پروژه ۶ هوش مصنوعی

حسام رمضانیان

۸۱۰۱۰۰۲۴۸

۱- در صورتی که داده ها نامتوازن بودند، چه مشکالتی در فرآیند خوشه بندی پیش می آمد؟ چه راهکاری را برای برطرف کردن این مشکل ارائه میدهید؟ توضیح دهید.

در صورتی که تعداد نمونه‌ها در هر کلاس به طور چشمگیری متفاوت باشد، چالش‌های زیادی پیش می‌آید. برای نمونه، مدل به سمت کلاس‌های خاصی سوگیری پیدا می‌کند و کلاس‌های پرنمونه‌تر را بهتر یاد می‌گیرد. در نتیجه، دقت پیش‌بینی برای کلاس‌های کم‌نمونه پایین می‌آید. همچنین مدل روی داده‌های آموزش کلاس‌های پر نمونه overfit می‌شود و توانایی تعمیم‌پذیری مدل کاهش می‌یابد.

برای مقابله با این چالش غیرمتوازن بودن داده‌ها، راهکارهای زیر وجود دارد:

۱) می‌توان الگوریتم‌هایی مثل undersampling و oversampling را برای متوازن‌سازی داده‌ها به کار برد.

۲) می‌توان با اعمال تکنیک‌هایی مانند بازنمونه‌برداری، تعداد نمونه‌های هر کلاس را برابر کرد. همچنین به جای تمرکز صرف بر روی معیارهایی مثل دقت کلی، از معیارهایی مانند F1-score استفاده نمود که به کلاس‌های کم‌نمونه‌تر هم توجه دارند.

توضیح تاثیر پیش‌پردازش ها :

با حذف نکردن اعداد فارسی مدل عملکرد ضیعف تری داشت اما حذف کردن اعداد انگلیسی باعث از دست رفتن داده های مهمی در دسته های مختلف میشد و مدل بدتر عمل میکرد ، همچنین استفاده از نرمالایز parsivar باعث می‌شد برخی کلمات به یکدیگر متصل شوند که باعث خرابی

دیتا می‌شد برخی از کاراکتر ها مانند نقطه در انتهای جملات نیز باعث خرابی دیتا می‌شد و حذف آن ها تاثیر چشم‌گیری در نتیجه مدل داشت.

۲ – در گزارش کار خود، جایگزین کردن کلمات با روش stemming یا lemmatization را توضیح دهید.

دو روش رایج برای ساده‌سازی کلمات، ریشه‌یابی(stemming) و پایه‌یابی(lemmatization) هستند.

ریشه‌یابی با حذف پسوند کلمات، آن‌ها را به ریشه‌شان تبدیل می‌کند. البته گاه نتیجه دقیق نیست. مثلاً کلمات "رفتم"، "رفتی" و "رفت" را به "رف" تبدیل می‌کند.

اما پایه‌یابی پیچیده‌تر است و با تجزیه‌ی ساختار کلمه، پایه یا مصدر دقیق آن را مشخص می‌کند. برای این کار به دانش زبان‌شناسی نیاز دارد. برای نمونه، کلمات بالا را به "رفتن" تبدیل می‌کند، چون "رفتن" مصدر و پایه‌ی این سه کلمه است.

مزیت پایه‌یابی، دقت بالاتر است. اما ریشه‌یابی سریع‌تر است و منابع کمتری می‌طلبد. بنابراین هرکدام بسته به نیاز، کاربرد دارند.

۳ – دلیل استفاده از بردار ویژگی و ویژگی های آن را در گزارش توضیح دهید.

بردارهای ویژگی ابزاری قدرتمند برای نمایش و درک داده‌ها هستند. این بردارها اطلاعات کمّی، عددی هر نمونه داده را در قالب مجموعه‌ای از اعداد ارائه می‌دهند. این کار پردازش داده‌ها در الگوریتم‌های یادگیری ماشین را ممکن می‌سازد.

علاوه بر این، بردارهای ویژگی با استانداردسازی داده‌های متنی، تصویری و دیگر انواع داده‌ها، امکان مقایسه بین آن‌ها را فراهم می‌کنند. همچنین محاسبه فواصل و شباهت‌ها بین نقاط داده را آسان می‌نمایند.

در نهایت الگوریتم‌ها می‌توانند الگوهای مفیدی را در این فضاهای برداری پیدا کرده و برای دسته‌بندی و پیش‌بینی به کار گیرند. در حالی‌که فشردگی و کارایی داده‌ها نیز حفظ می‌شود.

۴ – در مورد نحوه استفاده از word2vec و doc2vec و تبدیل متن به بردار ویژگی توضیح دهید.

word2vec و doc2vec دو روش کارآمد برای تبدیل متن به بردارهای عددی ویژگی هستند.

word2vec با بررسی الگوهای هم رخدادی(co-occurrence) کلمات در یک متن، برداری عددی برای هر کلمه تولید می‌کند به گونه‌ای که کلمات مشابه و مرتبط، بردارهای نزدیکی داشته باشند.

doc2vec همین کار را برای جمله‌ها و پاراگراف‌ها انجام می‌دهد و بردار ویژگی برای کل متن تولید می‌کند تا بتوان بر اساس آن متون را مقایسه کرد.

این بردارهای عددیِ معنادار، امکان درک محتوا و مفهوم متون را برای الگوریتم‌های یادگیری ماشین فراهم می‌کنند تا بتوانند تحلیل، خلاصه‌سازی، طبقه‌بندی و ... را بهتر انجام دهند.

۵ – در مورد روش های K-means و DBSCAN و مزایا و معایب این روش ها نسبت به هم توضیح دهید.

روش K-Means با اینکه ساده و سریع است اما نقاط ضعفی هم دارد. این روش برای خوشه‌بندی داده‌هایی که خوشه‌های آن‌ها شکل کروی دارند عملکرد بهتری نشان می‌دهد، اما زمانی‌که شکل خوشه‌ها نامنظم و غیرکروی باشد، K-Means نمی‌تواند خوشه‌بندی دقیقی انجام دهد. همچنین این روش به تعداد خوشه‌ها وابسته است و کاربر باید از ابتدا تعداد خوشه‌ها را مشخص کند که این می‌تواند محدودیتی باشد.

از طرفی روش DBSCAN نسبت به شکل خوشه‌ها انعطاف‌پذیرتر است و می‌تواند خوشه‌های با اشکال مختلف را تشخیص دهد. همچنین این روش برای شناسایی نقاط پرت و دورافتاده مناسب

است. اما DBSCAN در مقابل حجم بالای داده‌ها عملکرد ضعیف‌تری دارد و با افزایش حجم داده‌ها سرعت آن کاهش می‌یابد.

بنابراین می‌توان نتیجه گرفت که هر دو روش مزایا و معایب خاص خود را دارند و انتخاب روش مناسب بستگی به نوع داده‌ها و کاربرد خوشه‌بندی دارد. K-Means برای داده‌های بزرگ و خوشه‌های کروی‌شکل مناسب‌تر است در حالی‌که DBSCAN انعطاف‌پذیری بیشتری در شناسایی خوشه‌های نامنظم دارد.

۶ – خروجی حاصل از دو نوع خوشه بندی را با هم مقایسه کنید.

الگوریتم K-Means با توجه به معیار Homogeneity Score بالاتر، خوشه‌بندی بهتر و همگن‌تری نسبت به برچسب‌های واقعی داده‌ها انجام داده است. همچنین K-Means با داشتن Silhouette Score بالاتر، خوشه‌هایی با تفکیک‌پذیری درونی و بین خوشه‌ای بهتر تولید کرده است.

اما DBSCAN با توجه به ماهیت آن مبنی بر شناسایی نواحی پرتراکم به عنوان خوشه، خروجی متفاوتی ارائه می‌دهد. این الگوریتم خوشه‌ها را بر اساس تراکم تشکیل می‌دهد و ممکن است خوشه‌های نامنظم و غیرهمگن‌تری تولید کند.

بنابراین می‌توان گفت هر دو الگوریتم بسته به نوع داده‌ها و هدف از خوشه‌بندی، می‌توانند انتخاب‌های مناسبی باشند. K-Means برای داده‌هایی که خوشه‌بندی کروی‌شکل مطلوب است، مناسب‌تر است مانند پروژه ما اما DBSCAN می‌تواند خوشه‌های پیچیده‌تر و غیراستاندارد را شناسایی کند.

۷ – درباره PCA تحقیق کنید و نحوه عملکرد آن را به اختصار توضیح دهید.

PCA یک روش کاهش ابعاد در داده‌های چند متغیره است که به شرح زیر عمل می‌کند:

- ابتدا داده‌ها را استانداردسازی می‌کند تا میانگین آن‌ها صفر و انحراف معیارشان یک شود. این کار باعث می‌شود تمام متغیرها در یک مقیاس قرار بگیرند.

- سپس ماتریس همبستگی بین متغیرها را محاسبه می‌کند تا رابطه آن‌ها مشخص شود. متغیرهایی که همبستگی بالایی دارند، اطلاعات مشابهی را نشان می‌دهند.

- PCA متغیرهایی را که همبستگی بالایی دارند در یک مؤلفه اصلی خلاصه می‌کند. این مؤلفه‌های اصلی جدید، حاوی بیشترین اطلاعات داده‌ها هستند.

- در نهایت PCA تعدادی از مؤلفه‌های اصلی را که بیشترین واریانس داده‌ها را توضیح می‌دهند، انتخاب می‌کند تا ابعاد داده‌ها کاهش یابد.

- داده‌های اصلی در فضای کاهش یافته‌ی جدید نمایش داده می‌شوند.

درواقع PCA با حفظ بیشترین اطلاعات، تعداد ابعاد داده‌ها را کاهش می‌دهد و باعث ساده‌سازی داده‌ها می‌شود.

۸- در مورد نحوه محاسبه معیار silhouette و homogeneity توضیح دهید.

Silhouette Score به این صورت محاسبه می‌شود:

a: میانگین فاصله هر نقطه تا سایر نقاط همان خوشه

b: کمترین میانگین فاصله هر نقطه تا نقاط خوشه دیگر

$$s = (b - a) / \max(a, b)$$

سپس میانگین s برای تمام نقاط محاسبه می‌شود. مقادیر بالاتر از صفر نشان دهنده خوشه‌بندی بهتر است.

Homogeneity Score نشان دهنده درصد نقاطی است که درست خوشه‌بندی شده‌اند:

$C = \{c1, c2, ..., cn\}$ برچسب‌های واقعی

$K = \{k1, k2, ..., kn\}$ برچسب‌های پیش‌بینی شده

correct = تعداد نقاطی که ci == ki

Homogeneity = (correct /total)

مقادیر بالاتر از صفر و نزدیک یک نشان‌دهنده خوشه‌بندی بهتر است.

بنابراین هر دو معیار به نوعی کیفیت خوشه‌بندی را می‌سنجند، Silhouette Score تفکیک‌پذیری خوشه‌ها و Homogeneity Score صحت خوشه‌بندی را بررسی می‌کند.

۹- نتایج حاصل از معیارهای ذکر شده را برای هر یک از روشها گزارش کنید.

در فایل report.html موجود است

۱۰- راهکارهایی پیشنهاد کنید که بتوان عملکرد مدلها را بهبود داد.

برای بهبود عملکرد مدل‌های خوشه‌بندی متن، راهکار های زیر وجود دارد :

- بهینه‌سازی پارامترهای الگوریتم‌ها مانند پارامترهای تراکم در DBSCAN برای بهبود نتایج.

- افزایش داده‌های آموزشی برای بهبود یادگیری الگوریتم‌ها و غنی‌تر شدن بردارهای ویژگی.

- پیش‌پردازش بهتر داده‌ها برای حذف نویز و داده‌های پرت.

- استفاده از روش‌های مختلف برای برداری کردن متن مانند TF-IDF و Word2Vec به جای Bag-of-Words ساده. این روش‌ها معانی واژگان را در نظر می‌گیرند.

- بهینه‌سازی پارامترهای مدل‌های برداری کردن متن مانند اندازه بردارها در Doc2Vec. پارامترها باید متناسب با حجم داده‌ها تنظیم شوند.

- کاهش ابعاد داده‌ها قبل از خوشه‌بندی با استفاده از PCA. ابعاد پایین‌تر باعث بهبود عملکرد می‌شود.

به طور کلی باید پارامترها و الگوریتم‌ها را بر اساس داده‌های مورد استفاده بهینه کرد.