Lecture #21: Matrix Completion

Netflix Prize (2006)

movies 17770

users 480,184

1-5 star rating

Can you improve their predictor's RMSE by 10%?

Eventually won by Bellkor's Pragmatic Chaos team in 2009, many lessons about aggregating predictors

e.g. Memento vs. Forrest Gump

Popular abstraction:

1. Unknown matrix $M$ (low-rank, incoherent)

2. Observe $m$ entries of $M$ chosen u.a.r.

Can we recover $M$ exactly or approximately?

Natural, non-convex approach

$$\min \text{ rank}(x) \text{ s.t. } x_{ij} = M_{ij} \quad \forall (i,j) \in \text{observed entries}$$

Given a hypothesis $X$ we can write its SVD as

$$X = U \Sigma V^T$$
where \( \Sigma = \text{diag}(\sigma^2) \), then \( \text{rank}(X) = \| \Sigma \|_0 \)

The same way as in compressed sensing relaxation

\[ \ell_0\text{-minimization} \Rightarrow \ell_1\text{-minimization} \]

we can write:

\[
\min \|X\|_4 \text{ s.t. } X_{ij} = M_{ij} \forall (i,j) \in \Omega
\]

where \( \|X\|_4 \) = sum of singular values of \( X = \| \Sigma \|_4 \)

nuclear norm

This is now a convex relaxation, but does it work?

some intuition:

**Lemma**: The unit nuclear norm ball - i.e.

\[ P = \{ X | \|X\|_\ast \leq 1 \} = \text{conv} \{ a b^\top | a^\top 1 = 1, \| b \|_1 \leq 1 \} \]

**Proof**: For any \( X \in P \) we have

\[
X = \sum_{i=1}^{r} \sigma_i u_i v_i^\top
\]

(coefficients in convex combination)

Conversely \( \| \cdot \|_\ast \) is a convex function so

\[
\left\| \sum c_i a_i b_i^\top \right\|_\ast \leq \sum \lambda_i \left\| a_i b_i^\top \right\|_\ast \leq 1
\]

today, we will take an empirical processes approach:

Suppose the true \( \mathbb{M} \) satisfies \( \| \mathbb{M} \|_\ast \leq R \) and

\[
\| \mathbb{M} \|_0 \leq M \frac{R}{\sqrt{n_1 n_2}}
\]
Here $M$ is some type of incoherence — how aligned are the singular vectors of $M$ with standard basis.

Let $X$ be the minimizer of $(P)$, we have

$$\|X\|_1 \leq \|M\|_1 \leq r$$

Some terminology:

- **Training error:** $\frac{1}{|L_1|} \sum_{ij \in L_1} |X_{ij} - M_{ij}| = 0$.

- **Test error:** $\frac{1}{n_1 n_2} \sum_{ij} |X_{ij} - M_{ij}|$.

**Q:** If the training error is small, how can we bound the test error?

This must depend on the nuclear norm, following [Srebro, Shraibman]

$$\|X\|_*$ \leq r \Rightarrow \text{"simple" explanation for the observations}$$

Generalization error = test error - training error

Consider the following hypothesis class:

$$\mathcal{H} = \left\{ X: [n_1 \times n_2] \to \mathbb{R} \mid \|X\|_* \leq r, \|X\|_{1,1} \leq \frac{m}{\sqrt{n_2}} \right\}$$

interpreting $X$ as a function from domain $[n_1 \times n_2]$ to a prediction.
We will bound the following:

\[ \sup_{x \in \mathbb{R}} \left| \frac{1}{m} \sum_{i=1}^{m} |x_{ij} - M_{ij}| - \frac{1}{n} \sum_{i=1}^{n} |x_{ij} - M_{ij}| \right| \]

[Kolmogorov, [Bartlett, Mendelson], ...]

**Theorem:** Consider a set of classifiers

\[ g : \mathbb{Z} \rightarrow \mathbb{R} \]

with some distribution \( D \) on \( \mathbb{Z} \) with probability \( \geq 1 - \varepsilon \)

\[ \sup_{g \in G} \left| \mathbb{E}[g(Y)] - \frac{1}{m} \sum_{i=1}^{m} g(Y_i) \right| \] (1)

For \( g \) random from \( D \)

\[ \leq 2R_m(G) + (b-a) \sqrt{\frac{\log 2/\varepsilon}{2m}} \]

where \( R_m(G) = \mathbb{E}\left[ \mathbb{E}\left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \epsilon_i g(Y_i) \right] \right] \)

Rademacher complexity

**Thought experiment:** If \( g : \mathbb{Z} \rightarrow \{ \pm 1 \} \) then

Rademacher complexity = how well does best hypothesis in \( G \) agree with random function on \( S \)

Intuitively: If we can match random function \( \Rightarrow \) hypothesis class too expressive to generalize
How would we map this to matrix completion?

\[ Z = \text{domain} = [n_1 \times n_2 \times J] \]

\[ g(i,j) = |X_{ij} - M_{ij}| = \text{loss on prediction for } (i,j) \]

**Proof:** (most interesting part)

Let \( \hat{E}_S [g] = \frac{1}{m} \sum_{i=1}^{m} g(Y_i) \) then

\[
E \left[ \sup_{S'} \left( E [g] - \hat{E}[g] \right) \right] = E \left[ \sup_{S' \in G} \left( E [\hat{E}[g]] - \hat{E}[g] \right) \right]
\]

where \( S' = y_1, \ldots, y_m \) are "ghost samples".

Convexity

\[
\leq E \left[ \sup_{S, S'} \sup_{g \in G} \left( E [\hat{E}[g]] - \hat{E}[g] \right) \right]
\]

\[
= E \left[ \sup_{S, S'} \frac{1}{m} \left( \sum_{i} g(Y_i') - g(Y_i) \right) \right]
\]

\[
= E \left[ \sup_{S, S'} \frac{1}{m} \left( \sum_{i} \epsilon_i (g(Y_i') - g(Y_i)) \right) \right]
\]

\[
\leq E \left[ \sup_{S, S'} \frac{1}{m} \sum_{i} \epsilon_i g(Y_i') + \sup_{g \in G} \sum_{i} \epsilon_i (0, g(Y_i)) \right]
\]
= 2 R_m(G)

This bounds expected generalization error, can prove deviation bounds using McDiarmid's inequality.

Can also prove other side of bound.

Aside: If $G$ were finite, can bound generalization error through union bound + Chernoff bound.

But $G$ is infinite in our setting (and whenever you use a convex relaxation to get a predictor).

Can also bound the generalization error using

\[(\#) \leq 2 R_5(G) + 3(b-a) \sqrt{\frac{\log\frac{4}{\delta}}{2m}}\]

where $R_5(G) = \mathbb{E} \left[ \sup_{\sigma} \frac{1}{m} \sum_{i=1}^{m} \sigma \cdot g(Y_i) \right]$

empirical rademacher complexity

This way we're not making any assumptions about the distribution, except that $S$ comes from it.

But how do we bound the rademacher complexity?

Let's consider a simpler problem:

\[ R_m(H) = \mathbb{E} \left[ \mathbb{E} \left[ \sup_{x \in \mathcal{X}} \frac{1}{m} \sum_{(i,j) \in A \cap \mathcal{X}} \sigma_{ij} x_{ij} \right] \right] \]
Let's drop the \( \| \cdot \|_{\infty} \) constraints.

\[
\mathbb{E} \left[ \mathbb{E} \left[ \sup_{\| x \|_2 \leq r} \mathbb{E} \left[ \sup_{\| x \|_2 \leq r} \sup_{\| a, b \| \leq 1} \langle x, a b^T \rangle \right] \right] \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ \sup_{\| a, b \| \leq 1} \langle x, a b^T \rangle \right] \right] \\
= \mathbb{E} \left[ \| x \|_2 \| x \|_2 \right] \\
= \mathbb{E} \left[ \| x \|_2 \right] \\
\text{m nonzero entries chosen a.a.r., random } \pm 1s.
\]

The important point is:

\[
\mathbb{E} \left[ \| x \|_2 \right] \geq \| x \|_F \| a b^T \|_F = \| x \|_F = \sqrt{m} \\
\Rightarrow \text{good generalization}
\]