Lecturer: SVD and Tensors

Scribe: Ching-An Cheng, Shanshan Cao

1 Random projection vs. optimal projection

Let $A \in \mathbb{R}^{m \times n}$, where each row presents a point in $\mathbb{R}^n$. The goal of random/optimal projection is to find a rank-k matrix $\tilde{A} \in \mathbb{R}^{m \times n}$ with $k < n$ such that $\tilde{A}$ preserves the properties of A.

1.1 Review on linear algebra

Definition 1.1. *Given a set of vectors $a_1, \cdots, a_n \in \mathbb{R}^m$, the Gram matrix $X \in \mathbb{R}^{m \times m}$ of the vector set is defined as $X_{ij} = \langle a_i, a_j \rangle$.*

The Gram matrix preserves the pairwise distance bound.

Definition 1.2 (Eigenvectors). *$v \in \mathbb{R}^n$ is said to be an eigenvector of a square matrix $A \in \mathbb{R}^{n \times n}$ if there exists a $\lambda \in \mathbb{R}$ such that:*

$$Av = \lambda v.$$

Then λ is said to be an eigenvalue of A.

Definition 1.3 (Singular vectors). *Let $A \in \mathbb{R}^{m \times n}$, if there exists $\sigma \in \mathbb{R}$, $u \in \mathbb{R}^m$, and $v \in \mathbb{R}^n$ with $\|u\| = \|v\| = 1$, such that:*

$$Av = \sigma u, \quad A^T u = \sigma v.$$

Then σ is said to be a singular value of A, and (u, v) are the (left, right) singular vectors.

Fact 1. *Any right singular vector of A is an eigenvector of $A^T A$.*

Proof. Let v be the right singular vector of A corresponding to singular value σ. Thus:

$$A^T A v = A^T (\sigma u) = \sigma^2 v.$$

□

Exercise 1.1. *If matrix B is a real symmetric square matrix, then all the eigenvalues of B are real. Furthermore, if $B = A^T A$, all its eigenvalues are larger than equal or to zero (i.e. B is semi-positive definite).*

Let, v_1, the top right singular vector be defined as follows:

$$v_1 = \arg\max_{\|v\|=1} \|Av\|^2 = \arg\max_{\|v\|=1} \sum_{i=1}^{m} \langle A_i, v\rangle^2.$$

Then we have:

$$\arg\max_{\|v\|=1} \|Av\|^2 = \arg\max_{\|v\|=1} \sum_{i=1}^{m} \|A_i\|^2 - d(A_i, v)^2 = \arg\min_{\|v\|=1} \sum_{i=1}^{m} d(A_i, v)^2,$$

where $d(A_i, v)^2 = 1 + \|A_i\|^2 - \langle A_i, v\rangle^2$, is the distance from A_i to the 1-D subspace of span$\{v\}$.

In the above problem, the objective function is a function $f(v) = \|Av\|^2$ on unit sphere, and therefore at its local optimum the gradient is pointing along v. Thus, v is an singular vector of A^TA:

$$\nabla f(v) = \lambda v = 2A^TAv.$$

The top singular vector is the best fit in 1-dimensional subspace in terms of least squares error. In the following theorem, we will show the span of the top k singular vectors is the best fit in k-dimensional subspace.

Theorem 1.1. *Let* $v_1 = \arg\max_{\|v_1\|=1} \|Av\|^2$. *Let* $v_2 = \arg\max_{\|v_2\|=1, v_2 \perp v_1} \|Av\|^2$ *... Let* $v_k = \arg\max_{\|v_k\|=1, v_k \perp v_i, for\ i=1,..,k-1} \|Av\|^2$ *Then* $V_k = span\{v_1, ..., v_k\}$ *is the best fit in k-dimensional subspace.*

$$V_k = \arg\max_{dim(V)\le k} \sum_{i=1}^{m} \|\Pi_V(A_i)\|^2 = \arg\min_{dim(V)\le k} \sum_{i=1}^{m} d^2(A_i, v)$$

where Π_V *is the projection on the subspace* V.

Proof. k =1 is true.
Suppose the optimal subspace is $W_2 = \text{span}\{w_1, w_2\}$, where w_1 and w_2 are orthonormal. WLOG, we can pick $w_2 \perp v_1$ (since W_2 is 2-D while span$\{v_1\}$ is 1-D, in W_2, we can always find a vector w_2 which is orthogonal to v_1).

$$\|\Pi_W(A)\|^2 = \|Aw_1\|^2 + \|Aw_2\|^2 \le \|Av_1\|^2 + \|Aw_2\|^2$$

Then

$$w_2 = \arg\max_{v \perp v_1} \|Av\|^2$$

For general k-dimensional space $W_k = spw_1, ..., w_k$, it can proved similarly by picking $w_k \perp v_1, .., v_{k-1}$. Suppose the optimal k-dimensional subspace is $W_k = \text{span}\{w_1, w_2, \cdots, w_k\}$, where $\{w_i | i = 1, \ldots, k\}$ are orthonormal. WLOG, we can pick $w_k \perp v_i, i = 1, \cdots, k-1$ (similar to the argument for $k = 2$).

$$\|\Pi_W(A)\|^2 = \sum_{i=1}^{k-1} \|Aw_i\|^2 + \|Aw_k\|^2 \le \sum_{i=1}^{k-1} \|Av_i\|^2 + \|Aw_k\|^2$$

Then

$$w_k = \arg\max_{v \perp v_i, i=1,\cdots,k-1} \|Av\|^2$$

□

Theorem 1.2 (SVD). *Let $m \geq n$. For every matrix $A \in \mathbb{R}^{m \times n}$, there exists sets of orthonormal vector $\{v_i \in \mathbb{R}^n\}$ $\{u_i \in \mathbb{R}^m\}$ such that $A = \sum_{i=1}^n \sigma_i u_i v_i^T$ and $\sigma_i \geq 0$.*

Proof. Let $\{v_i \in \mathbb{R}^n\}$ be orthonormal and complete. Suppose $\{u_i \in \mathbb{R}^m\}$ satisfies $A = \sum_{i=1}^n \sigma_i u_i v_i^T$. Then first, $Av_j = \sigma_j u_j$ and therefore v_j is a singular vector. Second, $\{u_i \in \mathbb{R}^m\}$ is orthonormal, because $u_i^T u_j = \frac{1}{\sigma_i \sigma_j} v_j^T (A^T A) v_i = \frac{\sigma_i^2}{\sigma_i \sigma_j} v_j^T v_i = \delta_{i,j}$ which implies that the set is orthogonal. The sign of σ_i can be arbitrarily chosen and therefore can be chosen such that $\sigma_i \geq 0$

□

Theorem 1.3 (Eckart-Young). *Suppose $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$.*

$$\min_{D \in \mathbb{R}^{m \times n}, rank(D) \leq k} \|A - D\|_F^2 = \|A - A_k\|_F^2 = \| \sum_{i=k+1}^n \sigma_i u_i v_i^T \|_F^2 = \sum_{i=k+1}^n \sigma_k^2$$

Proof. According to theorems 1.1 and 1.2, A_k should be the projection of A on the span of $\{v_1, ..., v_k\}$, which is exactly: $AV_k V_k^T = A_k$. Thus

$$\|A - A_k\|_F^2 = \min_{D \in \mathbb{R}^{m \times n}, rank(D) \leq k} \|A - D\|_F^2 = \sum_{i=k+1}^n \sigma_k^2$$

□

Exercise 1.2 (Power iteration). *Let x_0 be a random unit vector. Let*

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|}.$$

Then the sequence converges to the vector corresponding to the top eigenvalue after $O(\log n/\epsilon)$ iterations, which means

$$\|Ax_k\| \geq (1 - \epsilon)\|Av_1\|.$$

2 Gaussian distribution

2.1 Covariance matrix of Gaussian

Let x be a Gaussian random variable with density

$$p(x) = \frac{1}{(\sqrt{2\pi})^n \sqrt{det(\Sigma)}} e^{-\frac{1}{2} x^T \Sigma^{-1} x}$$

where $\Sigma = E((x - \mu)(x - \mu)^T)$. Suppose $\mu = 0$. Then $\Sigma_{ij} = \sum E(x_i x_j)$ and along singular vector v of Σ, $E((v^T x)^2) = v^T \Sigma v$. That is, a zero mean Gaussian can be written as a product distribution of its principal components

$$p(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\lambda_i}} e^{-\frac{1}{2}\lambda_i^{-1}(v_i^T x)^2},$$

where $\lambda_i, i = 1, \ldots, n$, are the singular values of Σ and $v_i, i = 1, ..., n$ are the corresponding singular vectors.

2.2 Clustering and Learning a Mixture of Gaussian Distribution

Consider a mixture of Gaussian $F = w_1F_1 + ... + w_kF_k$, where $F_i = N(\mu_i, \sigma_i^2 I)$, $w_i \geq 0$, and $\sum w_i = 1$ (i.e. the Gaussian is assumed spherical and separable). Let $x_1, \ldots, x_m \in \mathbb{R}^n$ be i.i.d. samples from F.

Let $D = \{x_1, .., x_m\}$ be the samples from F. Given μ_i, σ_i, and w_i, the objective of clustering is to partition D according to the source of generation F_i. On the other hand, the objective of learning is to find the set of parameter μ_i, σ_i, and w_i such that it best describes the observations D.

2.2.1 Clustering

Suppose $n = 1$. To satisfy

$$P(|x - \mu| > t\sigma) \leq 2e^{-t^2/2} = \frac{2}{m}.$$

we take $t = \sqrt{2 \log m}$, i.e.$\|\mu_i - \mu_j\|^2 \geq \Omega(1)(\sigma_1^2 + \sigma_2^2)$. For general n, for $x \sim F_i$, $y \sim F_j$,

$$E(\|x - y\|^2) = n\sigma_i^2 + n\sigma_j^2 + \|\mu_i - \mu_j\|^2$$

Lemma 2.1.

$$P(|\|x - \mu\|^2 - n\sigma^2| > t\sqrt{n}\sigma^2) < 2e^{-t^2/8}$$

By the lemma, with high probability

$$\sqrt{n}\sigma - c\sigma \leq \|x - \mu\| \leq \sqrt{n}\sigma + c\sigma$$

Therefore, we can choose

$$\|\mu_i - \mu_j\|^2 \geq C\sqrt{n}(\sigma_i^2 + \sigma_j^2)$$

to separate pairwise distance. However, there is $\sqrt{n}$ compared to the 2-dimensional case.

The separability condition scales in $\sqrt{n}$, which can be larger than k. The following algorithm uses projection on a k dimensional subspace to decrease the factor.

Projection Algorithm

1. Center all the data (zero mean)
2. Project to top (k-1)-dimensional subspace
3. Clustering using pairwise distance in the projected subspace

The algorithm without the centering step replaces the second step with projection onto the top k-dimensional subspace.

Theorem 2.2. *After centering, $\{\mu_1, ..., \mu_k\} \subseteq V_{k-1}$, where V_{k-1} is the space of the top $k-1$ principle components of PCA.*

Proof. For $k = 1$, the first one is along its mean. The rest is symmetric and therefore can be chosen arbitrarily. For k, it should cover span$\{\mu_1, ..., \mu_k\}$ □

Therefore, for clustering, we have a relaxed condition

$$\|\mu_i - \mu_j\|^2 \geq C\sqrt{k}(\sigma_i^2 + \sigma_j^2)$$

Note that for general non-isotropic Gaussian mixtures PCA may not preserve the mean.

Theorem 2.3. *If each μ_i is separated from the span of the rest of the $\mu_j's$, and*

$$d(\mu_i, sp\mu_1, ..., \mu_k \backslash \{\mu_i\}) \geq C\sigma_i^2.$$

Then there exists a polynomial-time algorithm.

2.2.2 Learning

In contrast to clustering, learning does not requires separability.

Theorem 2.4. *Mixture of Gaussians are uniquely identifiable.*

Theorem 2.5. $O(n^{O(f(k))})$ *time and samples are needed to estimate w_i, μ_i, σ_i.*

Theorem 2.6. $O(2^{\Omega(k)})$ *samples, if time is not concerned (may be exponential)*

Theorem 2.7. *For spherical Gaussian, $\mu_1, \dots \mu_k$ are linearly independent, there is polynomial(n,k,1/ϵ) algorithm to estimate w_i, μ_i, σ_i within error ϵ.*

Using tensor algebra, the mean can also be identified using the following identities

$$E(X) = \sum_{i=1}^{k} w_i \mu_i$$

$$E(X \otimes X) = \sum_{i=1}^{k} w_i \sigma_i^2 I + w_i \mu_i \mu_i^T$$

$$E(X \otimes X \otimes X) = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i + \sum_{j=1}^{n} v \otimes e_j \otimes e_j + e_j \otimes v \otimes e_j + e_j \otimes e_J \otimes v_j$$

where $v = \sum_{i=1}^{k} w_i \sigma_i^2 \mu_i$ and e_j is the jth canonical unit vector.

Theorem 2.8. *Let $T = \sum \alpha_i \mu_i \otimes \mu_i \otimes \mu_i$. If μ_i is orthonormal, then given T, μ_i can be identified in polynomial time.*