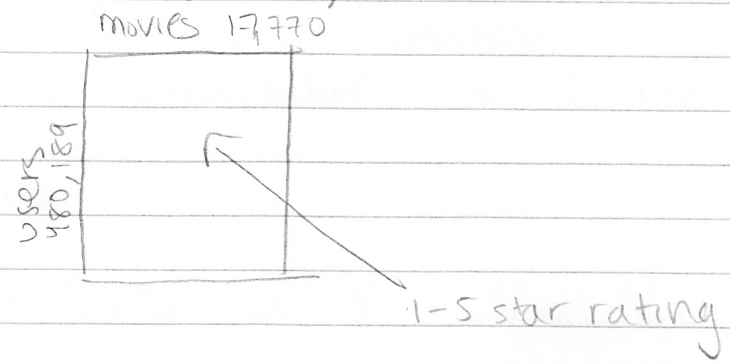
Lecture #21: Matrix Completion

Netflix Prize (2006)



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. Forest 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, nonconvex approach

observed entries

$$\min \ \mathrm{rank}(X) \quad \text{s.t.} \quad X_{ij} = M_{ij} \quad \forall (i,j) \in \Omega$$

Given a hypothesis X we can write its SVD as

$$X = U \Sigma V^T$$

where $\Sigma = \mathrm{diag}(\vec{\sigma})$. Then $\mathrm{rank}(X) = \|\vec{\sigma}\|_0$

The same way as in compressed sensing

$$\ell_0\text{-minimization} \xRightarrow{\text{relaxation}} \ell_1\text{-minimization}$$

we can write:

$$\min \|X\|_* \quad \text{s.t.} \quad X_{ij} = M_{ij} \quad \forall (i,j) \in \Omega$$

where $\underbrace{\|X\|_*}_{\text{nuclear norm}}$ = sum of singular values of $X = \|\vec{\sigma}\|_1$

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

some intuition:

<u>Lemma</u>: The unit nuclear norm ball – i.e.

$$P = \{X \mid \|X\|_* \leq 1\} = \mathrm{conv}\{ab^T \mid \|a\|, \|b\| \leq 1\}$$

<u>Proof</u>: For any $X \in P$ we have

$$X = \sum_{i=1}^{r} \underbrace{\sigma_i}_{\text{coefficients in convex combination}} u_i v_i^T$$

Conversely $\|\cdot\|_*$ is a convex function so

$$\left\|\sum_i \lambda_i a_i b_i^T\right\|_* \leq \sum \lambda_i \|a_i b_i^T\|_* \leq 1 \quad \blacksquare$$

Today, we will take an empirical processes approach:

Suppose the true $M \in \mathbb{R}^{n_1 \times n_2}$ satisfies $\|M\|_* \leq r$ and $\|M\|_\infty \leq M\sqrt{\frac{r}{n_1 n_2}}$

many choices in literature

Here μ 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

$$\|X\|_* \leq \|M\|_* \leq r$$

Some terminology:

by assumption

$$\text{training error}: \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} |X_{ij} - M_{ij}| = 0$$

$$\text{test error}: \frac{1}{n_1 n_2} \sum_{ij} |X_{ij} - M_{ij}|$$

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

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

$\|X\|_* \leq r \cong$ "simple" explanation for the observations

generalization error = test error − training error

Consider the following hypothesis class:

$$\mathcal{H} : \left\{ X : [n_1] \times [n_2] \to \mathbb{R} \;\middle|\; \|X\|_* \leq r, \|X\|_\infty \leq \mu \sqrt{\frac{r}{n_1 n_2}} \right\}$$

interpreting X as a function from domain $[n_1] \times [n_2]$ to a prediction

We will bound the following:

$$\sup_{X \in \mathcal{H}} \left| \frac{1}{n_1 n_2} \sum_{ij} |X_{ij} - M_{ij}| - \frac{1}{|\Omega|} \sum_{(ij) \in \Omega} |X_{ij} - M_{ij}| \right|$$

[Koltchinskii], [Bartlett, Mendelson], ...

Theorem: Consider a set of classifiers

$$G : Z \to [a,b]$$

with some distribution D on Z with probability $\geq 1-\delta$

$$\sup_{g \in G} \left| \mathbb{E}[\underset{\substack{\uparrow \\ \text{r.v from } D}}{g(Y)}] - \frac{1}{m} \sum_{i=1}^{m} g(Y_i) \right| \qquad (*)$$

$$\leq 2 R_m(G) + (b-a) \sqrt{\frac{\log 2/\delta}{2m}}$$

$$\text{where } \underset{\substack{\uparrow \\ \text{rademacher complexity}}}{R_m(G)} = \underset{\substack{S \\ \uparrow \\ \text{set of } m \text{ samples}}}{\mathbb{E}} \left[\underset{\sigma}{\mathbb{E}} \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \underset{\substack{\uparrow \\ \text{random } \pm 1 \text{ signs}}}{\sigma_i} g(Y_i) \right] \right]$$

Thought experiment: If $g : Z \to \{\pm 1\}$ 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?

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

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

<u>Proof:</u> (of most interesting part)

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

$$\mathbb{E}_S \left[\sup_{g \in G} \left(\mathbb{E}[g] - \hat{\mathbb{E}}[g] \right) \right] =$$

$$\mathbb{E}_S \left[\sup_{g \in G} \left(\mathbb{E}_{S'} [\hat{\mathbb{E}}_{S'}[g]] - \hat{\mathbb{E}}_S[g] \right) \right]$$

where $S' = Y_1' \ldots Y_m' \overset{iid}{\sim} D$ "ghost samples"

$$\underset{\text{of sup}}{\overset{\text{convexity}}{\leq}} \mathbb{E}_{S,S'} \left[\sup_{g \in G} \left(\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right) \right]$$

$$= \mathbb{E}_{S,S'} \left[\sup_{g \in G} \frac{1}{m} \left(\sum_i g(Y_i') - g(Y_i) \right) \right]$$

$$= \mathbb{E}_{\sigma,S,S'} \left[\sup_{g \in G} \frac{1}{m} \left(\sum_i \sigma_i \left(g(Y_i') - g(Y_i) \right) \right) \right]$$

$$\leq \mathbb{E}_{\sigma,S,S'} \left[\sup_{g \in G} \frac{1}{m} \sum_i \sigma_i g(Y_i') + \sup_{g \in G} \sum_i (-\sigma_i) g(Y_i) \right]$$

$= 2 R_m(G)$

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

Can also prove other side of bound. ∎

<u>Aside</u>: 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

$$(*) \leq 2 R_S(G) + 3(b-a)\sqrt{\frac{\log 4/\delta}{2m}}$$

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

<u>empirical</u> <u>rademacher</u> <u>complexity</u>

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:

$$R_m(\mathcal{H}) = \mathbb{E}_{\Omega}\left[\mathbb{E}_{\sigma}\left[\sup_{X \in \mathcal{H}} \frac{1}{m} \sum_{(i,j)\in\Omega} \sigma_{ij} X_{ij}\right]\right]$$

Let's drop the $\|\cdot\|_\infty$ constraints

$$\mathbb{E}_{\Omega}\left[\mathbb{E}_{\sigma}\left[\sup_{\|X\|_* \le r} \sum_{(i,j)\in\Omega} \sigma_{ij} X_{ij}\right]\right] \le r\, \mathbb{E}_{\Omega}\left[\underbrace{\mathbb{E}_{\sigma}\left[\sup_{\|X\|_* \le 1} \sum_{(i,j)\in\Omega} \sigma_{ij} X_{ij}\right]}\right]$$

$$= \mathbb{E}_{\Omega}\left[\mathbb{E}_{\sigma}\left[\sup_{\|a\|,\|b\|\le 1} \sum_{(i,j)\in\Omega} \sigma_{ij} a_i b_j\right]\right]$$

$$= \mathbb{E}_{\Omega}\left[\mathbb{E}_{\sigma}\left[\sup_{\|a\|,\|b\|\le 1} \langle \Sigma, ab^T\rangle\right]\right]$$

$$= \mathbb{E}\left[\|\Sigma\|\right]$$

m nonzero entries chosen u.a.r, random ± 1s

The important point is:

$$\mathbb{E}\left[\|\Sigma\|\right] \overset{\text{matrix concentration}}{\ll} \|\Sigma\|_F \, \|ab^T\|_F = \|\Sigma\|_F = \sqrt{m}$$

trivial bound

$\Rightarrow$ good generalization