



تمرین سری اول مبانی علم داده

مهلت ارسال 14 آبان

سوال 1: (5 نمره)

در مورد Estimation Likelihood Maximum تحقیق کنید و در دو پاراگراف به اختصار توضیح دهید.

سوال 2: (20 نمره)

فرض کنید مقادیر  $x_1, x_2, \dots, x_n$  از یک توزیع احتمال با چگالی زیر نمونه برداری شدند. با استفاده از Log-Likelihood و MLE پارامتر  $\theta$  را بر حسب  $n$  پیدا کنید.

$$f(x|\theta) = \begin{cases} \frac{-\theta x_i^{\theta-1}}{2^\theta}, & x < 0 \\ 0, & x \geq 0 \end{cases}$$


سوال 3: (20 نمره)

فردی 16 بار تاس انداخته و نتیجه به این شرح است: 15 بار عدد 1 و تنها 1 بار عدد 0 ظاهر شده است. این تاس یک تاس غیرمنصفانه است که احتمال آمدن عدد 1 با مقدار  $p$  تعیین می‌شود. با توجه به این داده‌ها:

الف) تخمین بزنید که احتمال آمدن عدد 1 با استفاده از روش Maximum Likelihood چقدر است؟ توضیح دهید که چرا این روش در این مثال استفاده می‌شود و چگونه می‌توان نتیجه را بر اساس آن تفسیر کرد.

ب) اگر داده‌های شما فقط شامل 8 بار پرتاب بود و همچنان 7 بار عدد 1 و 1 بار عدد 0 ظاهر می‌شد، تخمین شما برای احتمال  $p$  چگونه تغییر می‌کرد؟

مقایسه‌ای بین این دو مجموعه داده (16 بار و 8 بار) انجام دهید و در مورد تأثیر تعداد داده‌های مشاهده شده بر روی دقت تخمین صحبت کنید.



سوال عملی 1: (10 نمره)

در این سوال دیتای sample.csv در اختیار شما قرار داده شده است. بررسی های مورد نیاز روی دیتا را با توجه به نوت بوک انجام دهید.

سوال عملی 2: (10 نمره)

بر اساس توضیحات داده شده در نوت بوک و دیتای snapp.csv به مسئله مطرح شده پاسخ دهید. مسئله مطرح شده anomaly detection است که با استفاده از z-score حل می شود.

سوال عملی 3: (35 نمره)

طراحی سیستم پیشنهاد کتاب با استفاده از Collaborative Filtering

شما به عنوان یک داده کاو در یک فروشگاه آنلاین کتاب استخدام شده اید و از شما خواسته شده است تا یک سیستم پیشنهاد کتاب طراحی کنید. برای این کار، از دیتاست زیر استفاده کنید:

دیتاست:

یک فایل CSV به نام book\_ratings.csv با اطلاعات زیر:

user\_id: شناسه کاربر

book\_id: شناسه کتاب

rating: امتیاز کاربر به کتاب (از ۱ تا ۵)

وظایف:

بارگذاری و آماده سازی داده ها:

با استفاده از کتابخانه های pandas و numpy داده ها را بارگذاری کنید و یک ماتریس امتیازدهی بسازید.

تحلیل داده ها:

یک تحلیل ابتدایی از داده ها انجام دهید (تعداد کاربران، تعداد کتاب ها، توزیع امتیازها و غیره)



پیاده‌سازی: Collaborative Filtering

روش: User-Based

یک تابع بنویسید که شباهت بین کاربران را محاسبه کند (با استفاده از cosine similarity یا Pearson correlation).

با استفاده از شباهت‌ها، برای یک کاربر خاص، ۵ کتاب با بالاترین پیش‌بینی امتیاز را پیشنهاد دهید.

گزارش:

نتایج و تحلیل‌های خود را در یک نوت‌بوک مستند کنید و توضیح دهید که کدام روش بهتر عمل کرده و دلایل آن چیست.