Collaborative filtering :

Let $M \in R^{m \times n}$,

M = data matrix, m = rows, n = Columns.

rating = M_{ij} for $i \in [m]$ and $j \in [n]$

[i-th user given rating M_{ij} for j-th item]

all observable entries in $M = \Omega$:

$\Omega = \{(i, j) ; M_{ij}$ is observed$\}$

$|\Omega| \ll mn$ [all ratings are not received]

Observable entries :

$\Omega_i \subseteq [n]$ and $\Omega_j' \subseteq [m]$

Goal of collaborative filtering :

[recover complete matrix M]

→ Impossible to complete arbitrary M.

Assumption:

- M is very close to $m \times n$ rank-K matrix : $K < \min(m, n)$.
- Then complete matrix can be recovered by solving the optimization:

$$\min_{X \in R^{m \times n}} \| R_{\Omega}(M - X) \|_F^2, \text{ s.t. } \text{rank}(X) \leq K \quad (1)$$

$$\left[\| A \|_F^2 = \sum_{i,j} A_{ij}^2 \left[\text{squared Frobenious norm of matrix A} \right] \right]$$

$$R_{\Omega}(M - X) \quad [R_{\Omega}(A)]_{ij} = A_{ij} \text{ if } (i,j) \in \Omega$$
$$= 0, \text{ otherwise}$$

Eqn: (1) is non-convex, hence optimization is difficult.

Two standard approximation:

i) alternating minimization
ii) nuclear-norm minimization.

alternating minimization:

$$\min_{\substack{U \in R^{m \times k} \\ V \in R^{n \times k}}} \left\{ \| R_\Omega (M - UV^T) \|_F^2 + 2\lambda_U \|U\|_F^2 + 2\lambda_V \|V\|_F^2 \right.$$

[λ_U, λ_V → small positive constants to improve convergence]

Nuclear norm minimization:

$$\min_{X \in R^{m \times n}} \| R_\Omega (M - X) \|_F^2 + 2\lambda \|X\|_*$$

→ $\lambda > 0$, regularization parameter

$$\|X\|_* = \sum_{i=1}^{\text{rank}(X)} \sigma_i(X)$$

↑ nuclear norm of X (Convex surrogate of rank function)

σ_i = singular values.

Attack Model :

↳ How data poisoning can be done?

- for m user n item scenario, attacker is adding αm malicious users to training data matrix.
- Each malicious user is allowed to report preference on at most B items each bounded in a range $[-\lambda, \lambda]$.

Consider, $M \in R^{m \times n}$ [original matrix]

$\tilde{M} \in R^{m' \times n}$ [where $m' = \alpha m$ malicious users]

$\tilde{\Omega}$ = set of non-zero entries in $\tilde{M}$.

m' = no. of malicious users.

$|\tilde{\Omega}_i| < B$ for $i \in (1, \ldots m')$

$\|\tilde{M}\|_{max} = \max |\tilde{M}_{ij}| \leq \lambda.$

Attacker's Goal :

→ Attacker wants to divert the collaborative filtering algorithm to predict $\hat{M}$ on original dataset M for $\tilde{M}$.

→ $R.(\hat{M}, M)$ = attacker's utility that takes M and outputs $\hat{M}$.

To be successful optimal solution should be computed jointly :

$$\Theta_\lambda(\tilde{M}; M) = \arg\min \|R_\Omega(M - X)\|_F^2 + R_{\tilde{\Omega}}(\tilde{M} - \hat{X})\|_F^2 + 2\lambda\|(X; \hat{X})\|_*$$

[for nuclear norm minimization.]

→ similarly, objective function for alternating minimization can be updated with $\tilde{M}$, $\tilde{U}$ and $\tilde{\Omega}$.

Types of attacks:

1) Availability attack: attacker wants to maximize the error of the filtering system → system becomes useless.

$\bar{M}$ = prediction without error.

$$R^{av}(\hat{M}, M) = \| R_{\Omega^c}(\hat{M} - \bar{M}) \|_F^2$$

[Ω^c = unknown entries]

[utility function = total number of perturbations.]

2) Integrity Attack: Attacker wants to boost the popularity of a (subset) of items.

Let $J_0 \subseteq [n]$ ⇒ subset of items

$w : J_0 \to R$, specified weight vector by attacker

$$R_{J_0, w}(\hat{M}, M) = \sum_{i=1}^{m} \sum_{j \in J_0} w(j) \hat{M}_{ij}$$

Hybrid attack :

$$R^{hybrid}_{J_0,\omega,\mu}(\hat{M}, M) = \mu_1 R^{av}_{J_0,\omega}(\hat{M}, M) + \mu_2 R^{In}(\hat{M}, M)$$

[$\mu, \mu_2 \Rightarrow$ coefficients that trade off between availability and integrity]

Attack strategy

$$\tilde{M}^{(t+1)} = Proj_M\left(\hat{M}^t + s_t \cdot \nabla_{\hat{M}} R(\tilde{M}, M)\right)$$

→ Rate of change of utility R with perturbation.

→ use that in projected gradient ascent (PGA)

→ $Proj_M(\cdot)$ operator onto the feasible region M.