به نام خداوند بخشنده مهربان

# شبکه های اجتماعی و اقتصادی

دانشکده مهندسی کامپیوتر

مریم رمضانی
بهار ۱۴۰۴

سیستم های توصیه گر، گراف، شبکه | **تمرین چهارم** | تاریخ انتشار: ۱۰ خرداد ۱۴۰۴

۱. سوالات خود در مورد این تمرین را در کوئرا مطرح کنید.

۲. لطفا پاسخ خود را با توضیحات کافی ارائه دهید.

## سوالات تئوری (۸۰ نمره)

تاریخ تحویل: ۲۳ خرداد ۱۴۰۴

**پرسش ۱** (۲۰ نمره)

**سوال: تحلیل گراف‌های ناهمگن و کاربرد Meta-Path در خوشه‌بندی**

**مدل‌سازی گراف‌های ناهمگن:**

گرافی ناهمگن تعریف کنید که شامل نودهای متفاوت مانند مقاله‌ها، ژورنال‌ها و نویسندگان باشد. همچنین گراف دیگری شامل فیلم‌ها، کارگردان‌ها و بازیگران ایجاد نمایید. در هر دو گراف، نوع نودها و نوع یال‌ها (روابط میان نودها) را مشخص کنید.

**تعریف Meta-Path:**

مفهوم Meta-Path را در گراف‌های ناهمگن توضیح دهید. چرا این مفهوم برای تحلیل رابطه‌های غیر مستقیم بین نودها مهم است؟

**کاربرد Meta-Path در مدل‌سازی گراف:**

نحوه‌ی استفاده از Meta-Pathها برای مدل‌سازی روابط میان نودهای هم‌نوع (مثلاً نویسنده با نویسنده یا فیلم با فیلم) را توضیح دهید. چگونه می‌توان از این مدل‌سازی برای اعمال الگوریتم‌های Community Detection استفاده کرد؟

**طراحی Meta-Path برای گراف فیلم‌ها:** حداقل سه Meta-Path مختلف پیشنهاد دهید که برای خوشه‌بندی (Clustering) فیلم‌ها مفید باشند. برای هر Meta-Path توضیح دهید چه نوع ارتباطی میان فیلم‌ها برقرار می‌شود و چرا برای کشف جامعه‌های معنایی مناسب است.

**تعریف Meta-Path برای گراف مقالات جهت خوشه‌بندی نویسندگان:** چگونه می‌توان از Meta-Pathها برای تعریف شباهت میان نویسندگان استفاده کرد؟ چند نمونه Meta-Path ارائه دهید که خوشه‌بندی نویسندگان بر اساس آن‌ها معنادار باشد (مثلاً همکاری مشترک، انتشار در ژورنال مشابه و ...).

**پاسخ**

# ۱ مدل‌سازی گراف‌های ناهمگن

## ۱.۱ گراف مقالات

**نودها:**

- مقاله
- نویسنده
- ژورنال

**یال‌ها:**

- نویسنده $\rightarrow$ _نوشته است_ $\leftarrow$ مقاله
- مقاله $\rightarrow$ _منتشر شده در_ $\leftarrow$ ژورنال

## ۲.۱ گراف فیلم‌ها

**نودها:**

- فیلم
- کارگردان
- بازیگر

**یال‌ها:**

- کارگردان $\rightarrow$ _کارگردانی کرده_ $\leftarrow$ فیلم
- بازیگر $\rightarrow$ _بازی کرده در_ $\leftarrow$ فیلم

## ۲ تعریف Meta-Path

- **Meta-Path** دنباله‌ای از **انواع نودها** و **انواع یال‌ها** است که مسیرهای معنادار در گراف ناهمگن را تعریف می‌کند.
- Meta-Path امکان تحلیل روابط غیر مستقیم را فراهم می‌کند. در بسیاری موارد، نودها ارتباط مستقیم ندارند ولی از طریق زنجیره‌ی معنایی می‌توان مشابهت یا نزدیکی آن‌ها را سنجید.

## ۳ کاربرد Meta-Path در مدل‌سازی گراف

با تعریف Meta-Path‌ها، می‌توان **شباهت میان نودهای هم‌نوع** را مدل‌سازی کرد. این شباهت‌ها می‌توانند مبنای ایجاد **ماتریس شباهت (Similarity Matrix)** باشند. سپس الگوریتم‌های Community Detection روی این ماتریس اعمال می‌شوند و خوشه‌های معنایی استخراج می‌شوند.

## ۴ طراحی Meta-Path برای گراف فیلم‌ها

- فیلم_بازیگر_فیلم: یک بازیگر در هر دو فیلم بازی کرده است. هر بازیگر سبک خاصی دارد و مناسب بازی در برخی فیلم‌ها است.
- فیلم_کارگردان_فیلم: هر دو فیلم توسط یک کارگردان کارگردانی شده‌اند. کارگردان تأثیر زیادی در سبک فیلم دارد و یکی بودن کارگردان نشان‌دهنده شباهت زیاد بین دو فیلم است.
- فیلم_ژانر_فیلم: دو فیلم از یک ژانر هستند. در این حالت لازم است ژانر نیز به نودهای گراف اضافه شود.

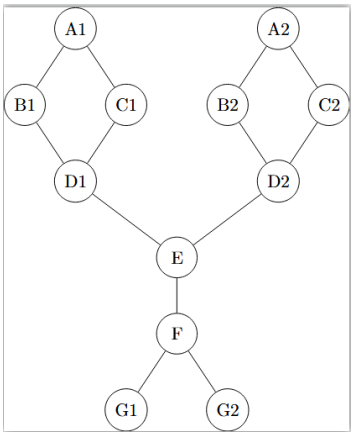
## ۵ تعریف Meta-Path برای گراف مقالات

در گراف مقاله_ژورنال_نویسنده، می‌توان Meta-Path های زیر را تعریف کرد:

- نویسنده_مقاله_نویسنده: اگر دو نویسنده یک کار مشترک داشته باشند به هم مرتبط می‌شوند.
- نویسنده_مقاله_ژورنال_مقاله_نویسنده: اگر دو نویسنده در یک ژورنال مشترک مقاله منتشر کرده باشند به هم مرتبط می‌شوند.

این روابط می‌توانند نشان دهند دو نویسنده حوزه‌های کاری نزدیک دارند و زیاد بودن این مسیرها احتمال هم‌حوزه بودن آن‌ها را تقویت می‌کند.

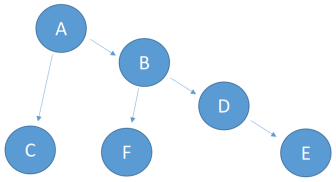
**پرسش ۲** (۲۰ نمره)




اجرای الگوریتم‌های Node2Vec، GraphSAGE و GAT را برای گراف داده شده به صورت گام‌به‌گام تا سه مرحله توضیح دهید. بر اساس پیچیدگی روابط بین رئوس، اندازه گراف و زمانی بودن توضیح دهید در هر حالت کدام الگوریتم مناسب‌تر است و چرا.

**پاسخ**

**پرسش ۳** (۲۰ نمره)



| یال | احتمال |
|---|---|
| $A \to B$ | ۰٫۹ |
| $A \to C$ | ۰٫۴ |
| $B \to D$ | ۰٫۶ |
| $B \to F$ | ۰٫۳ |
| $D \to E$ | ۰٫۵ |

**مرکزیت در فضای تعبیه‌شده (Embedding Space) در برابر مرکزیت ساختاری**

هر گره در یک فضای دوبعدی نهفته تعبیه شده که از طریق راهپیمایی‌های بایاس‌دار (Node2Vec) به دست آمده است. مختصات گره‌ها به صورت زیر است:

| گره | (x, y) |
|---|---|
| A | (9.0, 1.0) |
| B | (8.0, 2.0) |
| C | (7.0, 0.0) |
| D | (6.0, 3.0) |
| E | (5.0, 4.0) |
| F | (7.5, 2.5) |

آ) فاصله اقلیدسی گره F تا سایر گره‌ها را محاسبه کرده و میانگین فاصله‌ها را گزارش دهید.

ب) آیا مرکزیت در فضای تعبیه‌شده لزوماً با پتانسیل تأثیرگذاری در گراف ساختاری همبستگی دارد؟ با ذکر مثال مشخص از این گراف توضیح دهید.

**تحلیل انتشار تأثیر (مدل IC)**

آ) با استفاده از مدل انتشار مستقل (IC) تعداد مورد انتظار گره‌های فعال شده با شروع از گره A را محاسبه کنید. فرض کنید فرایند انتشار در حداکثر سه مرحله انجام می‌شود. فرمول $\text{Expected}(X \longrightarrow Y) = P(X \text{ activates } Y) \times P(X \text{ is active})$ همچنین موارد زیر را نیز محاسبه و تحلیل کنید:

- احتمال فعال شدن گره E
- تفاوت در انتظار تأثیر زمانی که گره B به جای A به‌عنوان seed انتخاب می‌شود

ب) توضیح دهید که چرا از الگوریتم‌های حریصانه در بیشینه‌سازی تأثیر استفاده می‌شود و به صورت خلاصه نشان دهید که تابع تأثیر زیرمجموعه‌گرا (submodular) است.

**دینامیک GRL – همرخدادی متنی در Node2Vec**

فرض کنید تعبیه‌های فوق با استفاده از الگوریتم Node2Vec به‌دست آمده‌اند با تنظیمات زیر: طول راهپیمایی: ۳، دو راهپیمایی برای هر گره، پارامتر بازگشت $p = ۱$، پارامتر in-out برابر $q = ۰/۲۵$ (تمایل به اکتشاف)

آ) دو راهپیمایی نمونه از گره A را شبیه‌سازی کنید. وزن انتخاب‌ها و مسیرها را مشخص کنید. مثال: $A \to B \to F \to B$

ب) بر اساس راهپیمایی‌ها، ماتریس همرخدادی بین گره‌ها را بسازید. این ماتریس چه تأثیری بر شباهت نقطه‌ای بین A و سایر گره‌ها در آموزش skip-gram دارد؟

ج) توضیح دهید که چگونه تغییر پارامتر $q$ در گراف‌های پراکنده و متراکم تعبیه‌ها را تحت تأثیر قرار می‌دهد. این مسئله چه ارتباطی با سیستم‌های توصیه‌گر حساس به تأثیر دارد؟

**ملاحظات راهبردی و نمایش چندوجهی**

آ) محدودیت‌های استفاده از تعبیه‌های ایستا (مانند Node2Vec یا DeepWalk) در فرآیندهای پویا مثل انتشار تأثیر را توضیح دهید. یک روش جایگزین پیشنهاد دهید (مانند GRL زمانی یا مدل‌های مبتنی بر توجه.)

ب) در گراف‌های چندوجهی که تأثیر تنها ساختاری نیست (مثلاً وابسته به محتوا یا زمان یا اعتماد اجتماعی است)، چگونه می‌توان معماری‌های GRL را گسترش داد؟ یک اسکچ از نحوه انتقال پیام در GNN را که شامل احتمال تأثیر است، ارائه دهید.

**به تمامی پرسش‌ها با استدلال دقیق پاسخ دهید. در صورت نیاز به مفاهیم نظری مانند زیرمجموعه‌گرایی، توزیع پایدار راهپیمایی تصادفی یا گرادیان‌های تابع هزینه skip-gram ارجاع دهید.**

**پاسخ**

## پرسش ۳ ـ مرکزیت در تعبیه، انتشار IC، و Vec۲Node

### (۱) مرکزیت در فضای تعبیه

مختصات تعبیهٔ دوبعدی فرضی:

$$A(۹,۱),\quad B(۸,۲),\quad C(۷,۰),\quad D(۶,۳),\quad E(۵,۴),\quad F(۷/۵,۲/۵).$$

برای هر $X \neq F$ داریم $d(F,X) = \sqrt{(x_F - x_X)^۲ + (y_F - y_X)^۲}$:

$$d(F,A) = \sqrt{(۷/۵ - ۹)^۲ + (۲/۵ - ۱)^۲} \approx ۲/۱۲۱۳,$$
$$d(F,B) = \sqrt{(۷/۵ - ۸)^۲ + (۲/۵ - ۲)^۲} \approx ۰/۷۰۷۱,$$
$$d(F,C) = \sqrt{(۷/۵ - ۷)^۲ + (۲/۵ - ۰)^۲} \approx ۲/۵۴۹۵,$$
$$d(F,D) = \sqrt{(۷/۵ - ۶)^۲ + (۲/۵ - ۳)^۲} \approx ۱/۵۸۱۱,$$
$$d(F,E) = \sqrt{(۷/۵ - ۵)^۲ + (۲/۵ - ۴)^۲} \approx ۲/۹۱۵۵.$$

میانگین فاصلهٔ $F$ از بقیه تقریباً ۱/۹۷۴۹ است.

**نکتهٔ تحلیلی.** مرکزی‌بودن در تعبیه لزوماً به معنی قدرتِ اثرگذاری در شبکهٔ انتشار نیست؛ چون انتشار به جهت یال‌ها و احتمال‌های فعال‌سازی وابسته است، نه صرفاً نزدیکی در تعبیه.

## (۲) مدل انتشار مستقل (IC) تا ۳ گام با بذر $A$

یال‌ها و احتمال‌ها:

$$A \to B : ۰/۹, \quad A \to C : ۰/۴, \quad B \to D : ۰/۶, \quad B \to F : ۰/۳, \quad D \to E : ۰/۵.$$

احتمال فعال‌بودن گره‌ها:

$$P(A) = ۱, \quad P(B) = ۰/۹, \quad P(C) = ۰/۴,$$
$$P(D) = ۰/۹ \times ۰/۶ = ۰/۵۴, \quad P(F) = ۰/۹ \times ۰/۳ = ۰/۲۷, \quad P(E) = ۰/۵۴ \times ۰/۵ = ۰/۲۷.$$

تعداد مورد انتظار فعال‌ها:

$$\boxed{۱ + ۰/۹ + ۰/۴ + ۰/۵۴ + ۰/۲۷ + ۰/۲۷ = ۳/۳۸}.$$

همچنین $\boxed{۰/۲۷} = ۰/۹ \times ۰/۶ \times ۰/۵ = P(E \mid A)$.

**اگر بذر $B$ باشد:**

$$P(B) = ۱, \quad P(D) = ۰/۶, \quad P(F) = ۰/۳, \quad P(E) = ۰/۶ \times ۰/۵ = ۰/۳ \Rightarrow \boxed{۱ + ۰/۶ + ۰/۳ + ۰/۳ = ۲/۲۰}.$$

بنابراین تفاوت انتظاری: $\boxed{۱/۱۸} = ۲/۲۰ - ۳/۳۸$.

**چرایی روش حریصانه.** تابع گسترشِ مورد انتظار در IC زیرمدولار است (بازده کاهنده)، لذا انتخاب حریصانهٔ $k$ بذر تضمین $۱ - ۱/e$ دارد.

## (۳) Vec۲Node با $p = ۱,\ q = ۰/۲۵$

در گذار مرتبهٔ دوم، وزن‌ها:

$$\alpha_{pq}(t,x) = \begin{cases} ۱/p & x = t, \\ ۱ & (t,x) \in E, \\ ۱/q & (t,x) \notin E. \end{cases} \quad \Rightarrow \quad ۱/q = ۴.$$

پس از گام $A \to B$، همسایه‌های $B$ برابر $\{A, D, F\}$ است و

$$w(A) = ۱, \quad w(D) = ۴, \quad w(F) = ۴ \ \Rightarrow \ P(A) = \tfrac{۱}{۹},\ P(D) = \tfrac{۴}{۹},\ P(F) = \tfrac{۴}{۹}.$$

**(الف) دو راه‌روی نمونهٔ طول ۳ از هر گره**

A: $A \to B \to F \to B$, $A \to B \to D \to E$
B: $B \to D \to E \to D$, $B \to F \to B \to D$
C: $C \to A \to B \to D$, $C \to A \to B \to F$
D: $D \to E \to D \to B$, $D \to B \to F \to B$
E: $E \to D \to B \to D$, $E \to D \to B \to F$
F: $F \to B \to D \to E$, $F \to B \to A \to B$

**(ب) ماتریس هم‌وقوعی با پنجرهٔ $w = ۲$**

جدول زیر با شمردن هم‌وقوعی گره‌ها در فاصلهٔ موضعی $\leq ۲$ داخل ۱۲ راه‌رو بالا به‌دست آمده است.

| F | E | D | C | B | A | |
|---|---|---|---|---|---|---|
| ۳ | ۰ | ۲ | ۲ | ۶ | ۰ | A |
| ۱۰ | ۶ | ۱۰ | ۲ | ۰ | ۶ | B |
| ۰ | ۰ | ۰ | ۰ | ۲ | ۲ | C |
| ۴ | ۸ | ۰ | ۰ | ۱۰ | ۲ | D |
| ۰ | ۰ | ۸ | ۰ | ۶ | ۰ | E |
| ۰ | ۰ | ۴ | ۰ | ۱۰ | ۳ | F |

**تفسیر.** هم‌وقوعی‌های بزرگ بین $(B, D)$، $(B, F)$ و $(A, B)$ دیده می‌شود؛ بنابراین در تعبیه، نمایش $A$ به این گره‌ها نزدیک‌تر خواهد شد. مقدار $q < ۱$ رفتارِ مشابهِ DFS و اکتشاف دورتر را تقویت می‌کند؛ درحالی‌که $q > ۱$ به رفتار محلی‌ترِ شبیه BFS منجر می‌شود.

**پرسش ۹ (۲۰ نمره) سیستم‌های پیشنهاددهنده**

یک گراف دو بخشی user-item را در نظر بگیرید که در آن هر یال بین کاربر $U$ و آیتم $I$ نشان‌دهنده‌ی این است که کاربر $U$ آیتم $I$ را پسندیده است. همچنین ماتریس امتیازدهی را برای این مجموعه از کاربران و آیتم‌ها با $R$ نمایش می‌دهیم، که در آن هر سطر از $R$ مربوط به یک کاربر و هر ستون مربوط به یک آیتم است. اگر کاربر $i$ آیتم $j$ را پسندیده باشد، آنگاه $R_{i,j} = 1$، در غیر این صورت $R_{i,j} = 0$. همچنین فرض می‌کنیم $m$ کاربر و $n$ آیتم داریم، بنابراین ماتریس R ابعادی برابر با $m \times n$ خواهد داشت.

ماتریس $P$ با ابعاد $m \times m$ را تعریف می‌کنیم به عنوان یک ماتریس قطری که در آن عنصر قطری $i$ام برابر است با درجه‌ی رأس کاربر $i$، یعنی تعداد آیتم‌هایی که کاربر $i$ آن‌ها را پسندیده است. به‌صورت مشابه، ماتریس $Q$ با ابعاد $n \times n$ یک ماتریس قطری است که در آن عنصر قطری $i$ام برابر است با درجه‌ی رأس آیتم $i$، یا به عبارتی تعداد کاربرانی که آیتم $i$ را پسندیده‌اند. برای مثال به شکل زیر توجه کنید.

**الف)**

ماتریس شباهت کاربر به صورت غیر نرمال شده را به صورت $T = RR^T$ تعریف می‌کنیم (یعنی حاصل‌ضرب ماتریس R و ترانهاده‌ی آن). مفهوم عناصر $T_{ij}$ (برای $i \neq j$) را از منظر ساختارهای گراف دو بخشی (مانند درجه‌ی رأس‌ها، مسیر میان رأس‌ها و ...) توضیح دهید (به شکل ۱ مراجعه کنید).

**شباهت کسینوسی:**

یادآوری می‌کنیم که شباهت کسینوسی بین دو بردار $\mathbf{u}$ و $\mathbf{v}$ به صورت زیر تعریف می‌شود:

$$\cos\text{-sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

**ب)**

ماتریس **شباهت آیتم** را به صورت $S_I$ با ابعاد $n \times n$ تعریف می‌کنیم، به‌طوری‌که عنصر سطر $i$ و ستون $j$ برابر است با شباهت کسینوسی بین **آیتم** $i$ و **آیتم** $j$ که متناظر با ستون‌های $i$ و $j$ از ماتریس R هستند. نشان دهید که:

$$S_I = Q^{-۱/۲} R^T R Q^{-۱/۲}$$

که در آن $Q^{-1/2}$ به صورت $1/\sqrt{Q_{rc}}$ برای تمام درایه‌های غیر صفر تعریف شده است و برای سایر درایه‌ها برابر صفر است.

همین سؤال را برای **ماتریس شباهت کاربران**، یعنی $S_U$ تکرار کنید؛ به‌طوری‌که عنصر سطر $i$ و ستون $j$ برابر است با شباهت کسینوسی بین **کاربر** $i$ و **کاربر** $j$ که متناظر با سطرهای $i$ و $j$ از ماتریس R هستند. عبارت $S_U$ را نیز به صورت یک عملیات ماتریسی بر حسب R، P و Q بیان کنید؛ توجه داشته باشید که نباید به صورت درایه به درایه $S_U$ را تعریف نمایید.

---

$$\mathbf{R} = \begin{bmatrix} ۱ & ۰ & ۱ \\ ۱ & ۰ & ۰ \\ ۱ & ۱ & ۱ \\ ۰ & ۰ & ۱ \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} ۲ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۳ & ۰ \\ ۰ & ۰ & ۰ & ۱ \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} ۳ & ۰ & ۰ \\ ۰ & ۱ & ۰ \\ ۰ & ۰ & ۳ \end{bmatrix}$$

*شکل ۱: گراف دو بخشی کاربر–آیتم.*

---

پاسخ شما باید نحوه‌ی استخراج این رابطه‌ها را نشان دهد. (نکته: برای جذر درایه‌به‌درایه‌ی یک ماتریس، می‌توان آن را به توان ۱/۲ نوشت.)

**(ج)**

روش پیشنهادی بر اساس فیلترینگ مشارکتی کاربر–کاربر برای کاربر $u$ به این صورت است: برای تمام آیتم‌ها $s$ محاسبه کنید

$$r_{u,s} = \sum_{x \in \text{کاربران}} \cos(\text{sim}(x,u)) \cdot R_{xs}$$

و سپس $k$ آیتمی را پیشنهاد دهید که مقدار $r_{u,s}$ در آن‌ها بیشینه است.

به‌صورت مشابه، روش پیشنهادی بر اساس فیلترینگ مشارکتی آیتم–آیتم برای کاربر $u$ به این صورت است: برای تمام آیتم‌ها $s$ محاسبه کنید

$$r_{u,s} = \sum_{x \in \text{آیتم‌ها}} R_{ux} \cdot \cos(\text{sim}(x,s))$$

و سپس $k$ آیتمی را پیشنهاد دهید که مقدار $r_{u,s}$ در آن‌ها بیشینه است.

ماتریس پیشنهادی $\Gamma$ را تعریف می‌کنیم که ابعادی برابر با $m \times n$ دارد، به‌طوری‌که $\Gamma(i,j) = r_{i,j}$. ماتریس $\Gamma$ را برای هر دو روش **فیلترینگ کاربر–کاربر** و **فیلترینگ آیتم–آیتم** برحسب $R$، $P$ و $Q$ بیابید. پاسخ نهایی شما باید شامل عملیات در سطح ماتریسی باشد و نه تعریف درایه به درایه.

**راهنمایی:** در حالت آیتم–آیتم داریم:

$$\Gamma = RQ^{-۱/۲}R^T.$$

پاسخ شما باید روند رسیدن به این عبارت‌ها را (حتی در مورد آیتم–آیتم که فرمول نهایی داده شده) توضیح دهد.

**(د)**

جدول امتیازدهی زیر را بین پنج کاربر و شش آیتم در نظر بگیرید:

| شناسه آیتم | ۱ | ۲ | ۳ | ۴ | ۵ | ۶ |
|---|---|---|---|---|---|---|
| ۱ | ۵ | ۶ | ۷ | ۴ | ۳ | ؟ |
| ۲ | ۴ | ؟ | ۳ | ؟ | ۵ | ۴ |
| ۳ | ؟ | ۳ | ۴ | ۱ | ۱ | ؟ |
| ۴ | ۷ | ۴ | ۳ | ۶ | ؟ | ۴ |
| ۵ | ۱ | ؟ | ۳ | ۲ | ۲ | ۵ |

جدول ۱: امتیازدهی پنج کاربر (ردیف‌ها) به شش آیتم (ستون‌ها).

(د–۱) مقادیر نامشخص امتیازدهی کاربر ۲ را با استفاده از الگوریتم فیلترینگ مشارکتی مبتنی بر کاربر پیش‌بینی کنید. از ضریب همبستگی پیرسون با میانگین‌گیری استفاده کنید. فرض کنید اندازه گروه همتایان حداکثر ۲ است و همبستگی‌های منفی را حذف نمایید.

(د–۲) مقادیر نامشخص امتیازدهی کاربر ۲ را با استفاده از الگوریتم فیلترینگ مشارکتی مبتنی بر آیتم پیش‌بینی کنید. از شباهت کسینوسی تعدیل‌شده استفاده کنید. فرض کنید اندازه گروه همتایان حداکثر ۲ است و همبستگی‌های منفی را حذف نمایید.

اکنون، یک سامانه پیشنهاددهنده فیلم را در نظر بگیرید که در آن فیلم‌ها با ژانرها مرتبط هستند و امتیازدهی یک کاربر خاص مشخص است:

| شناسه فیلم | کمدی | درام | عاشقانه | هیجانی | اکشن | ترسناک | دوست داشتن یا نداشتن |
|---|---|---|---|---|---|---|---|
| ۱ | ۱ | ۰ | ۱ | ۰ | ۰ | ۰ | دوست ندارد |
| ۲ | ۱ | ۱ | ۱ | ۰ | ۱ | ۰ | دوست ندارد |
| ۳ | ۱ | ۱ | ۰ | ۰ | ۰ | ۰ | دوست ندارد |
| ۴ | ۰ | ۰ | ۰ | ۱ | ۱ | ۰ | دوست دارد |
| ۵ | ۰ | ۱ | ۰ | ۱ | ۱ | ۱ | دوست دارد |
| ۶ | ۰ | ۰ | ۰ | ۰ | ۱ | ۱ | دوست دارد |
| آزمون–۱ | ۰ | ۰ | ۰ | ۱ | ۰ | ۱ | ؟ |
| آزمون–۲ | ۰ | ۱ | ۱ | ۰ | ۰ | ۰ | ؟ |

جدول ۲: رابطه ژانر–فیلم و بازخورد کاربر در قالب دوست داشتن یا نداشتن.

تمام قوانین انجمنی را با حداقل پشتیبانی ۳۳٪ و اطمینان ۷۵٪ استخراج کنید. بر اساس این قوانین، آیا آیتم **آزمون–۱** یا **آزمون–۲** را به کاربر پیشنهاد می‌کنید؟

**موارد مورد نیاز برای تحویل:**

(i) تفسیر $T_{ii}$ و $T_{ij}$.
(ii) بیان $S_I$ و $S_U$ برحسب ماتریس‌های $R$، $P$ و $Q$ به همراه توضیح.
(iii) بیان $\Gamma$ برحسب $R$، $P$ و $Q$ به همراه توضیح.
(iv) محاسبات مبتنی بر تکنیک‌های فیلترینگ مشارکتی.
(v) پاسخ به سوالات.

**پاسخ**

# پاسخ سؤال ۹

## (الف)

می‌دانیم $R_{ij} \in \{۰, ۱\}$ و $R_{ij} = R_{ji}^\top$. بنابراین می‌توانیم $T_{ii}$ و $T_{ij}$ را به صورت زیر محاسبه کنیم:

$$T_{ii} = \sum_{k=۱}^{n} R_{ik}\, R_{ki}^\top = \sum_{k=۱}^{n} R_{ik}^۲.$$

از آن‌جا که اگر $R_{ik} = ۱$ آنگاه $R_{ik}^۲ = ۱$ و در غیر این صورت ۰ است، خواهیم داشت

$$T_{ii} = \sum_{k=۱}^{n} R_{ik}.$$

پس $T_{ii}$ برابر است با تعداد آیتم‌هایی که کاربر $i$ آن‌ها را دوست دارد (درجهٔ گرهٔ کاربر $i$).
برای $i \neq j$:

$$T_{ij} = \sum_{k=۱}^{n} R_{ik}\, R_{kj}^\top = \sum_{k=۱}^{n} R_{ik}\, R_{jk}.$$

عبارت $R_{ik}R_{jk} = ۱$ تنها وقتی برقرار است که هر دو کاربر $i$ و $j$ آیتم $k$ را دوست داشته باشند؛ بنابراین $T_{ij}$ تعداد آیتم‌های مشترکی را می‌شمارد که هر دو کاربر $i$ و $j$ می‌پسندند (و معادل تعداد مسیرها بین $i$ و $j$ در گراف کاربر–آیتم است).

## (ب)

بگذارید $R_i^\top$ بردار سطری $i$ام در $R^\top$ (امتیازهای آیتم $i$) باشد. سپس $R^{۰\top}$ را طوری تعریف می‌کنیم که

$$R_{ij}^{:\top} = \frac{R_{ij}^\top}{\|R_i^\top\|}, \qquad \|R_i^\top\| = \sqrt{\sum_{j=۱}^{m} \left(R_{ij}^\top\right)^۲}.$$

از آن‌جا که $\|R_i^\top\|$ ریشهٔ تعداد کاربرانی است که آیتم $i$ را دوست دارند، و این تعداد همان درجهٔ گرهٔ آیتم $i$ است، می‌توان نوشت

$$\|R_i^\top\| = \sqrt{Q_{ii}}.$$

ماتریس $Q^{-۱/۲}$ را چنین تعریف می‌کنیم:

$$(Q^{-۱/۲})_{rc} = \begin{cases} \frac{۱}{\sqrt{Q_{rc}}} & \text{برای درایه‌های نامساوی با صفر،} \\ ۰ & \text{در غیر این صورت.} \end{cases}$$

در نتیجه

$$R^{۰\top} = Q^{-۱/۲} R^\top.$$

شباهت کسینوسیِ آیتم‌های $i$ و $j$ برابر است با

$$\frac{\sum_{k=۱}^{m} R_{ik}^\top R_{jk}^\top}{\|R_i^\top\|\,\|R_j^\top\|} = \frac{\sum_{k=۱}^{m} R_{ik}^\top R_{jk}^\top}{\sqrt{Q_{ii}}\sqrt{Q_{jj}}} = R_i^{:\top} \cdot \left(R_j^{:\top}\right)^\top.$$

پس ماتریس شباهت آیتم‌ها داریم

$$S_I = R^{۰\top}\left(R^{۰\top}\right)^\top.$$

از آن‌جا که $Q^{-۱/۲}$ متقارن است، $(R^{۰\top})^\top = (Q^{-۱/۲}R^\top)^\top = RQ^{-۱/۲}$. در نتیجه

$$S_I = Q^{-۱/۲} R^\top R\, Q^{-۱/۲}.$$

به طور مشابه، $R_i$ را بردار سطری $i$ام در $R$ (امتیازهای کاربر $i$) بگیرید و $R^{\star}$ را طوری تعریف کنید که

$$R^{\star}_{ij} = \frac{R_{ij}}{\|R_i\|}, \qquad \|R_i\| = \sqrt{\sum_{j=1}^{n} R_{ij}^{2}} = \sqrt{P_{ii}},$$

که در آن $P_{ii}$ درجهٔ کاربر $i$ است. بگذارید $P^{-1/2}$ همانند قبل تعریف شود؛ آنگاه $R^{\star} = P^{-1/2}R$ و شباهت کسینوسی کاربران $i$ و $j$ برابر است با

$$\frac{\sum_{k=1}^{n} R_{ik}R_{jk}}{\|R_i\|\,\|R_j\|} = \frac{\sum_{k=1}^{n} R_{ik}R_{jk}}{\sqrt{P_{ii}}\sqrt{P_{jj}}} = R^{\star}_{i} \cdot \left(R^{\star}_{j}\right)^{\top}.$$

پس ماتریس شباهت کاربران

$$S_U = R^{\star}\left(R^{\star}\right)^{\top} = P^{-1/2}RR^{\top}P^{-1/2}.$$

## (ج)

**(۱) فیلترگذاری مشارکتی کاربر–کاربر.** برای پیش‌بینی $r_{u,s}$ داریم

$$r_{u,s} = \sum_{x \in \text{users}} \text{cos-sim}(x,u)\, R_{x,s} = \sum_{x \in \text{users}} \left(P^{-1/2}RR^{\top}P^{-1/2}\right)_{u,x} R_{x,s} = \left(P^{-1/2}RR^{\top}P^{-1/2}\right)_{u} \cdot R_s.$$

در نتیجه، ماتریس توصیه

$$\Gamma \;=\; P^{-1/2}RR^{\top}P^{-1/2}\,R.$$

**(۲) فیلترگذاری مشارکتی آیتم–آیتم.** برای پیش‌بینی $r_{u,s}$ داریم

$$r_{u,s} = \sum_{x \in \text{items}} R_{u,x}\,\text{cos-sim}(x,s) = \sum_{x \in \text{items}} R_{u,x}\left(Q^{-1/2}R^{\top}RQ^{-1/2}\right)_{x,s} = R_u \cdot \left(Q^{-1/2}R^{\top}RQ^{-1/2}\right)_{s}.$$

پس ماتریس توصیه

$$\Gamma \;=\; RQ^{-1/2}R^{\top}RQ^{-1/2}.$$

**توضیح:** در این فرمول‌ها، $P$ و $Q$ ماتریس‌های قطریِ درجات کاربران و آیتم‌ها هستند.

## (د) (ترجمهٔ پاسخِ سؤال ۲: پیش‌بینی امتیازهای کاربر ۲ با جداول کامل)

فرض کنید جدول امتیازها بین پنج کاربر و شش آیتم به‌صورت زیر است (علامت «؟» یعنی مقدار نامشخص):

| ۶ | ۵ | ۴ | ۳ | ۲ | ۱ | ⇒ **Item-Id** |
|---|---|---|---|---|---|---|
| | | | | | | ⇓ **User-Id** |
| ؟ | ۳ | ۴ | ۷ | ۶ | ۵ | ۱ |
| ۴ | ۵ | ؟ | ۳ | ؟ | ۴ | ۲ |
| ؟ | ۱ | ۱ | ۴ | ۳ | ؟ | ۳ |
| ۴ | ؟ | ۶ | ۳ | ۴ | ۷ | ۴ |
| ۵ | ۲ | ۲ | ۳ | ؟ | ۱ | ۵ |

**گام اوّل: جدول میان‌مرکزشده (برای پیرسون و adjusted cosine).** برای هر کاربر، میانگین سطر را از امتیازهای همان سطر کم می‌کنیم:

| **Mean** | ۶ | ۵ | ۴ | ۳ | ۲ | ۱ | ⇒ **Item-Id** |
|---|---|---|---|---|---|---|---|
| ۵ | ؟ | -۲ | -۱ | ۲ | ۱ | ۰ | ۱ |
| ۴ | ۰ | ۱ | ؟ | -۱ | ؟ | ۰ | ۲ |
| ۲.۲۵ | ؟ | -۱.۲۵ | -۱.۲۵ | ۱.۷۵ | ۰.۷۵ | ؟ | ۳ |
| ۴.۸ | -۰.۸ | ؟ | ۱.۲ | -۱.۸ | -۰.۸ | ۲.۲ | ۴ |
| ۲.۶ | ۲.۴ | -۰.۶ | -۰.۶ | ۰.۴ | ؟ | -۱.۶ | ۵ |

**(الف) روشِ کاربر–کاربر Pearson) + .(mean-centering)** ضریب‌های پیرسون بین کاربر ۲ و سایر کاربران (بر اساس آیتم‌های مشترکِ ارزیابی‌شده) به‌صورت زیر محاسبه می‌شوند:

$$\text{Pearson}(۱,۲) = \frac{(۲)(-۱)+(-۲)(۱)}{\sqrt{۲^۲+(-۲)^۲}\sqrt{(-۱)^۲+۱^۲}} = -۱,$$

$$\text{Pearson}(۳,۲) \approx \frac{(-۱/۷۵)(-۱)+(-۱/۲۵)(۱)}{\sqrt{۱/۷۵^۲+(-۱/۲۵)^۲}\sqrt{(-۱)^۲+۱^۲}} = \frac{-۳}{۳/۰۴} \approx -۰/۹۹,$$

$$\text{Pearson}(۴,۲) \approx \frac{۱/۸}{۲/۹۵} \approx ۰/۶۱, \qquad \text{Pearson}(۵,۲) \approx \frac{-۰/۴-۰/۶}{\sqrt{۲}\sqrt{۱/۶^۲+۰/۴^۲+۰/۶^۲+۲/۴^۲}} \approx -۰/۲۳.$$

با توجه به شرط «همسایگی حداکثر ۲ و حذف ضرایب منفی»، نزدیک‌ترین همسایهٔ معتبر فقط کاربر ۴ است. بنابراین از سطرِ میان‌مرکزشدهٔ کاربر ۴ برای آیتم‌های نامشخص کاربر ۲ استفاده می‌کنیم:

$$\widehat{r}^{(mc)}_{۲,۲} = -۰/۸, \qquad \widehat{r}^{(mc)}_{۲,۴} = ۱/۲.$$

با افزودن میانگین کاربر ۲ (یعنی ۴):

$$\hat{r}_{۲,۲} = ۴+(-۰/۸) = ۳/۲, \qquad \hat{r}_{۲,۴} = ۴+۱/۲ = ۵/۲.$$

**(ب) روشِ آیتم–آیتم Cosine) .(Adjusted** برای آیتم ۲، نزدیک‌ترین آیتم‌ها ۳ و ۶ با شباهت‌های ۰/۹۹۷۷ و ۰/۷۰۷۱ هستند؛ با وزن‌دهیِ امتیازهای کاربر ۲ روی آن‌ها (۳ و ۴):

$$\hat{r}_{۲,۲} = \frac{۰/۹۹۷۷ \times ۳ + ۰/۷۰۷۱ \times ۴}{۰/۹۹۷۷ + ۰/۷۰۷۱} \approx ۳/۴۱.$$

برای آیتم ۴، نزدیک‌ترین آیتم‌ها ۱ و ۵ با شباهت‌های ۰/۷۹۱ و ۰/۹۴ هستند و امتیازهای کاربر ۲ روی آن‌ها ۴ و ۵ است:

$$\hat{r}_{۲,۴} = \frac{۰/۷۹۱ \times ۴ + ۰/۹۴ \times ۵}{۰/۷۹۱ + ۰/۹۴} \approx ۴/۵۴.$$

**جمع‌بندی.** در هر دو رویکرد، ترتیب ترجیح یکسان است ( $\hat{r}_{۲,۴} > \hat{r}_{۲,۲}$ )، ولی مقادیر مطلق متفاوت‌اند. برای تکمیل گزارش، می‌توانید در جدول اوّل (خام) مقادیر پیش‌بینی‌شدهٔ $\hat{r}_{۲,۲}$ و $\hat{r}_{۲,۴}$ را جای «؟»ها وارد کنید.

## (هـ) قوانین انجمنی و ترجیح کاربرِ فعال

**قوانین استخراج‌شده.**

- کمدی، درام ⇒ نپسندیدن
- کمدی، عاشقانه ⇒ نپسندیدن
- کمدی ⇒ نپسندیدن
- عاشقانه ⇒ نپسندیدن
- دلهره/هیجان‌انگیز ⇒ پسندیدن
- اکشن ⇒ پسندیدن
- ترسناک ⇒ پسندیدن
- دلهره/هیجان‌انگیز، اکشن ⇒ پسندیدن
- ترسناک، اکشن ⇒ پسندیدن

**نتیجه‌گیری.** به‌سادگی دیده می‌شود که Test-1 تمام قوانینی را که گزارهٔ نتیجهٔ آن‌ها «پسندیدن» است فعال می‌کند، در حالی که Test-2 همهٔ قوانینی را که نتیجه‌شان «نپسندیدن» است شلیک می‌کند. بنابراین کاربرِ فعال، Test-1 را به Test-2 ترجیح خواهد داد.

**توضیح تکمیلی .**

- وجود قوانین عام مانند «کمدی ⇒ نپسندیدن» باعث می‌شود قوانین خاص‌ترِ هم‌راستا (مثل «کمدی، عاشقانه ⇒ نپسندیدن») نیز بدون تناقض عمل کنند.
- اگر چند قانونِ هم‌زمان فعال شوند و همهٔ آن‌ها یک نتیجه داشته باشند (همه «پسندیدن» یا همه «نپسندیدن»)، تصمیم نهایی پایدار است. در مسائل واقعی، معمولاً برای حل تعارض یا اولویت‌بندی از وزن‌دهی بر اساس *confidence* و *support* هر قانون یا از رای‌گیریِ وزن‌دار استفاده می‌شود.

---

# سوالات عملی (۴۰ نمره)

تاریخ تحویل:

---

**پرسش ۱** (۲۰ نمره) به فایل جوپیتر Q1.ipynb مراجعه کنید.

**پاسخ**

**پرسش ۲** (۲۰ نمره) به فایل جوپیتر Q2.ipynb مراجعه کنید.

**پاسخ**