دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

دستورکار آزمایشگاه هوش محاسباتی

جلسه ۱۲
عامل‌بندی ماتریس و سیستمِ پیشنهاددهنده

استاد درس: دکتر مهران صفایانی

فصل ۱۲

عامل‌بندی ماتریس و سیستمِ پیشنهاددهنده

اهداف این جلسه

شما در این جلسه یاد خواهید گرفت که :

- یک سیستمِ پیشنهاددهنده [۱] بسازید.
- عملکرد این سیستم را ارزیابی کنید.
- عامل‌بندیِ ماتریس [۲] را با استفاده از SGD [۳] درک و پیاده‌سازی کنید.
- تعداد مناسبی از عامل ها، مانند عامل‌های منظم‌سازی، را انتخاب کنید.
- سیستم خود را با یکسری پایه [۴] مقایسه کنید.

در این جلسه، ما از از مجموعه دادهِ `movielens100k.csv` استفاده خواهیم کرد. همچنین شما می‌توانید از کدهای آماده‌ی مفیدی که برای شما آماده شده است، استفاده کنید.

---

[۱] System Recommender
[۲] Factorization Matrix
[۳] Descent Gradient Stochastic
[۴] baseline

۱.۱۲ ا یجاد و مصورسازیِ تقسیم‌بندیِ داده‌های آموزشی و آزمون

از آنجایی که هدف ما، پیشبینی کردنِ رتبه‌بندی های دیده نشده است، می‌توانیم یک مجموعه‌ی آزمون را با پنهان کردنِ برخی عناصر ماتریس، ماتریس ایجاد کنیم. منظور ما، انتخاب کردنِ تعدای رتبه‌بندی به صورت تصادفی از توی ماتریس است. این کار یک عمل بسیار معمولی در مجموعه داده‌های سیستم‌های پیشنهاد دهنده و کلان داده ها است.

تمرین اول

تابعِ `split_data` که درون فایل ژوپیترِ مربوط به این جلسه قرار دارد را به منظور تقسیم مجموعه داده به داده‌های آموزشی و آزمون، تکمیل کنید. ما فقط فیلمها و کاربرانی را در نظر می‌گیریم، که بیش از ۱۰ رتبه‌بندی داشته باشند و دیگر فیلم‌ها و کاربران را نادیده می‌گیریم. در میان رتبه‌بندی‌هایِ معتبرِ باقی‌مانده، شما باید بصورت تصادفی با احتمالِ ۱۰٪ یک رتبه‌بندی را برای داده‌ی آزمون و با احتمال ۹۰٪ یک رتبه‌بندی را برای داده‌ی آموزشی، انتخاب کنید. ما تعداد کمی از داده‌های آموزشی‌ای را برای کاربران و فیلم‌هایی که رتبه‌بندی کمی دارند، نگه می‌داریم.
می‌توان نتیجه‌ی تقسیم بندی را مانند شکل دو، نشان داد.

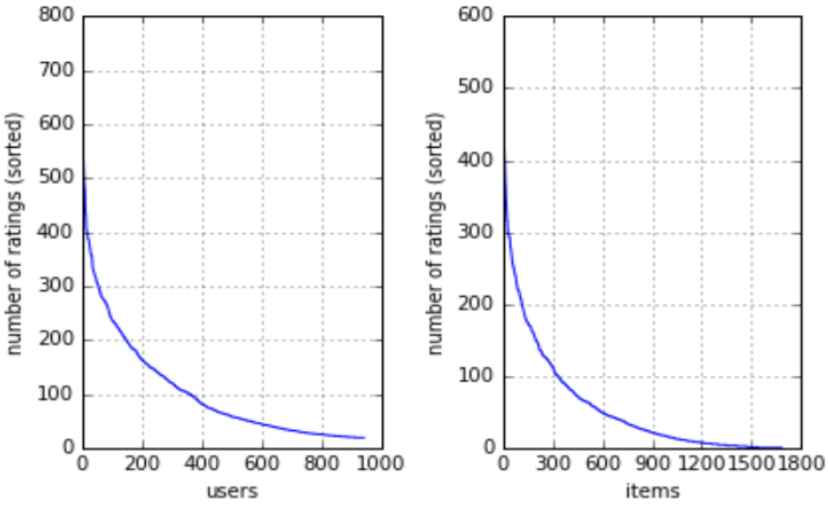
۲.۱۲ ارزیابی عملکرد

ما از مجذورِ میانگینِ مربعاتِ خطا [۱] (RMSE) برای ارزیابی عملکرد استفاده خواهیم کرد. با داشتنِ رتبه‌بندی صحیحِ x_{dn} برای d ـاُمین فیلم و n ـاُمین کاربر و همچنین پیش‌بینیِ $\hat{x}_{dn} = \boldsymbol{W}_{d:}\boldsymbol{Z}_{:n} = \left(\boldsymbol{W}\boldsymbol{Z}^\top\right)_{dn}$ ، می‌توانیم RMSE را بصورت زیر تعریف کنیم:

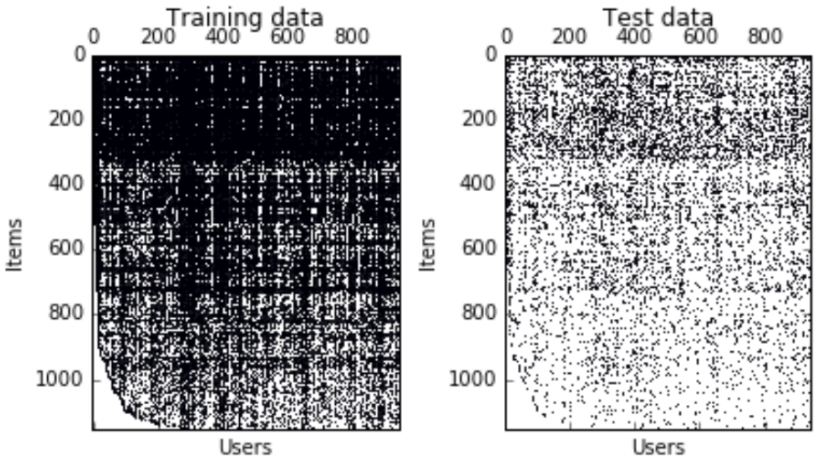
$$\mathrm{RMSE}(\boldsymbol{W}, \boldsymbol{Z}) := \sqrt{\frac{1}{|\Omega|}\sum_{(d,n)\in\Omega}\frac{1}{2}\left[x_{dn} - \left(\boldsymbol{W}\boldsymbol{Z}^\top\right)_{dn}\right]^2} \tag{۱.۱۲}$$

که $\boldsymbol{W} \in \mathbb{R}^{D\times K}, \boldsymbol{Z} \in \mathbb{R}^{N\times K}$. در اینجا $\Omega \subseteq [D] \times [N]$ مجموعه‌ی شماره‌های رتبه‌بندی های مشاهده شده‌ی ماتریسِ ورودیِ $\boldsymbol{X}$ می‌باشد. RMSE می‌تواند هم بر روی مجموعه آموزشیِ Ω محاسبه شود و هم بر روی یک مجموعه آزمونِ آماده به کار.

[۱] root mean-square error



شکل ۱.۱۲: شکل سمت چپ، تعداد رتبه‌بندیِ (مرتب شده) هر کاربر را نشان می‌دهد و شکل سمت راست، تعداد رتبه‌بندی (مرتب شده) فیلم‌ها را نشان می‌دهد. در هر دو نمودار، اعداد به منظور وضوح بیشتر، به صورت نزولی نمایش داده شده اند.



شکل ۲.۱۲: این تصویر، تقسیم‌بندیِ آموزشی و آزمونِ داده ها را نشان می‌دهد. نمودار سمت چپ، نشان‌دهنده‌ی داده های آموزشی و نمودار سمتِ راست، نشان‌دهنده‌ی داده های آزمون، می‌باشند. در هر نمودار، هر نقطه بیانگر یک زوجِ کاربر–فیلم با یک رتبه‌بندیِ غیر صفر، می‌باشد.

۳.۱۲ مدل‌های پایه

از مدل های ذیل، که برای پیش‌بینی، از میانگین بهره می‌برند، به عنوان پایه، استفاده خواهیم کرد.

$$\text{Mean: Global } \hat{x} := \frac{1}{|\Omega|}\sum_{(d,n)\in\Omega} x_{dn} = \frac{1}{|\Omega|}\sum_{n=1}^{N}\sum_{d\in\Omega_{:n}} x_{dn} = \frac{1}{|\Omega|}\sum_{d=1}^{D}\sum_{n\in\Omega_{d:}} x_{dn}, \tag{۲.۱۲}$$

$$\text{Mean: User } \hat{x}_n := \frac{1}{|\Omega_{:n}|}\sum_{d\in\Omega_{:n}} x_{dn}\ ، \tag{۳.۱۲}$$

$$\text{Mean: Movie } \hat{x}_d := \frac{1}{|\Omega_{d:}|}\sum_{n\in\Omega_{d:}} x_{dn}, \tag{۴.۱۲}$$

که Ω مجموعه‌ای از نشان‌گرهای غیرصفرِ رتبه‌بندی (d,n) در ماتریس داده‌های آموزشی و $\Omega_{:n}$ مجموعه‌ای از فیلم‌های رتبه‌بندی‌شده توسط n ـاُمین کاربر و $\Omega_{d:}$ مجموعه‌ی کاربرانی است که به فیلمِ d ـام، رتبه داده اند.

تمرین دوم

در ابتدا، سه پایه‌ی بالا را با هم مقایسه خواهیم کرد.

- به نظر شما، کدامیک از سه مدل، بهترین عملکرد را خواهد داشت ؟ چرا؟
- توابعِ `baseline_global_mean()` ، `baseline_user_mean()` و `baseline_item_mean()` را پیاده سازی کنید.
- مدل‌های حاصله را با یکدیگر مقایسه کنید. کدام مدل، کمترین RMSE آموزشی را به ما می‌دهد؟ کدامیک کمترین RMSE آزمون را به ما می‌دهد؟
  راهنمایی: شما می‌توانید مقداردهیِ تصادفیِ تقسیمِ آموزشی یا آزمون خود را تغییر دهید و در نتیجه، تعداد زیادی تخمین، ایجاد کنید.

۴.۱۲ عامل بندی ماتریس با استفاده از SGD

تمرین سوم

هدف پیادسازیِ نزول گرادیانی تصادفی است

- نزول گرادیانی نزولی را برای تابع هدفِ ،RMSE همانطورکه در معادله‌ی (۱) نشان داده شده است، پیاده سازی کنید.
- SGD را برای تابع هزینه‌ی منظم‌شده ی

$$\mathcal{L}(\boldsymbol{W},\boldsymbol{Z}) := \frac{1}{2}\sum_{(d,n)\in\Omega}\left[x_{dn} - \left(\boldsymbol{W}\boldsymbol{Z}^\top\right)_{dn}\right]^2 + \frac{\lambda_w}{2}\|\boldsymbol{W}\|_{\text{Frob}}^2 + \frac{\lambda_z}{2}\|\boldsymbol{Z}\|_{\text{Frob}}^2$$

پیاده‌سازی کنید. که در آن $\lambda_w, \lambda_z > 0$ اسکالر هستند.

- بصورت تجربی، بهترین اندازه گامِ γ را به منظور گرفتنِ کمترین خطای آموزشی، پیدا کنید.
- آیا با این کار، خطای آزمون نیز بصورت یکنواخت کاهش می‌یابد یا دوباره بعد از مدت زمانی، افزایش می‌یابد؟ در این مورد، انتخاب های دیگرِ K را نیز امتحان کنید.