Com S 631: Dimensionality Reduction

Lectures 14,15 Scribe: Madhavan

1 Dimensionality Reduction

Let S be a set of data points in high dimension space i.e. $S \subseteq \mathbb{R}^m$. The objective is to map S to a low dimension space while preserving certain *properties* of the data points. For example, consider the Euclidean distance of two points. Let $x, y \in \mathbb{R}^m$ and let $f : \mathbb{R}^m \to \mathbb{R}^p$ where $p << m$. We would like to design f such that $\|x - y\|_2 \approx f(x) - f(y)$.

1.1 AMS algorithm for estimating F2

Let us revisit the AMS algorithm for estimating the second frequency moment. Consider a stream Z over $U = \{1..m\}$. Let f_i be the number of times that element i appears. The second frequency moment $F_2(Z) = \sum_{i=1}^{m} f_i^2$. Let $F =< f_1, f_2, .., f_m >$ be a vector in $\mathbb{R}^m$. Equivalently, define $F_2(Z) = \|F\|_2^2$.

Algorithm 1 AMS Algorithm With Repititions

Input Stream Z over $U = \{1..m\}$
Output Estimate of $F_2(Z)$

procedure AMS
  For $i = 1..k$, generate 4-wise independent hash functions $h_i : U \to \{+1, -1\}$
  for $x \in S$ **do**
    for $i = 1$ to k **do**
      $c_i = c_i + h(x)$
    end for
  end for
  $\hat{F_2}(Z) = \frac{\sum_{i=1}^{k} c_i^2}{k}$
  return $\hat{F_2}(Z)$
end procedure

Idea: The above AMS algorithm is performing a dimensionality reduction. We will show that the estimate $\hat{F_2}(Z)$ is close to $F_2(Z)$ with high probability. If that is the case then we are mapping a vector $F =< f_1, f_2.., f_m > \to < \frac{c_1}{\sqrt{k}}, \frac{c_2}{\sqrt{k}}..\frac{c_k}{\sqrt{k}} >= C$.

Theorem 1. *If $k = O(\frac{1}{\epsilon^2})$ then $P\left[\|C\|_2^2 \in (1 \pm \epsilon)\|F\|_2^2\right] \geq 1 - \delta \geq \frac{2}{3}$. Here, $\delta = \frac{1}{3}$*

Consider the hash function $h_i : U \to \{+1, -1\}$. Define the random variable $X_{ij} = 1$ if $h_i(j) = 1$, otherwise $X_{ij} = -1$. Therefore, we define $c_1 = \sum_{j=1}^{m} X_{1j} f_j$, $c_2 = \sum_{j=1}^{m} X_{2j} f_j$ and so on. We can it as follows:

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{km} \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix}$$

The $(k \times m)$ matrix is called as the AMS sketch. Observe that the vector F can be expressed as the difference between two vectors a and b. In other words, if $a, b \in \mathbb{R}^m$ then $(a-b) \in \mathbb{R}^m$. Consider $S \subseteq \mathbb{R}^m$. Using AMS with k hash functions, we are able to claim $P\left[\|C\|_2^2 \in (1 \pm \epsilon)\|F\|_2^2\right] \geq 1-\delta$. This event corresponds to the distance between one pair of vectors. Can we bound the union of the events that correspond to the distance between every pair of vectors in S.

$$\begin{aligned} P\left[\forall x, y \in S, \|x-y\|_2^2 \notin (1 \pm \epsilon)\|x-y\|_2^2\right] &\leq n(n-1) P\left[x, y \in S, \|x-y\|_2^2 \notin (1 \pm \epsilon)\|x-y\|_2^2\right] (Union Bound) \\ &\leq n^2 \times O\left(\frac{1}{\epsilon^2}\right) \times \frac{1}{k} \end{aligned} \tag{1}$$

To bound the above probability by δ, we need $k = n^2 O\left(\frac{1}{\epsilon^2}\right)$. Therefore, $\delta = O\left(\frac{1}{\delta\epsilon^2}\right)$. We can reduce it to $O\left(log(\frac{1}{\delta})\frac{1}{\epsilon^2}\right)$, thereby reducing the number of hash functions.

2 Johnson Lindenstrauss Theorem

The Johnson-Lindenstrauss theorem states that there exists a mapping $f : \mathbb{R}^m \to \mathbb{R}^k$ such that the Euclidean distance of any two points in $\mathbb{R}^m$ is approximately preserved. We will use the following lemma to arrive at a precise definition and proof of the theorem.

Lemma 1. *Let A be a $k \times m$ matrix where each entry is independently drawn from $N(0,1)$. Then, $\forall x \in \mathbb{R}^m$, $P\left[\|Ax\|_2 \in (1+\epsilon)\|x\|_2\right] > 1-\delta$ where $k = O\left(log(\frac{1}{\delta})\frac{1}{\epsilon^2}\right)$*

Lemma 1 is called the distributional JL-Lemma. We use this lemma to state and prove the JL theorem.

Theorem 2. *Let $S \subseteq \mathbb{R}^m$ with n points. There exists $k \times m$ matrix A such that $\forall x, y \in S$, $\|Ax - Ay\|_2 \in (1 \pm \epsilon)\|x-y\|_2$ where $k = O\left(\frac{1}{\epsilon^2} logn\right)$*

Proof. Let A be the Gaussian Matrix from JL Lemma. Since $|S| = n$, there are $\binom{n}{2}$ pairs of points in S. We have:

$$P\left[\forall x, y, \|Ax - Ay\|_2 \in (1 \pm \epsilon)\|x-y\|_2\right] = 1 - P\left[\exists x, y, s.t \|Ax - Ay\|_2 \notin (1 \pm \epsilon)\|x-y\|_2\right] \tag{2}$$

$$\begin{aligned} P\left[\exists x, y, s.t \|Ax - Ay\|_2 \notin (1 \pm \epsilon)\|x-y\|_2\right] &\leq \sum_{x,y} P\left[\|Ax - Ay\|_2 \notin (1 \pm \epsilon)\|x-y\|_2\right] \\ &\leq \sum_{x,y} \delta (\text{From JL Lemma}) \\ &\leq n^2\delta = \frac{1}{2} \end{aligned} \tag{3}$$

The final equality is true if $\delta = \frac{1}{2n^2}$.

$$P\left[\forall x, y, \|Ax - Ay\|_2 \in (1 \pm \epsilon)\|x-y\|_2\right] \geq \frac{1}{2} \tag{4}$$

Therefore, there exists a matrix A such that $P\left[\forall x, y, \|Ax - Ay\|_2 \in (1 \pm \epsilon)\|x-y\|_2\right]$ is at least half. □

We will review probability distribution over reals and then proceed to prove JL Lemma.

2.1 Continuous Random Variables, Probability Density Function

A random variable X is continuous if there exists a function f such that:

$$Pr\left[a \leq X \leq b\right] = \int_a^b f(x)dx \tag{5}$$

f is called the Probability Distribution Function (PDF). f satisfies the following properties:

1. $f(x) > 0, \forall x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

If X is a continuous random variable that takes the value returned by the PDF, its expectation is defined as:

$$E[X] = \int_{-\infty}^{\infty} f(x)xdx \tag{6}$$

Gaussian Distribution The Gaussian distribution (a.k.a Normal Distribution) is defined by the probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{7}$$

Where μ is the mean (or expectation), σ is the standard deviation. The Gaussian distribution is written as $N(\mu, \sigma^2)$. If a random variable X is drawn according to the PDF of a Gaussian distribution with mean (or expectation) μ and standard deviation σ, we write it as $X \sim N(\mu, \sigma^2)$. The distribution has the following properties:

1. The variance $Var(X) = \sigma^2$.
2. If $X \sim N(0, \sigma^2)$ $Var(X) = E[X^2] - E[X]^2 = E[X^2]$
3. If $X \sim N_1(\mu_1, \sigma_1^2)$ and $Y \sim N_2(\mu_2, \sigma_2^2)$ are independently drawn then:
   - $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
4. If c is a constant and $X \sim N(\mu, \sigma^2)$ then $cX \sim N(\mu, c^2\sigma^2)$

2.2 Moment Generating Function

The Moment Generating Function of a random variable X, where $\lambda \in \mathbb{R}$:

$$\begin{aligned} MGF(X) &= E[e^{\lambda X}] \\ &= E[1 + \lambda X + \frac{\lambda^2 X^2}{2!} + ...] \end{aligned} \tag{8}$$

$$\begin{aligned} Pr[X > t] &= Pr[e^{\lambda X} > e^{\lambda t}] \\ &\leq \frac{E[e^{\lambda X}]}{e^{\lambda t}} \text{(Using Markov)} \\ &= \frac{MGF(X)}{e^{\lambda t}} \end{aligned} \tag{9}$$

$$
\begin{aligned}
MGF(X) &= E[e^{\lambda X}] \\
&= \int_{-\infty}^{\infty} e^{\lambda x} e^{\frac{-x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2+2\sigma^2\lambda x}{2\sigma^2}} dx
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
\frac{-x^2+2\sigma^2\lambda x}{2\sigma^2} &= \frac{-x^2+2\sigma^2\lambda x-\sigma^4\lambda^2+\sigma^4\lambda^2}{2\sigma^2} \\
&= \frac{-(x-\sigma^2\lambda)^2}{2\sigma^2} + \frac{\sigma^2\lambda^2}{2}
\end{aligned} \tag{11}
$$

Therefore,

$$
\begin{aligned}
MGF(X) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\sigma^2\lambda)^2}{2\sigma^2}} e^{\frac{\sigma^2\lambda^2}{2}} dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\sigma^2\lambda)^2}{2\sigma^2}} dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{\infty} \text{PDF of} N(\sigma^2\lambda, \sigma^2) dx \\
&= e^{\frac{\sigma^2\lambda^2}{2}}
\end{aligned} \tag{12}
$$

Therefore, we have $MGF(X) = e^{\frac{\sigma^2\lambda^2}{2}}$. Suppose we draw $X \sim N(0, \sigma^2)$ and we want to find the probability $Pr[X > t]$?

$$
\begin{aligned}
Pr[X > t] &= Pr[e^{\lambda X} > e^{\lambda t}] \\
&\leq \frac{MGF(X)}{e^{\lambda t}} (\text{Markov bound}) \\
&= \frac{e^{\frac{\sigma^2\lambda^2}{2}}}{e^{\lambda t}} \\
&= e^{\frac{\sigma^2\lambda^2}{2} - \lambda t}
\end{aligned} \tag{13}
$$

The above function is minimized when $\lambda = \frac{t}{\sigma^2}$. Therefore, $Pr[X > t] \leq e^{\frac{-t^2}{2\sigma^2}}$. Similarly, $Pr[-X > t] \leq e^{\frac{-t^2}{2\sigma^2}}$. This leads to the following theorem.

Theorem 3. $Pr[|X| > t] \leq 2e^{\frac{-t^2}{2\sigma^2}}$

2.3 Chi-Squared Distribution

χ^2 distribution is the with k degrees of freedom is the sum of the squares of k independent random variables drawn a Gaussian distribution with $\mu = 0, \sigma = 1$. Formally, if $X_1.X_2...X_k$ are independent random variables drawn from $N(0, 1)$ then the χ^2 distribution is Z where:

$$
Z = \sum_{i=1}^{k} X_i^2 \tag{14}
$$

Theorem 4. *Let $Z = \sum_{i=1}^{k} X_i^2$ where $X_1, X_2..X_k$ are random variables drawn independently from $N(0,1)$. Then, $Pr[|Z-k| \geq \epsilon k] \leq 2e^{-k(\epsilon^2 - d\epsilon^3)}$ for some $d > 0$.*

Proof. The expectation of Z, $E[Z] = \sum_{i=1}^{k} E[X_i^2] = k$. This is because $E[X_i^2] = \sigma^2 = 1$.

$$\begin{aligned} Pr[Z > k + \epsilon k] &= Pr[e^{\lambda Z} > e^{\lambda(k+\epsilon k)}] \\ &\leq \frac{E[e^{\lambda Z}]}{e^{\lambda(k+\epsilon k)}} \text{(Using Markov)} \end{aligned} \tag{15}$$

$$\begin{aligned} E[e^{\lambda Z}] &= E[e^{\lambda(X_1^2 + X_2^2 + ... + X_k^2)}] \\ &= E[e^{\lambda X_1^2} e^{\lambda X_2^2} ... e^{\lambda X_k^2}] \\ &= E[e^{\lambda X_1^2}] \times E[e^{\lambda X_2^2}] ... \times E[e^{\lambda X_k^2}] \end{aligned} \tag{16}$$

The last equality comes from the fact that if the random variables are independent, the expectation of the product is the product of expectations.

$$\begin{aligned} E[e^{\lambda X^2}] &= \int_{-\infty}^{\infty} e^{\lambda X^2} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{2\lambda x^2 - x^2}{2}} dx \end{aligned} \tag{17}$$

$$\begin{aligned} e^{\frac{2\lambda x^2 - x^2}{2}} &= e^{\frac{-x^2(1-2\lambda)}{2}} \\ &= e^{\frac{-x^2}{2(1-2\lambda)^{-1}}} \end{aligned} \tag{18}$$

$$\begin{aligned} E[e^{\lambda X^2}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2(1-2\lambda)^{-1}}} dx \\ &= \int_{-\infty}^{\infty} \frac{\sqrt{(1-2\lambda)^{-1}}}{\sqrt{2\pi}\sqrt{(1-2\lambda)^{-1}}} e^{\frac{-x^2}{2(1-2\lambda)^{-1}}} dx \\ &= \frac{1}{\sqrt{(1-2\lambda)}} \int_{-\infty}^{\infty} \text{PDF of } N(0,(1-2\lambda)^{-1}) dx \\ &= \frac{1}{\sqrt{(1-2\lambda)}} \end{aligned} \tag{19}$$

Plugging this back in, we get,

$$E[e^{\lambda X^2}] = \left(\frac{1}{\sqrt{(1-2\lambda)}}\right)^k = \left(\frac{1}{1-2\lambda}\right)^{\frac{k}{2}} \tag{20}$$

$$\frac{MGF(Z)}{e^{\lambda(k+\epsilon k)}} \leq \left(\frac{1}{1-2\lambda}\right)^{\frac{k}{2}} \times \frac{1}{e^{\lambda(k+\epsilon k)}} \tag{21}$$

The expression is minimized when $\lambda = \frac{\epsilon}{2(1+\epsilon)}$. Then, $1 - 2\lambda = \frac{1}{1+\epsilon}$. Therefore,

$$
\begin{aligned}
Pr[Z-k \geq \epsilon k] \leq \frac{MGF(Z)}{e^{\lambda(k+\epsilon k)}} &\leq \frac{(1+\epsilon)^{\frac{k}{2}}}{e^{\lambda(k+\epsilon k)}} \\
&= \frac{(1+\epsilon)^{\frac{k}{2}}}{e^{\frac{\epsilon k}{2}}} \\
&= \left[e^{ln(1+\epsilon)} \times e^{-\epsilon}\right]^{\frac{k}{2}} \\
&= \left[e^{\epsilon-\frac{\epsilon^2}{2}+c\epsilon^3} \times e^{-\epsilon}\right]^{\frac{k}{2}} \text{(Taylor series expansion)} \\
&= e^{-k(\epsilon^2-d\epsilon^3)}
\end{aligned} \tag{22}
$$

Where d is some constant. We can similarly prove $Pr[-(Z-k) > \epsilon k] \leq e^{-k(\epsilon^2-d\epsilon^3)}$. Therefore, $Pr[|Z-k| \geq \epsilon k] \leq 2e^{-k(\epsilon^2-d\epsilon^3)}$ for some $d > 0$. □

2.4 Proving JL-Lemma

Recall that Lemma 1 states that if A is a $k \times m$ matrix whose elements are independently drawn from $N(0,1)$ then $\forall x \in \mathbb{R}^m$, $P\left[\|Ax\|_2 \in (1+\epsilon)\|x\|_2\right] > 1-\delta$ when $k = O\left(log(\frac{1}{\delta})\frac{1}{\epsilon^2}\right)$. We can write Ax as:

$$
\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{km} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix}
$$

We will prove for $\|x\|_2 = 1$. Observe that $c_1 = A_{11}x_1 + A_{12}x_2 ... + A_{1m}x_m$. $c_1 \sim N(0, x_1^2) + N(0, x_2^2) + ... + N(0, x_m^2)$. Therefore, $c_1 \sim N(0, x_1^2 + x_2^2 + ... + x_m^2) = N(0,1)$. Let $\|Ax\|_2^2 = Z = c_1^2 + c_2^2 + ... + c_k^2$. We have:

$$
\begin{aligned}
Pr\left[\left|\frac{\|Ax\|_2^2}{k} - 1\right| \geq \epsilon\right] &= Pr\left[|Z-k| \geq \epsilon k\right] \\
&\leq 2^{\frac{-k(\epsilon^2-d\epsilon^3)}{2}} \\
&\leq \delta
\end{aligned} \tag{23}
$$

The last inequality is true when $k = O\left(log(\frac{1}{\delta})\frac{1}{\epsilon^2}\right)$. This completes the proof.