DATA STRUCTURE AND ALGORITHM FOR MASSIVE DATASET

MINSEOK SONG

Three ways to deal with massive dataset

(1) Dimensional reduction: the purpose is to minimize the loss of information.
(2) Compressed representation: present data in a compact form, but not necessarily predicated on the retainment of information. i.e., it may prefer higher compression rates.
(3) Interpolation: only use discrete information of the distribution f. This is useful since we do not have a full function f available. Remember we used finite element method in numerical PDE, and the right space of function to discuss numerical stability etc was Sobolev space.

- All in all, it focuses on achieving lower computational/statistical complexity.
- To clarify, computational complexity deals with the resources(time and space), while statistical complexity with the intricacy of models(in the sense of how simpler model represents reduced data).

Theorem 1. *(Johnson-Lindenstrauss Lemma) Let Q be a finite set of vectors in $\mathbb{R}^d$. Let $\delta \in (0,1)$ and n be large enough integer such that*

$$\epsilon = \sqrt{\frac{6\log(2|Q|/\delta)}{n}} \leqslant 3 \tag{1}$$

With probability of at least $1-\delta$ over a choice of a random matrix $W \in \mathbb{R}^{n,d}$ such that each element of W is distributed normally with zero mean and variance of $1/n$ we have

$$\sup_{x\in\mathbb{Q}} \left|\frac{\|Wx\|^2}{\|x\|^2} - 1\right| < \epsilon \tag{2}$$

- The proof leans on the following lemma, which uses the concentration property of χ^2.

Lemma 2. *Fix some $x \in \mathbb{R}^d$. Let $W \in \mathbb{R}^{n,d}$ be a random matrix such that each $W_{i,j}$ is an independent normal random variable. Then, for every $\epsilon \in (0,3)$ we have*

$$\mathbb{P}\left[\left|\frac{\|(1/\sqrt{n})Wx\|}{\|x\|} - 1\right| > \epsilon\right] \leqslant 2e^{-\epsilon^2 n/6} \tag{3}$$

- Note that $W : \mathbb{R}^d \to \mathbb{R}^n$, and the result does not depend on d. This suggests that we can conduct dimensionality reduction in very high-dimensional spaces without much cost(!).

Proof of Lemma 2. We can assume, WLOG, that $\|x\|^2 = 1$. Do note that $\|Wx\|^2$ has a χ_n^2 distribution by construction, so we may use concentration of χ^2 inequality to get the result. □

Proof. In order to deal with $|Q|$, use the union bound. We can find appropriate ϵ afterward. □

- This says that the random projections do not distort Euclidean distances too much.

PCA We aim at solving the problem

$$\arg \min_{W\in\mathbb{R}^{n,d}, U\in\mathbb{R}^{,\ltimes}} \sum_{i=1}^{m} \|x_i - UWx_i\|_2^2$$

- It is shown that $W = U^T$ and U is orthonormal.
- It is then shown that the optimal solution is caculated by computing the eigenvectors of $A = \sum_{i=1}^m x_i x_i^T = X^T X$. This is the right eigenvectors of SVD. Do note that x_i is each column of X.
- This means the complexity is given by $O(d^3 + md^2)$
  (1) $O(d^3)$ for computing the eigenvectors and eigenvalues of A.
  (2) $O(d^2 m)$ for computing the covariance matrix A.
- Instead of using XX^T, we can use the eigenvector of $X^T X$, that is, $A(X^T u) = \lambda(X^T u)$ where u is an eigenvector of B.
- This comes from the fact that $X^T X X^T u = \lambda X^T u$.
- Do note that B only requires inner products $\langle x_i, x_j \rangle$.
- This reduces our complexity to $O(m^3 + dm^2)$, which is useful when d is very large.

Compressed Sensing

Definition 1. *A matrix $W \in \mathbb{R}^{n,d}$ is $(\epsilon, s) - RIP$ if for all $x \neq 0$ s.t. $\|x\|_0 \leqslant s$ we have*

$$\left|\frac{\|Wx\|_2^2}{\|x\|_2^2} - 1\right| \leqslant \epsilon$$

- A particular theorem states that if W is an RIP (Restricted Isometry Property) matrix, then under certain conditions, the expression $\arg\min_{v:Wv=Wx} \|v\|_0$ evaluates to x. This implies that, for specific compression matrices and sparse data, the original data can be accurately recovered.
- Other theorems deal with L^1; it's because we are looking for solutions that are "almost sparse" rather than strictly sparse.