Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Fall 2020

Writing Assignment 3

Issued: Saturday 14th November, 2020 **Due:** Friday 27th November, 2020

3.1. (K-means) Given input data $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$, $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, the k-means clustering partitions the input into k sets $C_1, \ldots, C_k$ to minimize the within-cluster sum of squares:

$$\arg\min_{C} \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2,$$

where $\boldsymbol{\mu}_j$ is the center of the j-th cluster:

$$\boldsymbol{\mu}_j \stackrel{\text{def}}{=} \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}, \quad j = 1, \ldots, k.$$

(a) i. (2 points) Show that the k-means clustering problem is equivalent to minimizing the pairwise squared deviation between points in the same cluster:

$$\sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in C_j} \|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

ii. (2 points) Show that the k-means clustering problem is equivalent to maximizing the between-cluster sum of squares:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j| \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2.$$

(b) Define the distortion of k-means clustering as

$$J(\{c^{(i)}\}_{i=1}^{m}, \{\boldsymbol{\mu}_j\}_{j=1}^{k}) = \sum_{i=1}^{m} \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}\|^2.$$

i. (0.5 points) Show that the distortion J does not increase in each step of Lloyd's algorithm (refer to the lecture slides).

ii. (0.5 points) Does this algorithm always converge? Prove it or give a counterexample.

Solution:

$$\begin{aligned}
\sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2 &= \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} \left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{\mu}_j\|^2 - 2\langle \boldsymbol{x}, \boldsymbol{\mu}_j \rangle \right) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|^2 + \sum_{j=1}^{k} \left(|C_j| \|\boldsymbol{\mu}_j\|^2 - 2 \left\langle \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}, \boldsymbol{\mu}_j \right\rangle \right) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|^2 + \sum_{j=1}^{k} \left(|C_j| \|\boldsymbol{\mu}_j\|^2 - 2|C_j| \langle \boldsymbol{\mu}_j, \boldsymbol{\mu}_j \rangle \right) \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|^2 - \sum_{j=1}^{k} |C_j| \|\boldsymbol{\mu}_j\|^2.
\end{aligned}$$

Note that the first term is independent of the clustering results. Hence, the k-means algorithm is equivalent to maximizing

$$\sum_{j=1}^{k} |C_j| \|\boldsymbol{\mu}_j\|^2.$$

(a) i. Since

$$\begin{aligned}\sum_{\boldsymbol{x},\boldsymbol{x}'\in C_j} \|\boldsymbol{x}-\boldsymbol{x}'\|^2 &= \sum_{\boldsymbol{x}\in C_j}\sum_{\boldsymbol{x}'\in C_j} \left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{x}'\|^2 - 2\langle \boldsymbol{x},\boldsymbol{x}'\rangle\right) \\ &= 2|C_j| \sum_{\boldsymbol{x}\in C_j} \|\boldsymbol{x}\|^2 - 2\left\langle \sum_{\boldsymbol{x}\in C_j} \boldsymbol{x}, \sum_{\boldsymbol{x}\in C_j} \boldsymbol{x}' \right\rangle \\ &= 2|C_j| \sum_{\boldsymbol{x}\in C_j} \|\boldsymbol{x}\|^2 - 2|C_j|^2 \|\boldsymbol{\mu}_j\|^2,\end{aligned}$$

we have

$$\begin{aligned}\sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{\boldsymbol{x},\boldsymbol{x}'\in C_j} \|\boldsymbol{x}-\boldsymbol{x}'\|^2 &= \sum_{j=1}^{k} \frac{1}{2|C_j|} \left(2|C_j| \sum_{\boldsymbol{x}\in C_j} \|\boldsymbol{x}\|^2 - 2|C_j|^2\|\boldsymbol{\mu}_j\|^2\right) \\ &= \sum_{j=1}^{k} \left(\sum_{\boldsymbol{x}\in C_j} \|\boldsymbol{x}\|^2 - |C_j|\|\boldsymbol{\mu}_j\|^2\right) \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|^2 - \sum_{j=1}^{k} |C_j|\|\boldsymbol{\mu}_j\|^2,\end{aligned}$$

which is the same as the original objective function.

ii. Note that

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{k} |C_i||C_j| \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 &= \sum_{i=1}^{k}\sum_{j=1}^{k} |C_i||C_j| \left(\|\boldsymbol{\mu}_i\|^2 + \|\boldsymbol{\mu}_j\|^2 - 2\langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j\rangle\right) \\ &= \sum_{i=1}^{k} |C_i|\|\boldsymbol{\mu}_i\|^2 \sum_{j=1}^{k} |C_j| + \sum_{i=1}^{k} |C_i| \sum_{j=1}^{k} |C_j| \|\boldsymbol{\mu}_j\|^2 \\ &\quad - 2\left\langle \sum_{i=1}^{k} |C_i|\boldsymbol{\mu}_i, \sum_{i=1}^{k} |C_i|\boldsymbol{\mu}_j \right\rangle \\ &= 2|\mathcal{X}| \sum_{i=1}^{k} |C_i|\|\boldsymbol{\mu}_i\|^2 - 2|\mathcal{X}|^2\|\boldsymbol{\mu}\|^2\end{aligned}$$

where $\boldsymbol{\mu}$ is defined as the mean of all samples, i.e.,

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x},$$

and we have used the fact that

$$\sum_{i=1}^{k} |C_i|\boldsymbol{\mu}_i = \sum_{i=1}^{k} \sum_{\boldsymbol{x}\in C_i} \boldsymbol{x} = \sum_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{x} = |\mathcal{X}|\boldsymbol{\mu}.$$

Again, since the second term is independent of the choice of clustering, maximizing this object is equivalent to maximizing the first term, which is equivalent to the original optimization problem as illustrated at the beginning.

(b) i. The Lloyd's algorithm contains two steps:

$$c^{(i)} \leftarrow \arg\min_j \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j\|^2, i = 1, \ldots, m,$$

$$\boldsymbol{\mu}_j \leftarrow \frac{\sum_{i=1}^m \mathbf{1}\{c^{(i)} = j\}\boldsymbol{x}^{(i)}}{\sum_{i=1}^m \mathbf{1}\{c^{(i)} = j\}}, j = 1, \ldots, k.$$

In the first step, denote the updated c as $\hat{c}$, then we have

$$\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}\| \geq \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{\hat{c}^{(i)}}\|, \forall i = 1, \ldots, m.$$

Taking sum over i yields

$$J(\{c^{(i)}\}_{i=1}^m, \{\boldsymbol{\mu}_j\}_{j=1}^k) \geq J(\{\hat{c}^{(i)}\}_{i=1}^m, \{\boldsymbol{\mu}_j\}_{j=1}^k).$$

Similarly, in the second step, we have

$$\begin{aligned} J(\{c^{(i)}\}_{i=1}^m, \{\boldsymbol{\mu}_j\}_{j=1}^k) &= \sum_{i=1}^m \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}\|^2 \\ &= \sum_{j=1}^k \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2 \\ &\geq \sum_{j=1}^k \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j\|^2 \\ &= J(\{c^{(i)}\}_{i=1}^m, \{\hat{\boldsymbol{\mu}}_j\}_{j=1}^k), \end{aligned}$$

where $\hat{\boldsymbol{\mu}}$ is the updated $\boldsymbol{\mu}$.

As a result, the distortion J does not increase in each step.

ii. We have shown that the distortion $J > 0$ does not grow in each iteration; thus J must converge. To make the algorithm converge, we need one further assumption.

Note that the value of $\arg\min$ in the first step of the algorithm can be non-unique, which can eventually give different clustering results. Suppose that in this situation, we will always choose the smallest j that minimizes $\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j\|^2$. With this assumption, the convergence of J will imply the convergence of c and $\boldsymbol{\mu}$, since $(c, \boldsymbol{\mu})$ will be updated, if and only if the updated J is strictly less than the previous one.

3.2. (PCA) We will talk about a natural way to define PCA called Projection Residual Minimization. Suppose we have m samples $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)} \in \mathbb{R}^n\}$, then we try

to use the projections or image vectors to represent the original data. There will be some errors (projection residuals) and naturally we hope to minimize such errors.

(a) (1 point) First consider the case with one-dimentional projections. Let $\boldsymbol{u}$ be a non-zero unit vector. The projection of sample $\boldsymbol{x}^{(i)}$ on vector $\boldsymbol{u}$ is represented by $(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}$. Therefore the residual of a projection will be

$$\left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|$$

Please show that

$$\arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2 = \arg\max_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2$$

(b) (1 point) Follow the proof above and the discussion of the variance of projections in the lecture. Please show that minimizing the residual of projections is equivalent to finding the largest eigenvector of covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{u}^\star = \arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \frac{1}{m}\sum_{i=1}^{m} \left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2$$

then $\boldsymbol{u}^\star$ is the largest eigenvector of $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$

(c) (1 point) Now for a n-dimensional projection where the basis is a complete orthonormal set $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ that satisfies $\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

If we pick up a k-dimension projection, the residual will be the linear combination of the remaining bases.

$$\boldsymbol{x}^{(i)} - \sum_{j=1}^{k}(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j = \sum_{j=k+1}^{n}(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j$$

Please show that

$$\min_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k:\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}} \frac{1}{m}\sum_{i=1}^{m} \left\|\boldsymbol{x}^{(i)} - \sum_{j=1}^{k}(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j\right\|^2 = \sum_{i=k+1}^{n}\lambda_i,$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ is the eigenvalues of $\boldsymbol{\Sigma}$. It leads to the conclusion that the minimum average projection error is the sum of the eigenvalues of those eigenvectors that are orthogonal to the principal subspace .

Solution:

(a)

$$\left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2 = \boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(i)} - \left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2$$

Since $\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(i)}$ is a constant,

$$\arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2 = \arg\max_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2$$

(b) Therefore

$$\begin{aligned}\arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(i)}-(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2 &= \arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(i)}-\left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2 \\ &= \arg\max_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2 \\ &= \arg\max_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u}\end{aligned}$$

From Generalized Lagrange function that has been introduced in the class, we can find $\boldsymbol{u}^\star$ is the largest eigenvector.

(c)

$$\begin{aligned}&\min_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k:\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}} \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}^{(i)}-\sum_{j=1}^{k}\left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j\right)\boldsymbol{u}_j\right\|^2 \\ =&\min_{\boldsymbol{u}_{k+1},\ldots,\boldsymbol{u}_n:\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=k+1}^{n}\left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j\right)\boldsymbol{u}_j\right)^2 \\ =&\min_{\boldsymbol{u}_{k+1},\ldots,\boldsymbol{u}_n:\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=k+1}^{n}\left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j\right)^2 \\ =&\min_{\boldsymbol{u}_{k+1},\ldots,\boldsymbol{u}_n:\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}} \sum_{j=k+1}^{n}\boldsymbol{u}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u}_j\end{aligned}$$

Follow how we find the j-th principal component in the class, you will realize it is the summation of $(n-k)$ least eigenvalues. However the last step above is not that strict. The strict proof is a little bit complex with Lagrangian Multiplier Method. The good news is the conclusion is obvious. Let us omit the proof.

3.3. (Kernel PCA 2 ponit) Show that the conventional linear PCA algorithm is recovered as a special case of kernel PCA if we choose the linear kernel function given by $k(\boldsymbol{x},\boldsymbol{x}')=\phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}')=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'$.

Solution: The proof is also not that obvious. Thus, we want to introduce an intuitionistic explanation. It would be easy to understand. At least it is a common case that satisfies $\phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}')=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'$.

$$\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\mathrm{T}}=\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)})^{\mathrm{T}}$$

A bit more strict proof is to verify that $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ have the same eigenvalue.

$$\boldsymbol{\Sigma}_1=\frac{1}{m}\sum_{i=1}^{m}\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)})^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}_2 = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\mathrm{T}}$$

Note that the eigenvectors of $\boldsymbol{\Sigma_1}$ can be written as a linear combination of $\phi(\boldsymbol{x}^{(i)})$ and the eigenvectors of $\boldsymbol{\Sigma_2}$ can be written as a linear combination of $\boldsymbol{x}^{(i)}$

$$\begin{aligned}
\phi(\boldsymbol{x}^{(j)\mathrm{T}})\boldsymbol{\Sigma}_1\phi(\boldsymbol{x}^{(j)}) &= \frac{1}{m}\sum_{i=1}^{m} \phi(\boldsymbol{x}^{(j)\mathrm{T}})(\phi(\boldsymbol{x}^{(i)})\phi(\boldsymbol{x}^{(i)\mathrm{T}}))\phi(\boldsymbol{x}^{(j)}) \\
&= \frac{1}{m}\sum_{i=1}^{m} (\phi(\boldsymbol{x}^{(j)\mathrm{T}})\phi(\boldsymbol{x}^{(i)}))(\phi(\boldsymbol{x}^{(i)\mathrm{T}})\phi(\boldsymbol{x}^{(j)})) \\
&= \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{x}^{(j)\mathrm{T}}\boldsymbol{x}^{(i)})(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(j)}) \\
&= \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(j)\mathrm{T}}(\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\mathrm{T}})\boldsymbol{x}^{(j)} \\
&= \boldsymbol{x}^{(j)\mathrm{T}}\boldsymbol{\Sigma}_2\boldsymbol{x}^{(j)}
\end{aligned}$$

Therefore the sample covariance matrix is still the linear PCA covariance matrix.

3.4. (Bonus Question) (SVD) In the CCA and maximal correlation lecture, we used singular value decomposition (SVD)[1] to extract important features from data. The following exercise explores several properties of SVD in details.

Suppose a rank-r matrix $A \in \mathbb{R}^{m\times n}$ has the singular value decomposition: $A = U\Sigma V^{\mathrm{T}}$, where $U = [u_1, \ldots, u_r] \in \mathbb{R}^{m\times r}, \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), V = [v_1, \ldots, v_r] \in \mathbb{R}^{n\times r}, U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_r, \sigma_1 \geq \cdots \geq \sigma_r > 0$.

(a) (1 point) Show that $Av_i = \sigma_i u_i, A^{\mathrm{T}}u_i = \sigma_i v_i, i = 1, \ldots, r$.

(b) (1 point) The 2-norm of A is defined as

$$\|A\|_2 \overset{\text{def}}{=} \max_{x\in\mathbb{R}^n:\ \|x\|>0} \frac{\|Ax\|}{\|x\|}.$$

Prove that $\|A\|_2 = \sigma_1$. (*Hint: If $U^{\mathrm{T}}U = I$, then $\|Ux\| = \|x\|$.*)

Solution:

(a) Let e_i denote the i-th standard basis in $\mathbb{R}^r$, then

$$Av_i = U\Sigma V^{\mathrm{T}}v_i = U\Sigma e_i = \sigma_i U e_i = \sigma_i u_i.$$

Similarly, we have $A^{\mathrm{T}}u_i = \sigma_i v_i$.

[1] See `https://en.wikipedia.org/wiki/Singular_value_decomposition` for reference on SVD.

(b) $\forall\, x \in \mathbb{R}^n \setminus 0$,

$$\|Ax\| = \|U\Sigma V^{\mathrm{T}} x\| = \|\Sigma V^{\mathrm{T}} x\|$$

Let $\alpha = V^{\mathrm{T}} x \in \mathbb{R}^r$, then we can show that

$$\|\Sigma\alpha\| \leq \sigma_1 \|\alpha\| \leq \sigma_1 \|x\|.$$

To prove the first inequality, assume $\alpha = (\alpha_1, \ldots, \alpha_r)^{\mathrm{T}}$, then

$$\|\Sigma\alpha\| = \sqrt{\sum_{i=1}^{r} \sigma_i^2 \alpha_i^2} \leq \sqrt{\sum_{i=1}^{r} \sigma_1^2 \alpha_i^2} = \sigma_1 \|\alpha\|.$$

To prove the second inequality, we can expand $\{v_1, \ldots, v_r\}$ to an orthonormal basis $\{v_1, \ldots, v_r, \ldots, v_n\}$, then

$$\|\alpha\|^2 = \|V^{\mathrm{T}} x\|^2 = \sum_{i=1}^{r} \langle v_i, x\rangle^2 \leq \sum_{i=1}^{n} \langle v_i, x\rangle^2 = \|x\|^2.$$

As a consequence, we have

$$\max_{x \in \mathbb{R}^n :\, \|x\| > 0} \frac{\|Ax\|}{\|x\|} \leq \sigma_1 = \frac{\|Av_1\|}{\|v_1\|} \leq \max_{x \in \mathbb{R}^n :\, \|x\| > 0} \frac{\|Ax\|}{\|x\|},$$

which means

$$\max_{x \in \mathbb{R}^n :\, \|x\| > 0} \frac{\|Ax\|}{\|x\|} = \sigma_1.$$