XCS224N Assignment #1 Extra Credit: Understanding SVD (5 points)

1 Primer on Singular Value Decomposition

1.1 What is the SVD?

The following section serves as a primer about **Singular Value Decomposition** (SVD). All the material in this section will be presented as fact (without proof) and we will ask you to use it to answer the extra credit questions in the sections below. Suppose we have some matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank r. Then $\mathbf{A}$ has a decomposition given by

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{1}$$

which has the following (along with other) properties

- $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$, $\mathbf{V} \in \mathbb{R}^{d \times r}$.
- The matrix $\mathbf{D}$ is diagonal with positive entries sorted in descending order. We often let let $\sigma_i := \mathbf{D}_{ii}$, and we then have $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r > 0$. We call σ_i the ith singular value of the $\mathbf{A}$.
- The columns of $\mathbf{V}$ form an orthonormal basis for the row space of $\mathbf{A}$, and thus $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. The ith column $\mathbf{v}_i$ of $\mathbf{V}$ is known as the ith right singular vector of $\mathbf{A}$.
- The columns of $\mathbf{U}$ form an orthonormal basis for the column space of $\mathbf{A}$, and thus $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. The ith column $\mathbf{u}_i$ of $\mathbf{U}$ is known as the ith left singular vector of $\mathbf{A}$.

The above decomposition is known as the SVD of $\mathbf{A}$.

1.2 Projections

Recall that for a linear subspace $W \subset \mathbb{R}^p$, the projection of some vector $\mathbf{v}$ onto W is defined as

$$\text{proj}_W(\mathbf{v}) := \underset{w \in W}{\arg\min} ||w - \mathbf{v}||_2^2 \tag{2}$$

In the special case that W is a one dimensional vector space spanned by some $\mathbf{w}$, we call this projection the projection of $\mathbf{v}$ onto $\mathbf{w}$ which has simple closed form. We give it below, where we have let $\hat{\mathbf{w}} := \mathbf{w}/||\mathbf{w}||_2$ represent the unit vector pointing in the direction of $\mathbf{w}$.

$$\text{proj}_{\mathbf{w}}(\mathbf{v}) = \left(\frac{\mathbf{v}^T\mathbf{w}}{||\mathbf{w}||_2^2}\right)\mathbf{w} = (\mathbf{v}^T\hat{\mathbf{w}})\hat{\mathbf{w}} \tag{3}$$

Intuitively, the quantity $\mathbf{v}^T\hat{\mathbf{w}}$ captures the amount that $\mathbf{v}$ points in the direction of $\mathbf{w}$.

1.3 Best Fit Sub-spaces

Suppose we have a set of vectors $S = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ where each $\mathbf{w}_i \in \mathbb{R}^d$. We define the k-dimensional best-fit subspace for S to be the k-dimensional linear subspace W such that

$$W = \underset{\dim(W)=k}{\arg\min} \sum_{i=1}^{n} ||\mathbf{w}_i - \text{proj}_W(\mathbf{w}_i)||_2^2 \tag{4}$$

Intuitively, W is the k-dimensional subspace that is closest to the vectors in set S.

It is an important and interesting fact that, for a matrix $\mathbf{A}$ with SVD as given above, the subspace given by $V_k = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is the k-dimensional best-fit subspace for the rows of $\mathbf{A}$. Informally, σ_i tells us the degree to which $\mathbf{v}_i$ contributes to fitting the rows of $\mathbf{A}$. Thus the vector which best fits the rows of A is $\mathbf{v}_1$, the vector which best fits the rows of $\mathbf{A}$ when the contribution of $\mathbf{v}_1$ is accounted for is $\mathbf{v}_2$, etc.

2 Extra Credit Challenge (5 Points)

Suppose we have a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with SVD $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$, $\mathbf{V} \in \mathbb{R}^{d \times r}$.

(a) (1 point) Show that

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{5}$$

Solution: Note that

$$\begin{aligned}
(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T)_{k\ell} &= \sum_{i=1}^{r} \sigma_i (\mathbf{u}_i \mathbf{v}_i^T)_{k\ell} \\
&= \sum_{i=1}^{r} \sigma_i (\mathbf{u}_i)_k (\mathbf{v}_i)_\ell \\
&= \sum_{i=1}^{r} \sigma_i \mathbf{V}_{\ell i}^T \mathbf{U}_{ik} \\
&= \sum_{i=1}^{r} \mathbf{V}_{\ell i}^T (\mathbf{D}\mathbf{U})_{ik} \\
&= (\mathbf{V}^T \mathbf{D}\mathbf{U})_{\ell k} \\
&= (\mathbf{U}\mathbf{D}\mathbf{V}^T)_{k\ell} \\
&= \mathbf{A}_{k\ell}
\end{aligned}$$

This clearly implies (5).

(b) (1 point) Show that

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i \tag{6}$$

In particular, the components of $\mathbf{u}_i$ represent the size of the projection of the rows of $\mathbf{A}$ onto $\mathbf{v}_i$ (scaled by σ_i).

Solution: Recalling that the $\mathbf{v}_i$ are orthonormal, we can use the result from (a) to see that

$$\begin{aligned}\mathbf{A}\mathbf{v}_j &= \left(\sum_{i=1}^{r}\sigma_i\mathbf{u}_i\mathbf{v}_i^T\right)\mathbf{v}_j \\ &= \left(\sum_{i=1}^{r}\sigma_i\mathbf{u}_i\mathbf{v}_i^T\mathbf{v}_j\right) \\ &= \sigma_j\mathbf{u}_j\end{aligned}$$

so $\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i$ as desired.

(c) (1 point) One way of finding a reduced rank approximation of $\mathbf{A}$ is by hard-setting all but the k largest σ_i to 0. This approximation is called the truncated SVD, and by (a) we see it can be written as

$$\mathbf{A}_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{7}$$

From (a), we see the truncated SVD can also be written as $\mathbf{A}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$, where $\mathbf{U}_k \in \mathbb{R}^{n \times k}$, $\mathbf{V} \in \mathbb{R}^{d \times k}$ are the first k columns of $\mathbf{U}, \mathbf{V}$, and $\mathbf{D} \in \mathbb{R}^{k \times k}$ has the first k singular values.

Show that the rows of $\mathbf{A}_k$ are the projections of the rows of $\mathbf{A}$ onto the subspace of V_k spanned by the first k right singular vectors.

Hint: Recall that the projection of a vector $\mathbf{a}$ onto a subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$ where the $\mathbf{v}_i$ are pairwise orthogonal is given by the sum of projections of $\mathbf{a}$ onto the individual $\mathbf{v}_i$.

Solution: For an arbitrary vector $\mathbf{a}$, using that the $\mathbf{v}_i$ are orthonormal (see the hint for why this is important), we have that the projection of $\mathbf{a}$ onto $V_k = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is $\sum_{i=1}^{k} (\mathbf{a}^T \mathbf{v}_i)\mathbf{v}_i$ from Equation (3). Thus the matrix whose rows are the projections of each row of $\mathbf{A}$ onto V_k is given by

$$\sum_{i=1}^{k} \mathbf{A}\mathbf{v}_i \mathbf{v}_i^T$$

Using the result from (b) we have that $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ so

$$\sum_{i=1}^{k} \mathbf{A}\mathbf{v}_i \mathbf{v}_i^T = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \mathbf{A}_k$$

as desired.

(d) (2 points) The Frobenius norm of a matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ is defined as

$$||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij}^2} \tag{8}$$

Show that

$$\mathbf{A}_k = \underset{\text{rank}(\mathbf{B})=k}{\arg\min} ||\mathbf{A} - \mathbf{B}||_F \tag{9}$$

where the arg min is taken over matrices of rank k.

Hint: Use the fact that V_k is the best-fit k-dimensional subspace for the rows of $\mathbf{A}$.

Solution: For a matrix $\mathbf{M}$ and linear subspace V, let $\text{proj}_V(\mathbf{M})$ represent the matrix with rows $\text{proj}_V(m_i)$, where m_i is the ith row of $\mathbf{M}$.

Consider an arbitrary matrix $\mathbf{B} \in \mathbb{R}^{n \times d}$ of rank k. Then let the k-dimensional space W be the span of the rows of $\mathbf{B}$, so W has dimension k. We see that

$$||\mathbf{A} - \mathbf{B}||_F^2 = \sum_{i=1}^{n} ||a_i - b_i||_2^2$$

where a_i, b_i are the rows of $\mathbf{A}$ and $\mathbf{B}$. From Equation (2), we know that

$$||\mathbf{A} - \mathbf{B}||_F^2 = \sum_{i=1}^{n} ||a_i - b_i||_2^2 \geq \sum_{i=1}^{n} ||a_i - \text{proj}_W(a_i)||_2^2 = ||\mathbf{A} - \text{proj}_W(\mathbf{A})||_F^2$$

But since V_k is the best-fit k-dimensional subspace for the rows of $\mathbf{A}$, we know that

$$||\mathbf{A} - \text{proj}_{V_k}(\mathbf{A})||_F^2 = \sum_{i=1}^{n} ||a_i - \text{proj}_{V_k}(a_i)||_2^2 \leq \sum_{i=1}^{n} ||a_i - \text{proj}_W(a_i)||_2^2 = ||\mathbf{A} - \text{proj}_W(\mathbf{A})||_F^2$$

Part (c) tells us that $\text{proj}_{V_k}(\mathbf{A}) = \mathbf{A}_k$. Putting everything together gives

$$||\mathbf{A} - \mathbf{A}_k||_F^2 \leq ||\mathbf{A} - \text{proj}_W(\mathbf{A})||_F^2 \leq ||\mathbf{A} - \mathbf{B}||_F^2$$

Since the above inequality is true for any rank k matrix $\mathbf{B}$, it follows that

$$\mathbf{A}_k = \underset{\text{rank}(\mathbf{B})=k}{\arg\min} ||\mathbf{A} - \mathbf{B}||_F^2 = \underset{\text{rank}(\mathbf{B})=k}{\arg\min} ||\mathbf{A} - \mathbf{B}||_F$$

3 Interpretation

Suppose we have some data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with rows x_i. The x_i represent n data samples each with d entries. In the case that d is a very large number, we may want to try and reduce the dimensionality of our data. Suppose we want to reduce the dimension of each sample from d to $k < d$. How can we go about doing this?

Consider taking the truncated SVD $\mathbf{X}_k$ of our data. From (c) we know that we have transformed our data so all the data samples live within a k-dimensional linear subspace, in particular, the k-dimensional linear subspace that best-fit our original data. Part (d) tells us that under the Frobenius norm, our new matrix $\mathbf{X}_k$ is as close to $\mathbf{X}$ as it can be given that it has this property.

Our new approximate data samples (given by the rows of $\mathbf{X}_k$) have dimension d, but they certainly don't need to. Since each sample lives in the same k-dimensional linear subspace, each sample's place in this subspace relative to other samples can be determined by how much it "points" in k fixed directions which span the subspace. Thus it suffices to fix k basis vectors for the subspace (these are our k directions) and re-parametrize each sample based on how much it "points" in the direction of each of these basis vectors.

From (c) we know that we can let $\mathbf{v}_i, \ldots, \mathbf{v}_k$ be a basis of our subspace. From (b), we then know that the ith row of $\mathbf{U}$ denoted u_i is such that its jth component $(u_i)_j$ tells us how much the ith sample points in the direction of $\mathbf{v}_j$. Thus, based on the above discussion, we can use the vector $((u_i)_1, \ldots, (u_i)_k)$ as a substitute for the ith row of $\mathbf{X}_k$. This leaves us with a k dimensional embedding for each of our original d dimensional data samples. The end result is that we use $\mathbf{U}_k$ as our new data, and it represents an approximated and transformed $\mathbf{X}$.