G^TG} = \mathbf{I}$.
Sol2: Experimentally show that $\mathbf{WW}^T$ and $\mathbf{GG}^T$ are close to each other.
Other reasonable solutions will also be accepted as long as experimental justification is provided.

(g) For $d = 32, 64, 128$, the autoencoder without shared weights have the reconstruction errors of 129.7, 86.0 and 46.1, respectively. They are very close to the reconstruction errors of the autoencoder with shared weights. The autoencoder with shared weights is just the special case of the autoencoder without shared weights, and the gradient descent process may push the encoder and decoder to have same weights so as to approach the optimal network.

(h) The original data is not linearly separable. By building a nonlinear model using multiple layers and nonlinear activations, you are expected to achieve much lower reconstruction error compared with PCA and above two autoencoders.
This is one network setting. $batchsize = 256$, $epochs = 2000$, a 2-layer encoder and a dimension-symmetrical decoder with $tanh$ activation functions for the first layers and $relu$ for the last layer. Under this configuration, the reconstruction error is 47.7, which is much lower than 85.8 obtained by PCA.

5

You are expected to achieve the test accuracy better than 97%.
Here is one network setting: $epochs = 10$, $batchsize = 128$, $num\_hid\_units = 512$, $num\_hid\_units = 3$, $relu$ as activation function. Testing accuracy is 97.7%.