به نام خدا

گزارش تمرین ۵

ائلدار صمدزاده طریقت ۴۰۰۳۳۰۴۴

1. اگر داده‌ها غیرکروی یا دارای پراکندگی غیر یکنواخت باشند، K-Means نمی‌تواند خوشه‌بندی درستی ارائه دهد چون این الگوریتم بر اساس فاصله اقلیدسی و میانگین نقاط عمل می‌کند و فرض می‌کند خوشه‌ها شکل کروی و هم‌پراکندگی دارند. در چنین حالتی، K-Means ممکن است خوشه‌هایی ایجاد کند که در مرزها یا مرکزهای واقعی خوشه‌ها هم‌پوشانی نداشته باشند. راه‌حل: استفاده از الگوریتم‌هایی مانند DBSCAN یا Gaussian Mixture Model که توانایی تشخیص خوشه‌های با اشکال و چگالی مختلف را دارند.

2. K-Means ممکن است به مینیمم محلی برسد چون مقداردهی اولیه مراکز خوشه‌ها به صورت تصادفی انجام می‌شود. برای کاهش این مشکل می‌توان از K-Means++ برای مقداردهی اولیه استفاده کرد که مراکز اولیه را به گونه‌ای انتخاب می‌کند که پراکندگی بیشتری داشته باشند. اگر کد ضمیمه را با دو مقداردهی اولیه متفاوت اجرا کنید، نتایج نهایی (مراکز و برچسب‌ها) ممکن است متفاوت باشند چون تابع هزینه K-Means غیر محدب است و چندین مینیمم محلی دارد. این پدیده به ساختار تابع هزینه (sum of squared distances) برمی‌گردد که نسبت به مقداردهی اولیه حساس است.

3. افزایش تعداد خوشه‌ها همیشه منجر به بهبود واقعی عملکرد نمی‌شود، بلکه صرفاً خطای درون‌خوشه‌ای را کاهش می‌دهد .(overfitting) برای انتخاب مقدار بهینه K می‌توان از روش‌هایی مانند Elbow Method ، Silhouette Score یا Gap Statistic استفاده کرد.

4. اگر یکی از خوشه‌ها بسیار بزرگ‌تر از دیگری باشد، K-Means تمایل دارد که خوشه بزرگ‌تر را بهتر مدل کند و ممکن است خوشه کوچک‌تر را نادیده بگیرد یا با دیگری ادغام کند. چون الگوریتم برای کمینه کردن مجموع مربعات فاصله‌ها طراحی شده و نه برای توزیع متعادل نمونه‌ها در خوشه‌ها.

5. در K-Means مراکز خوشه‌ها با میانگین به‌روزرسانی می‌شوند چون میانگین، مقدار مینیمم مجموع مربعات فاصله‌ها (L2-norm) را فراهم می‌کند. اگر به‌جای میانگین از میانه استفاده شود، به الگوریتمی مشابه K-Medoids یا K-Medians می‌رسیم که در برابر نقاط پرت مقاوم‌تر است، اما پیچیدگی محاسباتی بیشتری دارد.

6. فرض استفاده از فاصله اقلیدسی معتبر است اگر ویژگی‌ها مستقل، عددی و هم‌مقیاس باشند. این فرض در شرایطی مثل داده‌های با مقیاس‌های متفاوت، داده‌های جهتی یا داده‌های دارای وابستگی ساختاری بین ویژگی‌ها (مانند داده‌های زمانی یا فضایی) نادرست می‌شود.

7. اگر ویژگی‌ها مقیاس‌های متفاوتی داشته باشند، ویژگی‌هایی با دامنه بزرگ‌تر تأثیر بیشتری در خوشه‌بندی خواهند داشت. برای حل این مشکل باید داده‌ها را نرمال‌سازی یا استانداردسازی کرد) مانند استفاده از StandardScaler یا MinMaxScaler در.(sklearn

8. K-Meansبرای داده‌های گسسته یا طبقه‌بندی شده مناسب نیست چون محاسبه میانگین برای ویژگی‌های گسسته بی‌معنی است. الگوریتم جایگزین مناسب برای داده‌های طبقه‌ای، K-Modesیا K-Prototypes (برای داده‌های ترکیبی) است که از معیارهایی مانند تطابق یا Gower distance استفاده می‌کنند.

9. K-Meansرا می‌توان برای کاهش بعد نیز استفاده کرد. پس از خوشه‌بندی، می‌توان هر نقطه داده را با مرکز خوشه متناظر جایگزین کرد یا از فواصل آن نسبت به مراکز خوشه‌ها به عنوان ویژگی‌های جدید استفاده کرد. همچنین، در برخی روش‌ها مانندspectral clustering ، از K-Means بر روی فضای برداری کاهش‌یافته استفاده می‌شود.

10. اگر مقدار K بیشتر از تعداد واقعی خوشه‌ها انتخاب شود، K-Means نمی‌تواند به‌صورت خودکار این خطا را اصلاح کند چون همیشه دقیقاً K خوشه را تولید می‌کند. در نتیجه، برخی خوشه‌ها ممکن است بسیار کوچک یا خالی شوند یا نمونه‌هایی را به‌طور تصادفی دسته‌بندی کنند. چون الگوریتم بدون دانش قبلی از ساختار واقعی داده، تنها به بهینه‌سازی تابع هزینه متکی است.