CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Nathaniel Hobbs

Lecture #14
November 7, 2017

In this lecture we begin to cover a generalization of clustering, **Matrix completion**.
Let $M_{n\times n}$ be some matrix that is unknown to us.
Input: Matrix $P_\Omega(M)$, where $\Omega \subseteq [n] \times [n]$ and

$$P_\Omega(M)_{i,j} = \begin{cases} M_{i,j} & \text{if } i,j \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

Goal: Recover M

In this lecture we will show an algorithm that can approximately recover M, and outline an algorithm for fully recovering M, assuming some further constraints.

1 A Motivating Example (Netflix Problem)

Given a sparse matrix of user movie preferences, i.e. users who have rated movies, Netflix would like to complete the matrix so it can predict which movies to recommend to its users.

2 A gradual start

What if M is a random matrix?

If M is random, then we can't use any of the entries in $P_\Omega(M)$ to infer anything about the missing entries. Recovering M would then need $|\Omega| = O(n^2)$, which is trivial.

In practice, we are more interested in cases when M is low rank, i.e. $rank(M) \ll n$.

What if M has rank 1?

If M is a rank 1 matrix, then the SVD of M would be $M = \vec{\sigma}_1 \cdot \vec{u}_1 \cdot \vec{v}_1^\intercal = \vec{w}_1 \cdot \vec{v}_1^\intercal$, where $\vec{w}_1 = \vec{\sigma}_1 \cdot \vec{u}_1$. More explicitly,

$$M = \vec{w}_1 \cdot \vec{v}_1^\intercal = \begin{bmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1n} \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \end{bmatrix} = \begin{bmatrix} w_{11}v_{11} & w_{11}v_{12} & \dots & w_{11}v_{1n} \\ w_{12}v_{11} & w_{12}v_{12} & \dots & w_{12}v_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n}v_{11} & w_{1n}v_{12} & \dots & w_{1n}v_{1n} \end{bmatrix}$$

If there exists a row i where none of the entries have been revealed, then we can say nothing about the corresponding w_{1i} value. Similarly, if there exists a column j where none of the entries have been revealed, then we can say nothing about the corresponding v_{1j} value.

Given Ω, we can solve for $\vec{w}$ and $\vec{v}$ such that $\forall i,j \in \Omega$, $w_{1i}v_{1j} = M_{i,j}$. This is a non-linear system of equations with $2n-1$ variables (-1 because $\vec{v}$ has to be unit length). Therefore, we need to see at least $|\Omega| \geq 2n-1$ entries before we can say that there's a unique solution to the system of equations. It turns out this number of values is both necessary and sufficient for learning a rank 1 matrix.

What is M has rank r?

If M is a rank r matrix, then the SVD of M would be $M = \sum_{i=1}^{r} \vec{\sigma}_i \cdot \vec{u}_i \cdot \vec{v}_i^{\intercal} = \sum_{i=1}^{r} \vec{w}_i \cdot \vec{v}_i^{\intercal}$, where $\vec{w}_i = \vec{\sigma}_i \cdot \vec{u}_i$.

There are n coordinates per vector, but then there are dependencies because we want the $\vec{v}_i$s to be unit length, so the number of variables that would need to be exposed is $2nr - r$.

In general, solving such a system of non-linear equations is NP-Hard. However, if we allow flipping coins, the problem becomes tractable.

3 Matrix Completion via Random Sampling

In the random model, each entry is revealed with some probability $p \in (0, 1)$, i.e. $\forall i, j \in [n] \times [n], i, j \in \Omega$ with probability p. Then $\mathbb{E}[|\Omega|] = n^2 p$. Typically we want the expected value $\ll n \log n$, so $p \approx \frac{\log n}{n}$.

Is low rank and coin flipping enough?

Note that even with randomness and low rank M, exact recovery is not always possible. As an example, consider the following rank-1 matrix

$$M = e_1 e_1^{\intercal} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

In this case, the first entry **must** be seen, otherwise exact recovery is not possible. This is demonstrative of a general problem for exact recovery, namely when entries are clustered around some area in the matrix; unless a large proportion of the cluster is sampled, then the inference of M will be severely hindered.

3.1 Approximate Recovery

Approximate recovery is always possible via SVD. The algorithm works as follows:

Input: $P_\Omega(M) \leftarrow$ noisy version of M (note, this is high rank)

Step 1: $\frac{1}{p} P_\Omega(M) = \sum_{i=1}^{n} \vec{\sigma}_i \cdot \vec{u}_i \cdot \vec{v}_i$

Step 2: Output $\tilde{M} = \sum_{i=1}^{r} \vec{\sigma}_i \cdot \vec{u}_i \cdot \vec{v}_i$, i.e. top r components of scaled $P_\Omega(M)$

Theorem 1 (approximate recovery)**.** *If M has rank-r, $|M_{i,j}| \leq M_{max}$, and $p \geq c\frac{\log^4 n}{n}$, then the spectral method can output in polytime $\tilde{M}$ such that with probability $\geq 1 - \frac{1}{n^3}$*

$$\frac{1}{n}\|\tilde{M} - M\|_F \leq M_{max} O\Big(\sqrt{\frac{r}{np}}\Big)$$

Our proof will make use of the following theorem:

Theorem 2 (bounded spectral norm)**.** *If R is a random matrix with $\mathbb{E}[R_{i,j}] = 0, |R| \leq 1, \sigma^2 = \max_{i,j} Var(R_{i,j})$ then the spectral norm $\|R\| \leq 3\sigma\sqrt{n}$ w.h.p. provided $\sigma^2 \geq c\frac{\log^4 n}{n}$*

Proof. First, note that $\mathbb{E}[\frac{1}{p}P_\Omega(M)] = M$. Then let view $P_\Omega(M)$ as a noisy version of M, so the noise factor $R = M - \frac{1}{p}P_\Omega(M)$, and so $pR = pM - P_\Omega(M)$ and

$$(pR)_{i,j} = \begin{cases} (p-1)M_{i,j} & \text{with probabilty} p \\ pM_{i,j} & \text{with probabilty} 1-p \end{cases}$$

The variance is then $Var\big((pR)_{i,j}\big) = p(p-1)^2 M_{i,j}^2 + (1-p)p^2 M_{i,j}^2 \leq p(1-p)M_{max}^2$. Then from Theorem **??**, $\|pR\| \leq M_{max}\sqrt{np} \implies \|R\| \leq M_{max}\sqrt{\frac{n}{p}}$.

Now that we've bounded the spectral norm, we want to bound the Frobenius norm because it's usually much larger. In general, this is done by looking at the lower rank components. We know $\|\tilde{M} - M\|_2 \leq M_{max}\sqrt{\frac{n}{p}}$. Next, we have

$$\begin{aligned} \|\tilde{M} - M\|_F &\leq \sqrt{2r}\|\tilde{M} - M\|_2 && (1) \\ &\leq \sqrt{2r}\big(\|\tilde{M} - \frac{1}{p}P_\Omega(M)\| + \|\frac{1}{p}P_\Omega(M) - M\|\big) && (2) \\ &\leq 2\sqrt{2r}\|\frac{1}{p}P_\Omega(M) - M\|_2 && (3) \\ &\leq M_{max}O(\sqrt{\frac{rn}{p}}) && (4) \end{aligned}$$

Where (1) is because for any rank r matrix A, $\|A\|_F \leq \sqrt{r}\|A\|_2$, (2) uses triangle inequality to introduce $P_\Omega(M)$, and (3) is because both $\tilde{M}$ and M are rank r matrices and $\tilde{M}$ is the best rank r approximation and (4) follows by the spectral norm bound. □

In the next lecture we will present a method for exact matrix recovery.