زهرا گروسی

تمرین جلسه دوم

درس آمار در نوروالکتروفیزیولوژی – Local Outlier Factor

الگوریتم عامل دور افتاده محلی یا Local Outlier Factor (LOF) در سال ۲۰۰۰ برای کشف ناهنجاری‌ها و داده‌های پرت موجود در نقاط داده با اندازه‌گیری انحراف محلی یک نقطه داده با توجه به همسایه‌های آن ارائه شد. این الگوریتم(LOF) ، در برگیرنده برخی مفاهیم موجود در الگوریتم‌های DBSCAN و OPTICS مانند فاصله از مرکز (core distance) و فاصله دسترسی‌پذیری (reachability distance) است که برای تخمین چگالی محلی مورد استفاده قرار می‌گیرند.

الگوریتم عامل دورافتاده محلی بر مبنای مفهوم چگالی محلی بنا شده و در آن محلی بودن بر اساس k نزدیک‌ترین همسایگی تعیین می‌شود که فاصله آن‌ها برای تخمین چگالی مورد استفاده قرار می‌گیرد. با مقایسه چگالی محلی یک شی با چگالی‌های همسایه‌های آن می‌توان نواحی دارای چگالی مشابه و نقاطی که اساسا چگالی کمتری نسبت به همسایه‌های خود دارند را تعیین کرد.

این موارد به عنوان دورافتادگی (داده پرت) در نظر گرفته می‌شوند. چگالی محلی به وسیله فاصله معمولی که یک نقطه داده توسط همسایه‌های خود «دسترسی‌پذیر» است تخمین زده می‌شود.

بیان ریاضی

فرض می‌شود k-distance(P) فاصله شی P از k اُمین نزدیک‌ترین همسایه است. توجه به این نکته لازم است که مجموعه k نزدیک‌ترین همسایه‌ها شامل همه اشیا موجود در این فاصله است که در صورت وقوع «گره» ممکن است بیش از k شی باشند. مجموعه k نزدیک‌ترین همسایگی به صورت $N_k(P)$ نشان داده می‌شود.

مراحل محاسبه

۱. انتخاب پارامتر k: تعداد همسایه‌های مورد بررسی تعیین می‌شود.
۲. محاسبه فاصله k-امین همسایه: برای هر نقطه مانند P فاصله‌ای به نام k-distance(P) محاسبه می‌شود.
۳. فاصله دسترسی پذیری (Reachability Distance): پایدارسازی محاسبات، فاصله بین دو نقطه به صورت اصلاح‌شده تعریف می‌شود.

$$o \in N_k(P)$$
$$RD_k(P,o) = max\big(d(P,o), k-distance(o)\big)$$

۴. چگالی دسترسی پذیری محلی (LRD): برای هر نقطه p ، چگالی محلی با استفاده از فاصله‌های دسترسی پذیری به همسایگان محاسبه می‌شود.

$$LRD_k(p) = \left(\frac{\sum_o RD_k(p,0)}{|N_k(P)|}\right)^{-1}$$

۵. محاسبه LOF: نسبت میانگین چگالی همسایگان به چگالی نقطه مورد نظر محاسبه میگردد.

$$LOF_k(P) = \frac{\sum_o \frac{LRDk(o)}{LRD_k(P)}}{|N_k(P)|}$$

تفسیر مقدار LOF:

اگر $LOF_k(p) \approx 1$ آنگاه چگالی p مشابه همسایه‌هاست.

اگر $LOF_k(p) > 1$ آنگاه چگالی p کمتر از همسایه‌هاست.

هرچه مقدار LOF بزرگتر باشد احتمال پرت بودن قوی‌تر است.

مزایا:

ماهیت نسبی: داده‌های پرت تنها در مقایسه با همسایگان مشخص می‌شوند.

کاربرد در داده‌های چند بعدی: بدون نیاز به ساده سازی یا فروض خاص درباره توزیع داده‌ها

انعطاف پذیری: خوشه‌های پرت ولی معتبر را به عنوان داده‌های پرت شناسایی نمی‌کند.

معایب:

حساسیت به انتخاب پارامتر *k*: انتخاب نامناسب می‌تواند موجب کاهش دقت شود.

پیچیدگی محاسباتی: به دلیل نیاز به محاسبه فاصله بین نقاط متعدد، هزینه زمانی بالایی دارد.

محدودیت در داده‌های کم‌تراکم: در صورت یکنواخت بودن کل داده‌ها، کارایی کاهش می‌یابد.