بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

درس : *یادگیری ماشین*

استاد : *خانم دکتر زرین بال ماسوله*

موضوع : *clustering validity index*

دانشجو : *فرزاد محسنی*

مقطع : *کارشناسی ارشد*

رشته تحصیلی : *مهندسی آیتی*

کد دانشجویی :

دانشگاه : *صنعتی امیرکبیر*

بهار 1404

Method ---------- Page

Silhouette Score

امتیاز سیلوئت (Silhouette Score) چیست؟

- Silhouette Score یک معیار برای ارزیابی کیفیت خوشه بندی است. این معیار بررسی می‌کند که هر نقطه داده (Data Point) تا چه حد به خوشه‌ی خودش تعلق دارد و چقدر از خوشه های دیگر جداست.

چه زمانی از Silhouette Score استفاده می‌شود؟

- برای ارزیابی کیفیت خوشه بندی بدون نیاز به برچسب های واقعی (Unsupervised).
- برای مقایسه‌ی تعداد خوشه ها: مثلاً اگر نمی‌دانی چند خوشه مناسب‌تر است، می‌توانی سیلوئت را برای مقادیر مختلف K محاسبه کنی و بهترین را انتخاب کنی.
- در الگوریتم های خوشه بندی مثل K-Means، DBSCAN، Agglomerative Clustering و غیره کاربرد دارد.

مزایا

- بدون نیاز به برچسب (Label-Free)
- ساده و قابل فهم
- امکان بررسی خوشه بندی به‌صورت نقطه به نقطه

معایب

- محاسبه اش می‌تواند برای دیتاست های بزرگ کند باشد.
- به معیار فاصله‌ی مورد استفاده حساس است (مثلاً فاصله اقلیدسی، کسینوسی، و غیره).

فرمول امتیاز سیلوئت

برای هر نقطه داده i ، امتیاز سیلوئت به‌صورت زیر محاسبه می‌شود:

$$s(i) = \frac{b(i) - a(i)}{\max(\ a(i),\ b(i)\ )}$$

که در آن:

- a(i): میانگین فاصله‌ی نقطه i تا تمام نقاط در همان خوشه (درونی بودن خوشه).
- b(i): کمترین میانگین فاصله‌ی نقطه i تا نقاط خوشه‌ی نزدیک دیگر (بیرونی بودن خوشه).

مقدار امتیاز سیلوئت

- s(i) همیشه بین ۱- تا ۱+ است:
  - اگر نزدیک به ۱ باشد: یعنی نقطه به خوشه‌ی خودش بسیار خوب تعلق دارد.
  - اگر نزدیک به ۰ باشد: یعنی نقطه در مرز بین دو خوشه است.
  - اگر منفی باشد: یعنی احتمالاً نقطه به خوشه‌ی اشتباهی تخصیص داده شده است.

Calinski-Harabasz Index

شاخص کالینسکی–هاراباز چیست؟

- شاخص Calinski-Harabasz که به آن Variance Ratio Criterion نیز می‌گویند، یک معیار برای ارزیابی کیفیت خوشه بندی است. این شاخص نسبتِ پراکندگی بین خوشه ها (بین‌خوشه‌ای) به پراکندگی درون خوشه ها (درون‌خوشه‌ای) را اندازه‌گیری می‌کند.

چه زمانی استفاده می‌شود؟

- زمانی که می‌خواهی خوشه بندی بدون برچسب واقعی را ارزیابی کنی (unsupervised evaluation).
- برای تعیین تعداد بهینه خوشه ها (با مقایسه‌ی مقدار شاخص برای مقادیر مختلف k).
- کاربرد در الگوریتم هایی مثل K-Means، K-Medoids و الگوریتم های سلسله مراتبی (hierarchical clustering).

مزایا

- محاسبه سریع و مؤثر
- بدون نیاز به داده های برچسب خورده
- قابل تفسیر ساده: مقدار بیشتر = خوشه بندی بهتر

معایب

- به شکل خوشه ها و اندازه‌ی داده ها حساس است
- فرض می‌کند خوشه ها شکل کروی و اندازه مشابهی دارند

فرمول شاخص کالینسکی-هارابازش

$$\mathrm{CH} = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \cdot \frac{N-k}{k-1}$$

که در آن:

- $\mathrm{Tr}(B_k)$ مجموع مربعات بین خوشه ها (Between-cluster dispersion)
- $\mathrm{Tr}(W_k)$ مجموع مربعات درون خوشه ها (Within-cluster dispersion)
- **N**: تعداد کل نمونه ها
- **k**: تعداد خوشه ها

تفسیر شاخص

- مقدار بالاتر بهتر است.
- یعنی خوشه ها به خوبی از هم جدا هستند و داده ها درون هر خوشه فشرده و نزدیک به هم اند.
- شاخص CH مقدار منفی ندارد و می‌تواند از صفر تا بی‌نهایت باشد.
- اگر خوشه ها با هم ادغام یا پراکنده باشند، مقدار شاخص کاهش می‌یابد.

Davies-Bouldin Index

شاخص دیویس-بولدین چیست؟

- Davies-Bouldin Index یا DBI یکی دیگر از معیارهای ارزیابی کیفیت خوشه بندی است. این شاخص بررسی می‌کند که خوشه ها چقدر از هم جدا هستند و چقدر درون خود فشرده اند.

چه زمانی استفاده می‌شود؟

- برای ارزیابی کیفیت خوشه بندی در حالت بدون نظارت (بدون نیاز به برچسب ها).
- برای مقایسه خوشه بندی ها با تعداد مختلف خوشه و انتخاب بهترین k.
- کاربرد در الگوریتم هایی مثل K-Means، DBSCAN، Agglomerative Clustering و غیره.

مزایا

- مستقل از برچسب
- محاسبه ی نسبتاً ساده
- ترکیب خوبی از فشردگی داخلی و جدایی بین خوشه ها را می‌سنجد

معایب

- به شکل و توزیع خوشه‌ها حساس است
- اگر یک خوشه بسیار بزرگ یا پراکنده باشد، می‌تواند مقدار کل شاخص را خراب کند

فرمول شاخص دیویس-بولدین

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{s_i + s_j}{d_{ij}}\right)$$

که در آن:

- **k**: تعداد خوشه ها
- s_i میانگین فاصله نقاط در خوشه i تا مرکز خوشه (درون‌خوشه ای یا فشردگی خوشه)
- d_{ij} فاصله بین مراکز خوشه های i و j

تفسیر شاخص DBI

- مقدار کمتر بهتر است
- چون نشان می دهد خوشه ها از یکدیگر جدا هستند و داده های هر خوشه نزدیک به هم هستند.
- مقدار DBI همیشه بزرگ تر یا مساوی صفر است.
- در حالت ایده آل، هر خوشه باید کمترین هم پوشانی را با خوشه های دیگر داشته باشد.

Adjusted Rand Index

شاخص Adjusted Rand Index (ARI) چیست؟

- شاخص رند تعدیل شده معیاری برای ارزیابی کیفیت خوشه بندی است در صورتی که برچسب های واقعی (Ground Truth Labels) در دسترس باشند.
- این شاخص بررسی می‌کند که چقدر خوشه بندی به‌دست آمده با برچسب های واقعی هماهنگ و منطبق است.

تفاوت با Rand Index معمولی؟

- شاخص رند ساده فقط میزان تطابق بین دو دسته بندی را بررسی می‌کند، ولی ARI تعدیل شده برای تطابق تصادفی است.
- یعنی اگر دو دسته بندی کاملاً تصادفی باشند، ARI نزدیک به صفر خواهد بود، نه مقدار بالا.

چه زمانی استفاده می‌شود؟

- فقط زمانی که برچسب های واقعی داده ها موجود باشند (برای ارزیابی دقیق).
- مقایسه‌ی عملکرد الگوریتم های خوشه‌بندی در داده های labeled.
- بسیار پرکاربرد در ارزیابی K-Means، Agglomerative Clustering و سایر الگوریتم ها با ground truth.

مزایا

- دقت بالا در مقایسه با معیار های ساده
- تعدیل شده برای جلوگیری از نتایج گمراه کننده در حالت تصادفی
- قابل اعتماد برای مقایسه با ground truth

معایب

- فقط در صورت وجود برچسب های واقعی قابل استفاده است
- ممکن است محاسبات برای داده های بسیار بزرگ کمی کند شود

فرمول کلی

$$\text{ARI} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}$$

جزئیات کامل فرمول با استفاده از ماتریس احتمال هم قابل نمایش است، اما ایده‌ی اصلی این است:

- ARI = 1 ⟶ خوشه بندی کاملاً مطابق با برچسب های واقعی
- ARI ≈ 0 ⟶ تطابقی بیش‌تر از حالت تصادفی ندارد
- ARI < 0 ⟶ بدتر از تصادف (نادر)

ویژگی ها

- دامنه: از -1 تا 1
- ۱: تطابق کامل
- ۰: تصادفی
- کمتر از صفر: بدتر از حالت تصادفی

Normalized Mutual Information – NMI

NMI چیست؟

- (NMI) یک معیار برای مقایسه دو خوشه بندی است، معمولاً بین خوشه بندی حاصل از الگوریتم و برچسب های واقعی (Ground Truth).
- این معیار می سنجد که چه مقدار اطلاعات بین دو دسته بندی مشترک است.

ایده‌ی اصلی:

- اطلاعات متقابل (Mutual Information - MI) میزان اشتراک اطلاعات بین دو متغیر تصادفی (در اینجا: دو خوشه بندی) را اندازه می‌گیرد.
- اما MI به تعداد نمونه ها و اندازه‌ی خوشه ها حساس است. برای رفع این مشکل، آن را نرمال سازی (Normalization) می‌کنند.

چه زمانی استفاده می‌شود؟

- زمانی که برچسب های واقعی در دسترس هستند
- برای مقایسه‌ی خوشه بندی ها با ground truth یا حتی بین دو الگوریتم خوشه بندی
- پرکاربرد در ارزیابی K-Means، Spectral Clustering و سایر الگوریتم ها

مزایا

- مقیاس بندی شده و قابل مقایسه بین داده های مختلف
- برابر است در صورت تعویض برچسب خوشه ها (invariant to label permutations)
- مناسب برای خوشه بندی های با تعداد متفاوت خوشه

معایب

- فقط برای داده های دارای برچسب قابل استفاده است
- نیاز به محاسبه آنتروپی و اطلاعات متقابل (کمی پیچیده تر از برخی معیارها)

فرمول NMI

$$\mathrm{NMI}(U,V) = \frac{2 \cdot I(U;V)}{H(U) + H(V)}$$

که در آن:

- **U**: خوشه بندی پیش بینی شده
- **V**: خوشه بندی واقعی
- **I(U;V)**: اطلاعات متقابل بین
- **H(U) و H(V)**: آنتروپی خوشه بندی ها (میزان عدم قطعیت)

دامنه و تفسیر

- NMI ∋ [0, 1]
- ۱ = تطابق کامل بین خوشه بندی و برچسب های واقعی
- ۰ = هیچ ارتباط اطلاعاتی بین دو دسته بندی وجود ندارد

جدول مقایسه شاخص های اعتبار خوشه بندی

شاخص	نوع ارزیابی	نیاز به برچسب واقعی؟	دامنه مقدار	معیار بهتر بودن	کاربرد اصلی	مزایا	معایب
Silhouette Score	درونی (Internal)	ندارد	1- تا 1+	مقدار بیشتر بهتر	ارزیابی فشردگی درون خوشه‌ای و جدایی بین خوشه‌ای	شهودی، بدون نیاز به برچسب، قابل تفسیر	در دیتاست های بزرگ کند می‌شود
Calinski-Harabasz Index	درونی (Internal)	ندارد	0 تا ∞	مقدار بیشتر بهتر	انتخاب تعداد بهینه خوشه ها	سریع، ساده، بدون نیاز به برچسب	فرض کروی بودن خوشه ها
Davies-Bouldin Index	درونی (Internal)	ندارد	0 تا ∞	مقدار کمتر بهتر	بررسی جدایی و فشردگی خوشه ها	ترکیب فشردگی و جدایی، بدون نیاز به برچسب	به خوشه های پراکنده حساس است
Adjusted Rand Index	بیرونی (External)	دارد	1- تا 1+	مقدار بیشتر بهتر	ارزیابی دقت نسبت به برچسب واقعی	تعدیل شده برای تصادف، دقیق	فقط در صورت داشتن برچسب کاربرد دارد
Normalized Mutual Information - NMI	بیرونی (External)	دارد	0 تا 1	مقدار بیشتر بهتر	مقایسه خوشه بندی با برچسب یا خوشه دیگر	نرمال شده، مستقل از ترتیب برچسب ها	نیاز به محاسبه آنتروپی و MI