

Clustering لتقييم ال Internal vs External Matrices مقارنة

نظرة عامة

Internal Validation Metrics

- باستخدام البيانات المستخدمة في ال Clustering **التعريف**: مقاييس تقييم جودة ال Clustering نفسها
- والانفصال بينها Clusters **المبدأ**: تقييم التماسك داخل ال
- True labels أو Ground truth: لا تحتاج

External Validation Metrics

- مع التصنيف الحقيقي المعروف مسبقاً Clustering **التعريف**: مقاييس تقارن نتائج ال
- الأصلية Classes المكتشفة مع ال Clusters **المبدأ**: قياس مدى توافق ال
- True labels أو Ground truth: تحتاج

المقارنة التفصيلية

الهدف والاستخدام

الجانب	Internal Metrics	External Metrics
الهدف الأساسي	Clustering قياس الجودة الداخلية للـ	مقارنة بالحقيقة Clustering قياس مدى صحة الـ
متى نستخدمها	عندما لا نعرف التصنيف الحقيقي	عندما نعرف التصنيف الحقيقي مسبقاً
نوع التقييم	Unsupervised evaluation	Supervised evaluation
الاعتماد	على خصائص البيانات فقط	على المقارنة مع المرجع الحقيقي

متطلبات الاستخدام

المتطلب	Internal Metrics	External Metrics
Ground Truth	غير مطلوب ❌	مطلوب بشدة ✅
البيانات الأصلية	مطلوبة ✅	مطلوبة ✅
Distance Matrix	مطلوبة عادة ✅	غير مطلوبة ❌
Domain Knowledge	غير مطلوب ❌	مفيد للتفسير ✅

الأنواع والأمثلة

Internal Validation Metrics

1. Silhouette Score

- المدى: $[-1, 1]$
- الأفضل: قيم أعلى (قريبة من 1)
- يقيس: التماسك الداخلي والانفصال الخارجي

2. Calinski-Harabasz Index

- المدى: $(0, \infty)$
- الأفضل: قيم أعلى
- إلى التباين داخلها Clusters يقيس: نسبة التباين بين الـ

3. Davies-Bouldin Index

- المدى: $(0, \infty)$
- الأفضل: قيم أقل
- Clusters يقيس: متوسط التشابه بين الـ

4. Dunn Index

- المدى: $(0, \infty)$
- الأفضل: قيم أعلى
- cluster إلى أكبر مسافة داخل clusters يقيس: نسبة أقل مسافة بين

External Validation Metrics

1. Adjusted Rand Index (ARI)

- المدى: $[-1, 1]$
- الأفضل: قيم أعلى (1 = perfect match)
- والتصنيف الحقيقي Clustering يقيس: التشابه بين الـ

2. Adjusted Mutual Information (AMI)

- المدى: $[0, 1]$
- الأفضل: قيم أعلى (1 = perfect match)
- يقيس: المعلومات المتبادلة المعدلة

3. Fowlkes-Mallows Index

- المدى: [0, 1]
- الأفضل: قيم أعلى
- يقيس: المتوسط الهندسي للدقة والاستدعاء

4. V-Measure

- المدى: [0, 1]
- الأفضل: قيم أعلى
- Homogeneity و Completeness يقيس: المتوسط التوافقي لـ

المزايا والعيوب

Internal Metrics

المزايا

- يمكن استخدامها في أي موقف: **Ground Truth** لا تحتاج
- سرعة الحساب: عادة أسرع في التنفيذ
- تقييم هيكلي: تركّز على الهيكل الداخلي للبيانات
- عملية: مناسبة للتطبيقات الحقيقية

العيوب

- مفيد للمهمة Clustering قد تكون مضللة: لا تضمن أن الـ
- معينة clusters حساسة للشكل: قد تفضل أشكال
- لا تقيس الصحة: تقيس الجودة وليس الصحة
- متحيزة للخوارزمية: قد تفضل خوارزميات معينة

External Metrics

المزايا

- تقييم حقيقي: تقارن بالحقيقة المعروفة
- موثوقة: أكثر موثوقية في قياس الأداء
- مستقلة عن الخوارزمية: لا تتحيز لخوارزمية معينة
- مفهومة: أسهل في التفسير والفهم

العيوب

- غير متاحة دائماً: **Ground Truth** تحتاج

- قد يكون مكلف Labels مكلّفة: الحصول على
- قد تكون خاطئة Ground Truth قد تكون غير كاملة: ال
- محدودة الاستخدام: فقط عند توفر التصنيف الحقيقي

متى نستخدم كل نوع؟ 🎯

عندما Internal Metrics استخدم:

- لا تعرف التصنيف الحقيقي ?
- تستكشف البيانات لأول مرة 🔍
- تريد تقييم سريع ⚡
- تقارن بين خوارزميات مختلفة 🔄
- الأمثل Clusters تحدد عدد ال 📊

عندما External Metrics استخدم:

- تعرف التصنيف الحقيقي ✅
- Clustering تريد تقييم دقة ال 🎯
- تقييم أداء نموذج معين 📄
- تجري دراسة علمية 🧪
- معروف baseline تقارن مع 📊

استراتيجية التقييم المثلى 🔄

نهج متكامل:

1. لفهم هيكل البيانات Internal Metrics ابدأ ب.
2. للتحقق من الصحة (إن أمكن) External Metrics استخدم.
3. اجمع بين عدة مقاييس لتقييم شامل.
4. فسر النتائج في سياق المشكلة الأصلية.

مثال عملي:

1. Clusters لتحديد عدد ال Silhouette Score استخدم.
2. Ground Truth للمقارنة مع ARI استخدم.
3. لتقييم جودة الانفصال Davies-Bouldin استخدم.
4. لقياس المعلومات المحفوظة AMI استخدم.



الخلاصة

Internal Metrics

- الأنسب لـ: الاستكشاف الأولي وتحسين الخوارزميات
- القوة: العملية والسرعة
- الضعف: قد لا تعكس الفائدة الحقيقية

External Metrics

- الأنسب لـ: التحقق من الصحة والتقييم النهائي
- القوة: الموثوقية والوضوح
- Ground Truth الضعف: تحتاج

التوصية النهائية:

للتحقق والتقييم النهائي External Metrics للاستكشاف والتطوير، و Internal Metrics استخدم عندما تكون متاحة.