گزارش تکلیف دوم درس الگوریتم های علوم داده (بخش اول)

نام و نام خانوادگی: فاطمه ترودی

شماره دانشجویی: ۴۰۳۴۲۲۰۴۸

نام استاد: دکتر سعیدرضا خردپیشه

نیمسال دوم ۰۴-۱۴۰۳

فهرست مطالب

۱. سوالات تئوری..... ۳

۱.۱. سوال ۱..... ۳

۱.۲. سوال ۲..... ۳

۱.۳. سوال ۳..... ۴

۱.۴. سوال ۴..... ۴

۱.۵. سوال ۵..... ۵

۱.۶. سوال ۶..... ۵

۱.۷. سوال ۷..... ۵

۱.۸. سوال ۸..... ۶

۱.۹. سوال ۹..... ۶

۱. سوالات تئوری

۱/۱. سوال ۱

همبستگی زمانی است که دو متغیر با هم حرکت می کنند اما یکی لزوما باعث دیگری نمی شود. علیت بیانگر این است که یک متغیر مستقیماً بر دیگری تأثیر می گذارد.

مثال: در مجموعه داده YouTube، ممکن است یک همبستگی مثبت بین طول ویدیو و بازدیدها پیدا کنیم. اما این بدان معنا نیست که ویدیوهای طولانی‌تر باعث بازدید بیشتر می‌شوند - عوامل دیگری مانند کیفیت محتوا یا موضوع ممکن است مسئول باشند. اشتباه همبستگی برای علیت می تواند منجر به نتیجه گیری نادرست شود.

۱/۲. سوال ۲

(الف) مسائل عمده در داده های خام:

- مقادیر از دست رفته: می تواند نتایج را سوگیری کند یا عملکرد مدل را کاهش دهد.
- قالب بندی ناسازگار: به عنوان مثال، قالب های تاریخ، رشته و غیره.
- موارد پرت: تحلیل آماری اریب.
- رکوردهای تکراری: معیارها را افزایش دهد یا افزونگی ایجاد کند.
- نویز: داده های نامربوط که نتایج را تحت تاثیر قرار می‌دهد.

(ب) چهار وظیفه اصلی پیش پردازش داده ها:

پاکسازی داده ها (Data Cleaning)، یکپارچه سازی داده ها (Data Integration)، تبدیل داده ها (Data Transformation) و کاهش داده ها (Data Reduction)

(ج) روش‌های رسیدگی به مقادیر گم‌شده:

- سطرها/ستون های دارای داده های از دست رفته را حذف کنیم.
- وارد کردن با استفاده از میانگین/میانه/مد

- مقادیر گمشده را با استفاده از مدل ها پیش بینی کنیم.
- از دسته "ناشناخته" استفاده کنیم (برای متغیرهای رسته ای)

۱/۳. سوال ۳

Binning روشی برای حل مشکل داده های نویز با گروه بندی مقادیر در bin ها و جایگزینی آنها با یک مقدار معرف (مانند میانگین یا میانه) است. مثال: مدت زمان ویدئو را به bin ها تبدیل کنیم:

۲-۰ دقیقه: "کوتاه"

۱۰-۲ دقیقه: "متوسط"

۱۰+ دقیقه: "طولانی"

این کمک می کند تا الگوها را بدون اینکه تحت تأثیر عوامل پرت قرار بگیرند آشکار شوند.

۱/۴. سوال ۴

الف) اهمیت کیفیت داده ها:

داده های ضعیف منجر به نتیجه گیری نادرست می شود. مسائلی مانند نقاط پرت، ناسازگاری ها، یا مقادیر تکراری قابلیت اطمینان را کاهش می دهند.

(ب) سناریو: اگر ویدیویی با ۱ میلیارد بازدید به اشتباه به عنوان دارای ۱ میلیون ثبت شود، ممکن است تحلیل به اشتباه روند نزولی را در ویدیوهای محبوب نشان دهد.

(ج) تکنیک های EDA برای شناسایی مسائل:

- خلاصه های آماری (Summary Statistics)
- نمودارهای جعبه ای برای نقاط پرت

- هیستوگرام ها و نمودارهای توزیع
- ماتریس های همبستگی برای تشخیص ناهنجاری ها

۱/۵. سوال ۵

نرمال سازی داده های عددی را تا محدوده استاندارد، اغلب [۱، ۰] یا با میانگین ۰ مقیاس می کند. دلیل اهمیت اینکار این است که ویژگی ها را در مقیاس قابل مقایسه می کند و عملکرد و همگرایی الگوریتم را بهبود می بخشد. سه روش عبارتند از مقیاس بندی Min-Max، استانداردسازی و Decimal Scaling

۱/۶. سوال ۶

هدف: کاهش اندازه داده ها با حفظ الگوهای کلیدی برای تصویرسازی ساده تر و ...

تکنیک ها:

- کاهش ابعاد (PCA، t-SNE)
- تجمیع (Aggregation)
- نمونه برداری (Sampling)
- انتخاب ویژگی (Feature Selection)

۱/۷. سوال ۷

(الف) تجسم داده ها به عنوان ابزار داستان سرایی:

- درک داده های پیچیده را آسان می کند.
- به برجسته کردن روندها، الگوها و ناهنجاری ها کمک می کند.

- مخاطب را از نظر عاطفی و شناختی درگیر می کند.

(ب) تجسم Traditional در مقابل داستان سرایی:

Storytelling	Traditional
نمودار های پویا و تعاملی	نمودار های ایستا
تفسیر ساده	تفسیر سخت

مثال: نمودار میله ای ساده در مقابل نمودار خطی با حاشیه نویسی که روند رشد ترندها را پس از رویدادها نشان می دهد (مانند همکاری ها یا زمان انتشار). عناصر طراحی که داستان سرایی را تقویت می کند:

عنوان و زیرنویس، حاشیه نویسی ها، هایلایت (رنگ، فوکوس)، Clear Axes and legends

۱/۸. سوال ۸

(الف) عوامل انتخاب نوع نمودار:

- نوع داده (رسته‌ای در مقابل عددی)
- تعداد متغیرها
- روابط برای نشان دادن (روندها، توزیع ها، مقایسه ها)

(ب) نمودارهای توزیع در EDA: به درک پراکندگی، چولگی و وجود نقاط پرت کمک می‌کند.

(ج) نقشه حرارتی ماتریس همبستگی:

قدرت و جهت روابط بین چندین متغیر را نشان می دهد. به شناسایی متغیرهای چند خطی و پیش بینی کننده کمک می کند.

۱/۹. سوال ۹

نمودارهای میله ای برای مقایسه دسته های مختلف مناسب هستند. آنها تشخیص تفاوت‌ها در گروه‌ها را آسان می‌کنند، مانند مقایسه تعداد بازدیدها در دسته‌های مختلف ویدئو در YouTube.

نمودارهای خطی برای تجسم روندها در طول زمان ایده آل هستند. آنها به درک اینکه چگونه معیارهای خاصی - مانند تعامل یا بازدیدها - در طول روز، هفته یا ماه تغییر می کنند کمک می کنند.

نمودارهای دایره ای هنگام نمایش روابط جزئی به کل، مانند نسبت ویدیوهای متعلق به هر دسته، مفید هستند. با این حال، آنها معمولاً در EDA در مقایسه با نمودارهای میله ای یا خطی کمتر مؤثر هستند زیرا وقتی دسته بندی های زیادی یا تفاوت های ظریف در مقادیر وجود دارد تفسیر آنها دشوار است.