داده‌کاوی ـ ترم پاییز ـ سال ۹۸
دانشگاه صنعتی خواجه نصیرالدین طوسی ـ دانشکده ریاضی

مسائل برای حل (مهلت تحویل تا ۲۱ دیماه)

(Frequent Itemsets)

۱. مسائل زیر را از کتاب مرجع Han, Kamber, Pei حل کنید.

6.1, 6.2, 6.4, 6.6 (only for Apriori), 6.14 (a)

۲. توضیح دهید که با استفاده از الگوریتم Count-Min که در کلاس توضیح داده شد، چگونه عناصر پرتکرار (آنهایی که بیشتر از ϵn بار تکرار شده‌اند) را پیدا کنیم. پیچیدگی زمانی و فضای مصرفی راه حل شما چند است؟

(Maximum Likelihood Estimation)

۳. فرض کنید احتمال آمدن خط برای سکه‌ای p باشد. این سکه را آزمون اول ۱۰ بار پرتاب کرده‌ایم، ۳ بار خط آمده است. در آزمون دوم ۲۰ بار پرتاب کرده‌ایم، ۸ بار خط آمده است. در آزمون سوم ۳۰ بار پرتاب کرده‌ایم، ۱۲ خط آمده است. با در نظر گرفتن این آزمونها، محتملترین مقدار برای p حدودا چند است؟

۴. مانند سوال بالا، فرض کنید آزمون پرتاب سکه را n بار اجرا کنیم. در هر آزمون سکه را t بار پرتاب میکنیم و تعداد رخداد آمدن خط را ثبت می‌کنیم. فرض کنید تعداد خط ها در n آزمون به ترتیب $t_1, ..., t_n$ باشد. با فرض اینکه یکی از آزمونها نتیجه بسیار پرت و متفاوت با بقیه بدهد، توضیح دهید آزمون پرت را چگونه پیدا می‌کنید؟

۵. فرض کنید $X_1, \ldots, X_n$ یک سری نمونه از یک توزیع پیوسته با تابع چگالی زیر باشند.

$$f(x, \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \qquad x > 0, \theta > 0$$

تابع تخمین درست نمایی را برای پارامتر θ بدست آورید. با توجه به مشاهدات زیر تخمینی برای θ بدست آورید.

$$x_1 = 0.50, x_2 = 1.5, x_3 = 4.00, x_4 = 3.00$$

۶. فرض کنید $X_1, \ldots, X_n$ داده شده است. هر X_i مقداری تصادفی از بازه $[0, \theta]$ است که بصورت یکنواخت و مستقل نمونه برداری شده است. پارامتر θ مجهول است. با استفاده از تابع تخمین درست نمایی، یک تخمین برای θ بدست آورید.

(Graph Laplacian)

۷. مقادیر ویژه ماتریس لاپلاسین برای گراف K_n (گراف کامل با n راس) را بدست آورید.
راهنمایی: نشان دهید مقادیر ویژه برابر با

$$(\lambda_1, \lambda_2, \ldots, \lambda_n) = (0, n, \ldots, n)$$

است.

۸. مقادیر ویژه ماتریس لاپلاسین برای گراف C_n (دور با n راس) را بدست آورید.
راهنمایی: نشان دهید.

$$\lambda_k = 2 - 2cos(\frac{2\pi k}{n})$$

۹. مقادیر ویژه ماتریس لاپلاسین برای گراف S_n (گراف ستاره با n راس) را بدست آورید.
راهنمایی: نشان دهید مقادیر ویژه برابر با

$$(\lambda_1, \lambda_2, \ldots, \lambda_n) = (0, 1, \ldots, 1, n)$$

است.

۱۰. مقدار ratioCut را برای هر کدام از گرافهای بالا بدست آورید. مقدار بدست آمده چه رابطه‌ای با λ_2 دارد؟

(LSH)

۱۱. برای هر دو بردار $x, y \in \mathbb{R}^n$ فاصله زاویه‌ای بین دو بردار x, y بصورت زیر تعریف می‌شود.

$$d_\theta(x, y) = \cos^{-1}(\frac{x.y}{\|x\|\|y\|})$$

(در اینجا $\cos^{-1}$ زاویه‌ای در بازه $[0, \pi]$ را برمی‌گرداند.) تشابه زاویه‌ای دو بردار x, y بصورت زیر تعریف می‌شود.

$$s_\theta(x, y) = 1 - \frac{d_\theta(x, y)}{\pi}$$

می‌خواهیم یک LSH برای معیار تشابه زاویه‌ای تعریف کنیم. برای این منظور، یک خانواده تابع هش H تعریف می‌کنیم. برای هر $h_\sigma \in H$ یک بردار تصادفی $\sigma \in \mathbb{R}^n$ (وقتی که $\|\sigma\|^2 = 1$) انتخاب می‌کنیم و تعریف می‌کنیم

$$h_\sigma(x) = sign(x.\sigma)$$

نشان دهید که

$$Pr(h_\sigma(x) = h_\sigma(y)) = s_\theta(x, y)$$

(Clustering)

۱۲. الگوریتم زیر را برای مسئله $k - center$ در نظر بگیرید. در این الگوریتم هر بار شعاع بهینه را حدس می‌زنیم و سعی می‌کنیم نقاط را با توجه به حدسمان خوشه بندی کنیم. اگر تعداد خوشه‌های بدست آمده از k بیشتر شد، سراغ حدس بعدی می‌رویم. فرض کنید قطر نقاط $D = 2^d$ باشد و کمترین فاصله بین دو نقطه برابر با 1 است. الگوریتم را برای حدسهای مختلف (از کوچک به بزرگ)

$$1 = r_0 \leq r_1 \leq r_2 \leq \ldots \leq r_m = D$$

اجرا می‌کنیم. فرض کنید حدس فعلی r باشد. اگر نقطه جدید در فاصله حداکثر $2r$ از نزدیکترین مرکز کنونی واقع شد، نقطه را در خوشه نزدیکترین مرکز قرار می‌دهیم، در غیر این صورت اگر نقطه جدید از همه مراکز کنونی فاصله‌اش بیشتر از $2r$ باشد، یک خوشه جدید با مرکزیت نقطه جدید ایجاد میکنیم. در هر نقطه تعداد مراکز از k بیشتر شد، اجرا را متوقف کرده و سراغ حدس بعدی می‌رویم. (دقت کنید که اولین نقطه همیشه به عنوان یک مرکز انتخاب می‌شود.)
با توجه به توصیف بالا به دو سوال زیر پاسخ دهید:

(آ) فرض کنید OPT شعاع خوشه بندی بهینه باشد. نشان دهید اگر حدس فعلی $r = OPT$ باشد آنگاه حداکثر k مرکز ایجاد می‌شود. به عبارت دیگر با فرض $r = OPT$، یک خوشه بندی با k خوشه و شعاع حداکثر $2OPT$ بدست می‌آید.

(ب) با استفاده از استراتژی بالا، الگوریتمی طراحی کنید که با یک گذر روی داده، مقدار r را بدست آورد بطوریکه

$$OPT \leq r \leq 4OPT$$

فضای مصرفی الگوریتم شما باید $O(k \log D)$ باشد.