بسمه تعالی

الگوشناسی آماری ۷۲۵-۴۰ (گروه ۲) **نیمسال دوم ۹۲-۹۱**

پروژه پایانی: سیستم‌های ارائه پیشنهاد (Recommender Systems) **مدرس: سلیمانی**

توضیحات :

- تمامی کدها باید به زبان Matlab نوشته شوند.
- برای هر فاز باید گزارش کامل به همراه کد ضمیمه گردد.
- کد شما برای نمره‌دهی مطالعه خواهد شد. بنابراین باید توضیحات کامل در مورد شیوه پیاده‌سازی در گزارش آورده شود یا کد دارای توضیح‌گذاری مناسب باشد.
- کد و گزارش باید در یک فایل ZIP با نام SPR_P#N_#SN (به جای #N شماره فاز و به جای #SN شماره دانشجویی جایگزین نمایید.) به آدرس spr.ce.sut@gmail.com ارسال گردد.

سیستم ارائه پیشنهاد[1]: رویکرد آماری

بیان مسئله:

فرض کنید تعدادی آیتم (مثل فیلم، کتاب، ...) در یک بازار خرید آنلاین موجود است. هدف ما طراحی سیستمی خودکار برای پیشنهاد دادن آیتم‌های مناسب به هر کاربر است. کاربر در سیستم خرید آنلاین به تعدادی از کالاها رتبه داده است. در رویکرد فیلتر اشتراکی[2] برای ارائه آیتم‌های مناسب به یک کاربر می‌توانیم از رتبه‌بندی بقیه‌ی کاربرها و رتبه‌هایی که خود کاربر به کالاهای مختلف داده است، استفاده کنیم تا میزان رای کاربر به آیتم‌های مختلف را پیش‌بینی نموده و در نتیجه پیشنهادهای موثرتری به کاربر داشته باشیم. تاکنون روش‌های متعددی برای حل مساله‌ی پیش‌بینی میزان رای یک شخص به کالاها ارائه شده است.

در فاز اول این پروژه ابتدا از روش‌های ساده مبتنی بر همسایگی برای سیستم ارائه‌ی پیشنهاد کار را شروع کرده و سپس با معرفی فضای نهان[3] سعی در پیدا کردن فضایی داریم که کاربرها و آیتم‌ها می‌توانند در این فضای مشترک واقع شوند و با استفاده از بردارهای کاربر و آیتم مقدار رای پیش‌بینی شود. سپس در فاز دوم یک روش مبتنی بر مدل آماری را مورد مطالعه قرار داده و عمل‌کرد آن را بر روی مساله‌ی مربوطه بررسی می‌نماییم.

مجموعه داده:

مجموعه داده‌ی موردنظر برای این پروژه، پایگاه داده MovieLens[4] است که شامل رای هزار کاربر به هزار و هفتصد فیلم می‌باشد. در این پایگاه داده هر رای یک عدد صحیح در بازه [5-1] است. علاوه بر میزان رای کاربر به آیتم، زمان ثبت رای نیز موجود است.

---

[1] recommender system
[2] collaborative filter
[3] latent space
[4] http://www.grouplens.org/node/12

فاز اول پروژه:

زمان تحویل: ۱۳ خرداد

نمره : ۵۵ درصد

هدف : بررسی رویکرد تجزیه ماتریس رای‌ها و مقایسه آن با روش ساده مبتنی بر همسایگی

ابتدا قصد داریم الگوریتم k نزدیکترین همسایه را به عنوان یک روش ساده‌ی مبتنی بر حافظه[5] برای مساله‌ی پیش‌بینی میزان رای بررسی کنیم. برای این منظور پیاده‌سازی را در دو حالت زیر انجام دهید:

- مبتنی بر آیتم[6]: برای تخمین رای یک کاربر به یک آیتم، از رای کاربر به آیتم‌های مشابه استفاده شود. همچنین برای تعیین میزان شباهت آیتم‌ها معیارهای ضریب همبستگی Pearson و شباهت کسینوسی تنظیم شده[7] مورد بررسی قرار گیرد.
- مبتنی بر کاربر[8]: برای تخمین رای یک کاربر به یک آیتم، از رای کاربران مشابه به آن آیتم استفاده شود. مشابه حالت قبل برای تعیین میزان شباهت کاربرها معیارهای ضریب همبستگی Pearson و شباهت کسینوسی تنظیم شده مورد بررسی قرار گیرد.

(برای آشنایی بیشتر با معیارهای شباهت بالا می‌توانید مرجع [۱] را مطالعه نمایید)

یک ایده‌ی جدیدتر و کاراتر این است که به جای محاسبه‌ی شباهت‌های آماری (نظیر ضریب همبستگی) بین کاربرها یا بین آیتم‌ها به صورتی که در بالا ذکر شد، از طریق ماتریس رای‌ها، فضای فاکتورهای نهان (معمولاً با ابعاد کم) را پیدا کنیم. کاربران و آیتم‌ها در این فضا واقع می‌شوند و رای کاربر به آیتم می‌تواند از طریق ضرب داخلی بردار ویژگی کاربر و آیتم (در این فضا) تاحد خوبی تخمین زده شود. در واقع فاکتورهای نهان (یا پایه‌های فضای نهان) در این مساله می‌توانند مفاهیمی نظیر ژانر فیلم، موضوع فیلم یا ... باشند. به عبارت دیگر اگر ماتریس $\boldsymbol{A}$ شامل رای n کاربر به m آیتم را در نظر بگیریم (a_{ij} رای کاربر i به آیتم j را نشان می‌دهد)، قصد داریم ماتریس‌های $\boldsymbol{W}$ و $\boldsymbol{H}$ را که به ترتیب بردارهای ویژگی کاربران و آیتم‌ها را شامل می‌شوند به‌نحوی پیدا کنیم که مقدار تابع هزینه‌ی زیر حداقل شود:

$$J(\boldsymbol{W},\boldsymbol{H}) = \sum_{i=1}^{n}\sum_{j=1}^{m} loss\left(\boldsymbol{w}_i^T \boldsymbol{h}_j, a_{ij}\right) \qquad (۱)$$

تابع $loss$ در عبارت بالا می‌تواند تابعی نظیر مجذور خطا باشد.

SVD[9] یک روش شناخته شده برای تجزیه‌ی ماتریس به فاکتورها است. در این روش برای ماتریس $\boldsymbol{A}$ با ابعاد $m \times n$، ماتریس‌های unitary[10] $\boldsymbol{U}$ و $\boldsymbol{V}$ به ترتیب با ابعاد $n \times r$ و $r \times m$ و همچنین ماتریس قطری $\boldsymbol{S}$ با ابعاد $r \times r$ به‌گونه‌ای پیدا می‌شوند که $\boldsymbol{A} = \boldsymbol{USV}^T$. عناصر روی قطر ماتریس $\boldsymbol{S}$ یا همان مقادیر تکینه، جذر مقادیر ویژه‌ی ناصفر ماتریس $\boldsymbol{AA}^T$ هستند و به همین ترتیب

---

[5] Memory-based
[6] Item-based
[7] Adjusted cosine similarity
[8] User-based
[9] Singular Value Decomposition (SVD)
[10] Unitary: ماتریس‌هایی که ستون‌هایشان مجموعه‌ای از بردارهای orthonormal است

ستون‌های ماتریس $\boldsymbol{U}$ را بردارهای ویژه‌ی ماتریس $\boldsymbol{AA^T}$ (متناظر با مقادیر ویژه‌ی ناصفر) و ستون‌های ماتریس $\boldsymbol{V}$ را بردارهای ویژه‌ی ماتریس $\boldsymbol{A^TA}$ (متناظر با مقادیر ویژه‌ی ناصفر) تشکیل می‌دهند.

می‌توان نشان داد اگر از بین مقادیر ویژه $\boldsymbol{AA^T}$ تنها k مقدار بزرگتر در ماتریس $\boldsymbol{S_k}$ لحاظ شوند و به طور متناظر ماتریس‌های $\boldsymbol{U_k}$ و $\boldsymbol{V_k}$ با استفاده از بردارهای ویژه‌ی متناظر با k مقادیر ویژه‌ی بزرگتر پیدا شوند، ماتریس $\boldsymbol{A' = U_kS_kV_k^T}$ از بین ماتریس‌های با رتبه k کمترین فاصله (نرم Frobenious) را با ماتریس $\boldsymbol{A}$ خواهد داشت ($\boldsymbol{A'}$ جواب مساله‌ی $\min_{\boldsymbol{B},rank(\boldsymbol{B})=k}\|\boldsymbol{A}-\boldsymbol{B}\|_F$ است). به این حالت تجزیه truncated-SVD گفته می‌شود ($k < r$).

نشان دهید در حالتی‌که تابع *loss* مجذور خطا باشد، ماتریس‌های $\boldsymbol{W}$ و $\boldsymbol{H}$ در رابطه‌ی (۱) می‌توانند به راحتی به صورت $\boldsymbol{W = U_kS_k^{1/2}}$ و $\boldsymbol{H = V_kS_k^{1/2}}$ محاسبه شوند. به این ترتیب می‌توان به‌طور همزمان کاربران و آیتم‌ها را در یک فضای نهان k-بعدی نمایش داد.

یک مشکل در این روش آن است که ماتریس اولیه‌ای که قرار است تجزیه شود باید همه‌ی درایه‌هایش معلوم باشند، در حالی‌که در بیش‌تر درایه‌های ماتریس رای $\boldsymbol{A}$ نامشخص هستند. یک روش ابتدایی برای حل این مشکل آن است که مقادیر نامشخص در ماتریس، اول با صفر پر شوند و سپس SVD اعمال شود. ابتدا این روش ساده را پیاده‌سازی کنید (پارامتر k مناسب باید پیدا شود).

سپس تابع هدف مناسب‌تر زیر را در نظر بگیرید:

$$J(\boldsymbol{W},\boldsymbol{H}) = \sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij} \times loss\left(\boldsymbol{w}_i^T\boldsymbol{h}_j, a_{ij}\right) \qquad (۲)$$

که مقادیر s_{ij} در صورتی‌که کاربر i به آیتم j رای داده باشد مقدار ۱ وگرنه مقدار صفر دارد. اما بهینه‌سازی در این حالت حتی اگر تابع *loss* مجذور خطا درنظر گرفته شود نمی‌تواند مشابه حالت قبل به راحتی انجام شود.

نشان دهید که در حالتی‌که تابع *loss* مجذور خطا باشد، برای کمینه کردن تابع هزینه (۲) با ثابت نگه‌داشتن هر یک از ماتریس‌های $\boldsymbol{H}$ و $\boldsymbol{W}$ می‌توان رابطه‌ی مربوط به ماتریس دیگر را پیدا کرد. سپس روش بهینه‌سازی را پیاده‌سازی نمایید که با استفاده از یک رویکرد تکراری (تا رسیدن به همگرایی) در هر دور ابتدا $\boldsymbol{W}$ را ثابت نگه دارد و $\boldsymbol{H}$ را بهینه کند و سپس بالعکس ($\boldsymbol{H}$ را ثابت نگه داشته و $\boldsymbol{W}$ را پیدا کند)، تا در نهایت ماتریس‌های $\boldsymbol{W}$ و $\boldsymbol{H}$ حاصل را پیدا کند.

در قسمت بعد برای جلوگیری از بیش‌برازش[11] به تابع هزینه‌ی موجود در رابطه‌ی (۲) جملات منظم‌سازی[12] را به صورت زیر اضافه نمایید:

$$J(\boldsymbol{W},\boldsymbol{H}) = \sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij} \times loss\left(\boldsymbol{w}_i^T\boldsymbol{h}_j, a_{ij}\right) + \lambda_1\|\boldsymbol{W}\|^2 + \lambda_2\|\boldsymbol{H}\|^2 \qquad (۳)$$

و روابط به‌روزرسانی مربوط در هر دور تکرار الگوریتم را مشابه بالا برای حالتی‌که تابع *loss* مجذور خطا باشد، پیدا کنید. سپس این الگوریتم را نیز پیاده‌سازی نمایید.

[11] ovefitting
[12] regularization

معیار و نحوه‌ی ارزیابی

- معیار ارزیابی RMSE[۱۳] در نظر گرفته شود.
- برای انتخاب پارامترهای هریک از روش‌ها از 10-fold cross validation استفاده نمایید.
- RMSE روی مجموعه آزمون را به ازای روش‌های حاصل (با پارامترهای انتخاب شده) مشخص نمایید و نتایج به دست آمده برای روش‌ها را مقایسه و تحلیل نمایید.
- در بخش نتایج ارائه نمودارهایی که متناظر با انتخاب پارامترها با استفاده از cross-validation، تغییرات عملکرد را به ازای مقادیر مختلف پارامترها نشان دهند، لازم است.

برای این‌که در گزارش ارجاع به روش‌هایی که پیاده‌سازی کردید، یکنواخت باشد، نام‌گذاری روش‌های پیاده‌سازی شده به ترتیب معرفی در متن بالا به صورت زیر باشد: kNN-Item، kNN-User، Matrix-Fact-Zero، Matrix-Fact-AO، Matrix-Fact-AO-Regul

فاز دوم : پیاده سازی یک روش آماری مبتنی بر مدل

زمان تحویل: ۱۱ تیر

نمره : ۴۵ درصد

هدف : به‌کارگیری یک مدل احتمالی

در این فاز قصد داریم یک مدل احتمالی برای پیدا کردن فضای نهان (که در فاز قبل در مورد آن بحث شد) استفاده نماییم. یکی از روش‌های مطرح برای سیستم‌های ارائه‌ی پیشنهاد، مدل کردن رای‌دهی به صورت احتمالی است. از مدل‌های پایه‌ای که برای این منظور ارائه شده می‌توان مدل تحلیل معنایی نهان احتمالی[۱۴] را نام برد [۲]. فرض اصلی این مدل برقراری ارتباط کاربر و آیتم با استفاده از متغیر واسطی است که نهان می‌باشد (در ادامه حالت forced prediction معرفی شده در مقاله [۲] مورد بررسی قرار می‌گیرد).

فرض کنید کاربر، آیتم و رای را به ترتیب با u، i و v نشان دهیم. مقدار لگاریتم درست‌نمایی[۱۵] روی درایه‌هایی از ماتریس $\boldsymbol{A}$ که موجود هستند، به ازای پارامترهای $\boldsymbol{\theta}$ به صورت زیر محاسبه می‌شود:

$$\sum_{\substack{s_{ui}=1 \\ i,u}} \log p(v = a_{ui} | u, i; \boldsymbol{\theta}) \qquad (۴)$$

ماتریس $\boldsymbol{S}$ به ازای زوج کاربر و آیتم‌هایی که رای در ماتریس $\boldsymbol{A}$ وجود دارد یک و به ازای بقیه صفر است. a_{ui} درایه‌ی از ماتریس $\boldsymbol{A}$ است که رای کاربر u به آیتم i را نشان می‌دهد.

در روش مورد بررسی، احتمال رای کاربر به آیتم با استفاده از مدل ترکیبی گاوسی[۱۶] مدل می‌شود (z متغیر نهان):

$$p(v|u,i) = \sum\nolimits_z p(v|i,z) P(z|u) \qquad (۵)$$

که چگالی احتمال شرطی $p(v|i,z)$ گاوسی فرض شده است $p(v|i,z) \sim N(v|\mu_{i,z}, \sigma_{i,z}^2)$.

---

[13] Root Mean Square Error
[14] probabilistic Latent Semantic Analysis
[15] likelihood
[16] Gaussian mixture models

در واقع در این مدل، میزان رایی که کاربر به آیتم می‌دهد وابسته به یک عامل نهان z است. برای مثال اگر بدانیم کاربر فیلم‌های نوع ترسناک را دوست دارد و فیلم i ترسناک است، می‌توانیم مستقل از خود شخص کاربر میزان رای کاربر به این فیلم را مشخص کنیم. به این ترتیب عامل نهان ضمن کشف توزیع رای به آیتم‌ها می‌تواند میزان رای به آیتم را از کاربر مستقل نماید.

برای بیشینه کردن لگاریتم درست‌نمایی از الگوریتم برآورد–بیشینه‌سازی[۱۷] (EM) استفاده می‌شود. در این الگوریتم تارسیدن به همگرایی در هر دور گام‌های برآورد (E-step) و بیشینه‌سازی (M-step) تکرار می‌شوند. در گام E، احتمال پسین[۱۸] z به صورت زیر محاسبه می‌شود:

$$P(z|u,v,i;\boldsymbol{\theta}^{old}) = \frac{p(v|i,z;\boldsymbol{\theta}^{old})P(z|u;\boldsymbol{\theta}^{old})}{\sum_{z'} p(v|i,z';\boldsymbol{\theta}^{old})P(z'|u;\boldsymbol{\theta}^{old})} \quad (۶)$$

و سپس در گام M پارامترهای $\boldsymbol{\theta}$ به‌گونه‌ای یافت می‌شوند که تابع زیر حداکثر شود:

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{old}) = \sum_{u,i} P(z|u,v,i;\boldsymbol{\theta}^{old})[\log p(v|i,z;\boldsymbol{\theta}) + \log P(z|u;\boldsymbol{\theta})] \quad (۷)$$

پس از یادگیری پارامترها از طریق الگوریتم EM برآورد رای کاربر u به آیتم i می‌تواند به صورت $E[v|u,i] = \sum_z P(z|u)\mu_{i,z}$ انجام شود.

در این فاز هدف آن است که ابتدا با پاسخ‌گویی به سوالات نظری (عمدتاً از طریق مطالعه‌ی مرجع [۲]) به مفاهیم پایه‌ی مورد نظر برای استفاده از این رویکرد آماری مسلط شوید و سپس بخش عملی مربوطه را پیاده سازی نمایید و مساله‌ی تخمین مقدار رای را به صورت احتمالی حل نمایید.

سوالات نظری:

- مفهوم متغیر نهان در روش مطرح در مرجع [۲] چیست؟ مشخص کنید که در این روش استفاده از متغیر نهان منجر به خوشه‌بندی کاربرها می‌شود یا آیتم‌ها؟ توضیح دهید.
- اگر به جای رابطه‌ی (۵) به صورت $p(v|u,i) = \sum_z p(v|i,z)P(z|u)$ قرار می‌دادیم $p(v|u,i) = \sum_z p(v|u,z)P(z|i)$ و سپس $p(v|u,z) \sim N(v|\mu_{u,z}, \sigma^2_{u,z})$ چه تغییری در تعبیر روش ایجاد می‌شد؟
- روابط مربوط به گام M الگوریتم را به صورت کامل بنویسید.
- در این مقاله هدف از به‌کارگیری نرمال سازی کاربری[۱۹] چه بوده است؟
- روش‌های مختلف برای regularization مدل را شرح دهید. (به بخش ۵.۳ مقاله و همچنین به مرجع [۳] مراجعه شود.)
- (**نمره اضافه**) عیب‌های عمده مدل را بیان نمایید. (می‌توانید مراجع [۴–۶] را مطالعه نمایید.)

سوالات عملی

- روش ارائه شده در مقاله را در دو حالت بدون نرمال‌سازی کاربری و با نرمال‌سازی کاربری پیاده‌سازی نمایید و با روش پایه‌ی ارائه شده در مقاله مقایسه نمایید.
- (۱۰ درصد **نمره اضافه**) الگوریتم ارائه شده در مقاله را بهبود دهید (از مراجع [۴–۶] استفاده نمایید)

---

[17] Expectation Maximization (EM)
[18] posterior
[19] user normalization

معیار و نحوه‌ی ارزیابی

- معیار ارزیابی RMSE در نظر گرفته شود.
- برای تعیین تعداد گروه‌های (خوشه‌های) مناسب از 10-fold cross validation استفاده کنید. نمودار تغییرات دقت به دست آمده برای تعداد گروه‌های مختلف را رسم نمایید.
- RMSE روی مجموعه آزمون را به ازای این روش مشخص نمایید. نتایج به دست آمده را با نتایجی که در فاز قبل به‌دست آمد، مقایسه نمایید.

فاز سوم (اختیاری) : استفاده از روابط اعتماد و عدم اعتماد

زمان تحویل فاز: ۱۱ تیر

نمره : ۲۵ درصد (اضافه)

هدف : تحقیق، پژوهش و ابتکار

در بعضی از پایگاه‌های داده‌ای مثل Epinions[20] برای کاربران علاوه بر رتبه‌بندی آیتم‌ها، امکان رتبه‌بندی کاربران دیگر نیز فراهم شده که به رتبه‌های بین کاربران روابط اعتماد و عدم اعتماد گفته می‌شود. در واقع می‌توان روابط بین کاربران و آیتم‌ها را با استفاده از یک شبکه وزن‌دار ناهمگون[21] مدل کرد [۷]. با استفاده از مطالبی که در فازهای قبل پروژه فرا گرفتید، مدلی برای بهبود پیش‌بینی رای یک کاربر به یک آیتم و هم‌چنین پیش‌بینی رابطه اعتماد بین دو فرد ارائه دهید. برای دانلود پایگاه داده به http://www.trustlet.org/wiki/Downloaded_Epinions_dataset مراجعه نمایید. برای مطالعه مقاله‌هایی در این زمینه به http://www.trustlet.org/wiki/Epinions_datasets مراجعه نمایید.

مراجع

[1] P. Melville, V. Sindhwani, "Recommender Systems", Encyclopedia of Machine Learning, Springer, 2010.

[2] T. Hofmann, "Collaborative filtering via gaussian probabilistic latent semantic analysis", in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 259-266.

[3] T. Hofmann, "Latent semantic models for collaborative filtering", ACM Transactions on Information Systems (TOIS), vol. 22, pp. 89-115, 2004.

[4] B. Marlin, "Modeling user rating profiles for collaborative filtering", Advances in Neural Information Processing Systems, vol. 16, 2003.

[5] B. Marlin, "Collaborative filtering: A machine learning perspective", University of Toronto, 2004.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[7] J. Han, "Mining heterogeneous information networks: the next frontier", in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 2-3.

[20] http://www.trustlet.org/wiki/Epinions_datasets

[21] Heterogeneous