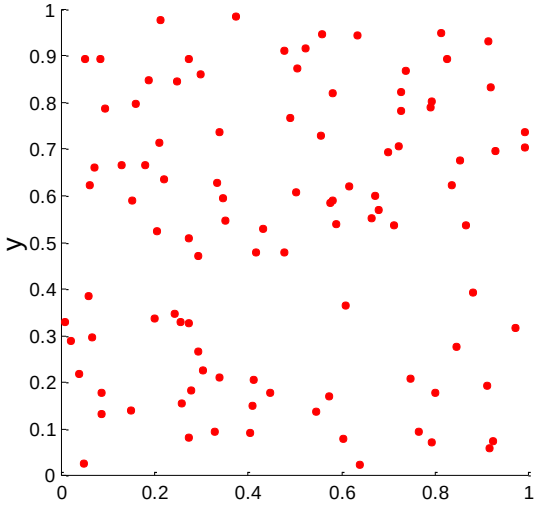
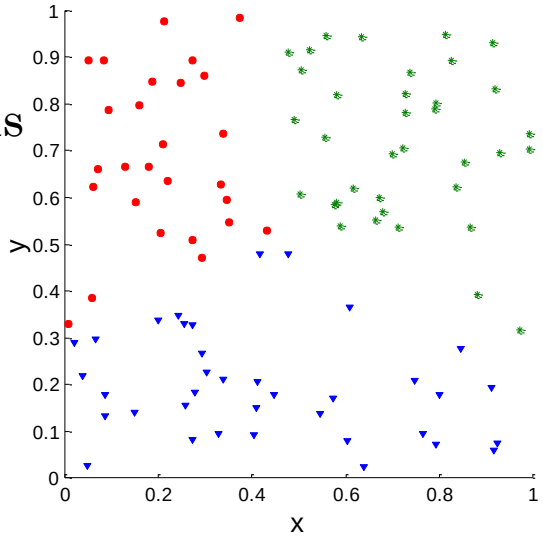
اعتبارسنجی (ارزیابی) خوشه‌بندی
Clustering Validation

خوشه‌بندی ایجادشده بر روی داده‌های تصادفی

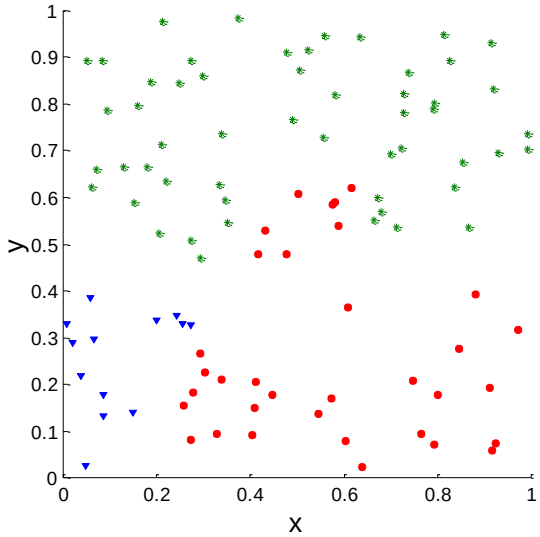
Random Points



K-means



Complete Link



جنبه‌های مختلف اعتبارسنجی خوشه‌بندی

- تعیین قابلیت خوشه‌بندی – آیا داده‌های مورد آزمایش ساختار غیرتصادفی دارند؟
- مقایسه نتایج خوشه‌بندی با اطلاعات اضافه‌ای (نظیر برچسب داده‌ها) که در دسترس است.
- ارزیابی نتایج خوشه‌بندی بدون داشتن اطلاعات اضافه
- مقایسه دو روش خوشه‌بندی
- تعیین تعداد خوشه

روش‌های اعتبار سنجی

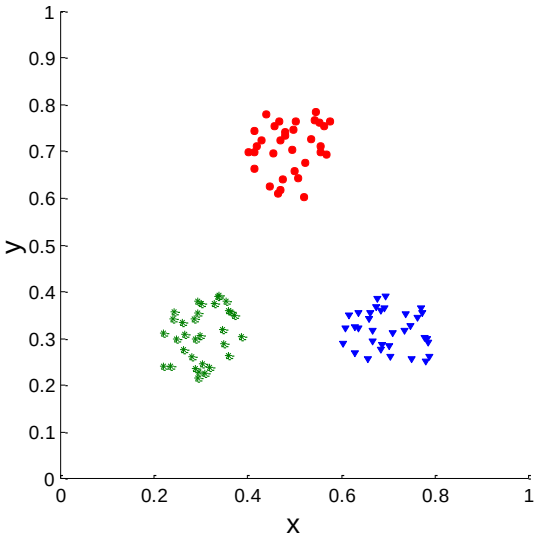
- روش اول: اطلاعات جانبی نوع کلاس داده در دسترس است(External Index).
  - **آنتروپی(Entropy)**
  - **بازیابی(Recall)**
  - **درستی(Precision)**
- روش دوم: اطلاعات جانبی در دسترس نیست(Internal Index).
  - **جمع مربعات خطا – Sum of Squared Error-SSE**

اعتبار سنجی از طریق همبستگی (CORRELATION)

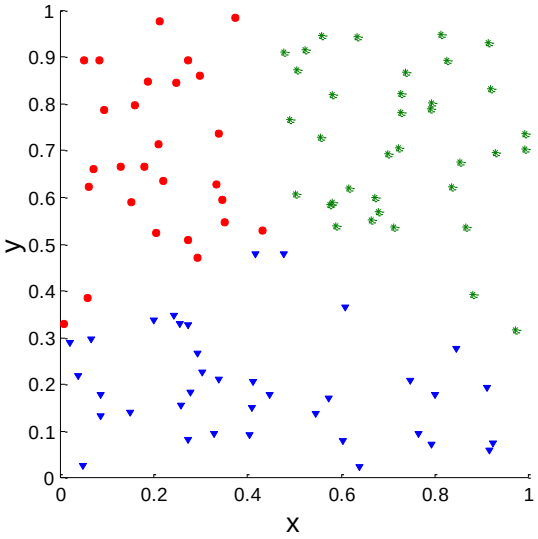
- اعتبار سنجی را می‌توان از روی محاسبه همبستگی بین دو ماتریس محاسبه نمود.
  - ماتریس مجاورت (proximity)(مشابهت یا فاصله)
    - هر داده در یک سطر و ستون است.
    - مقدار ماتریس میزان مشابهت (یا فاصله) دو داده است.
  - ماتریس برخورد (Incidence matrix)
    - هر داده در یک سطر و ستون است.
    - اگر دو داده در یک خوشه باشند مقدار یک و در غیر آن مقدار درایه صفر است.
  - برای برخی خوشه‌بندیهای مبتنی بر چگالی یا همسایگی مناسب نیست.

اعتبار سنجی از طریق همبستگی (CORRELATION)

- همبستگی بین دو ماتریس مجاورت و برخورد برای دو نوع داده زیر که با kmeans خوشه‌بندی شده‌اند.



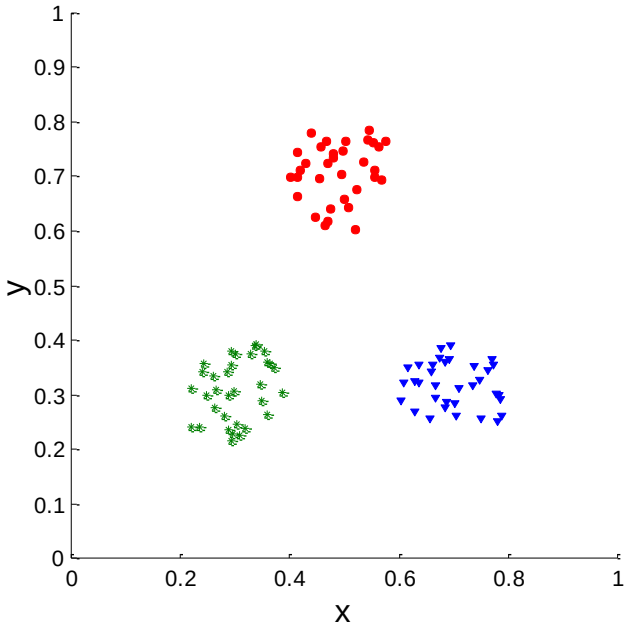
Corr = -0.9235

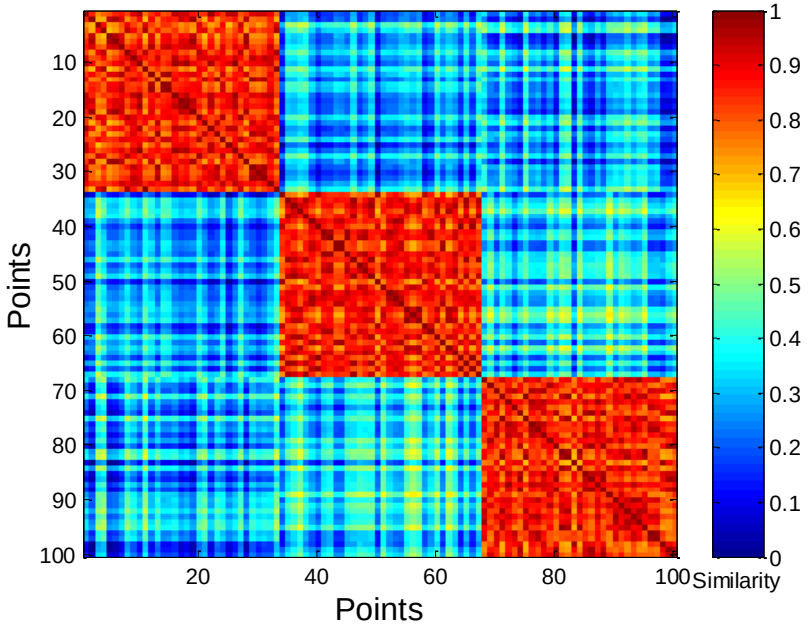


Corr = -0.5810

استفاده از ماتریس مشابهت در اعتبارسنجی خوشه‌بندی

- داده‌ها بر اساس خوشه مرتب شود و به صورت دیداری نتایج بررسی شود.



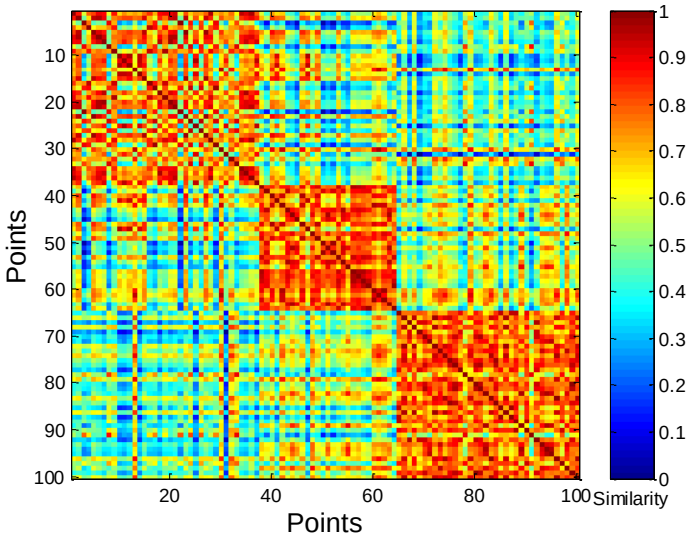


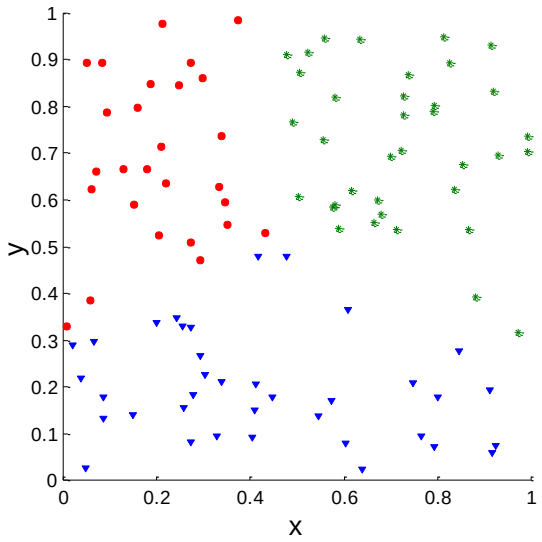
$$sim(i,j) = 1 - \frac{d_{ij} - d_{min}}{d_{max} - d_{min}}$$

استفاده از ماتریس مشابهت در اعتبارسنجی خوشه‌بندی

- خوشه‌ها در داده‌های تصادفی چندان واضح نیست.

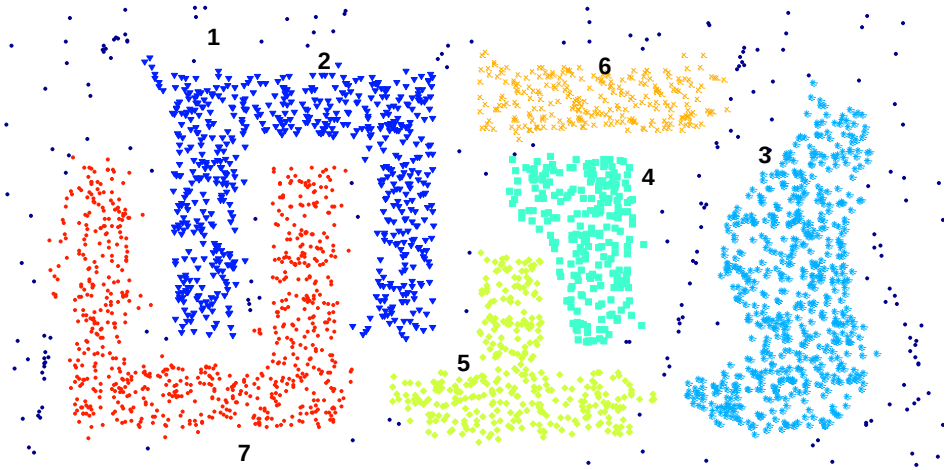
K-means

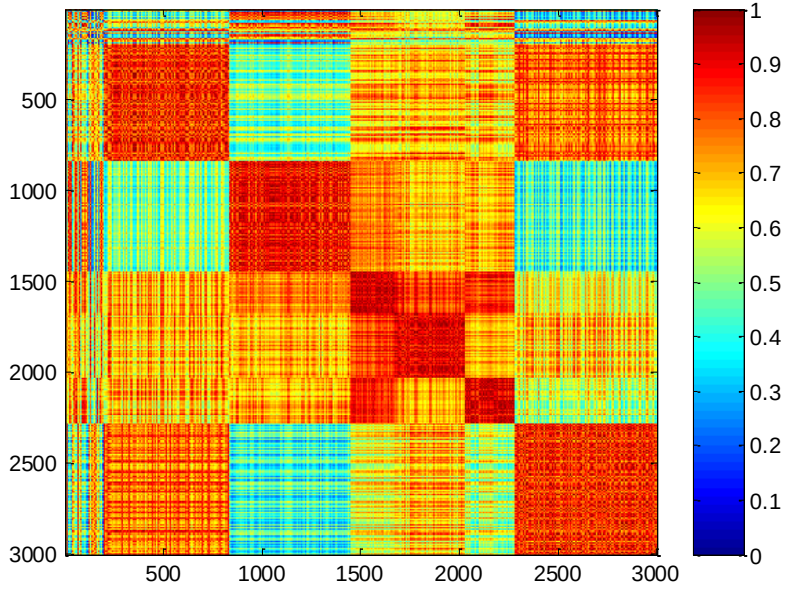




استفاده از ماتریس مشابهت در اعتبارسنجی خوشه‌بندی

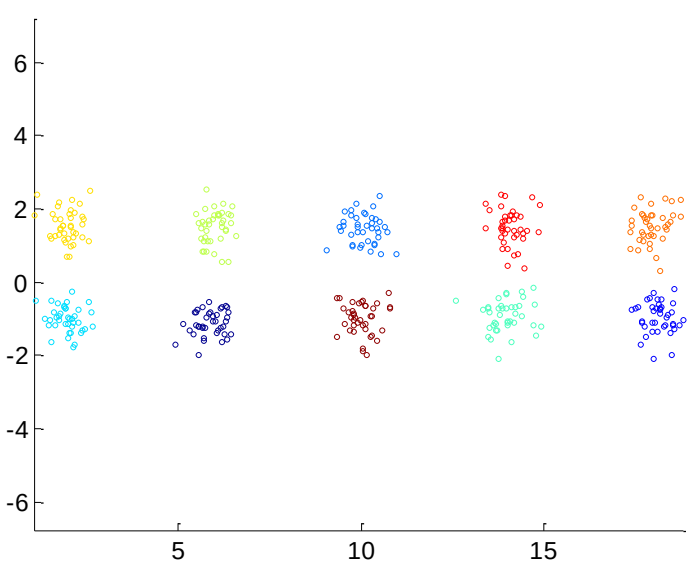
نمای دیداری خوشه‌ها در داده‌های پیچیده

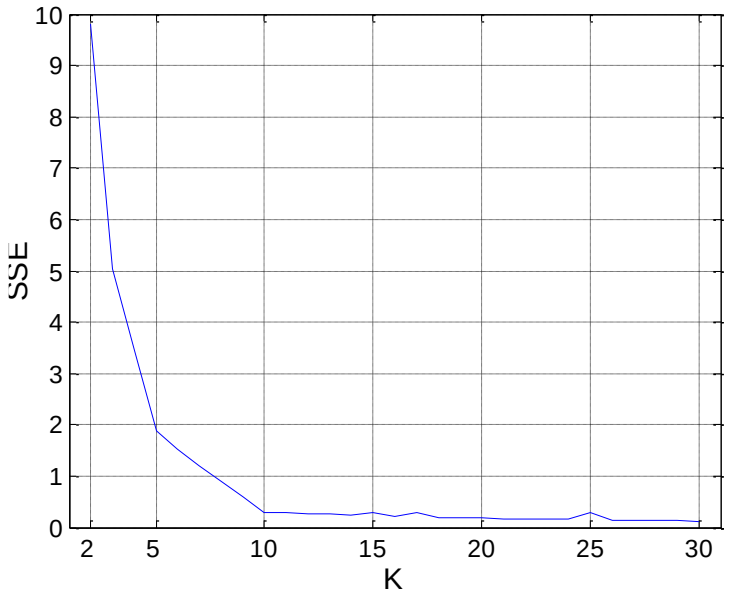




معیار درونی SSE

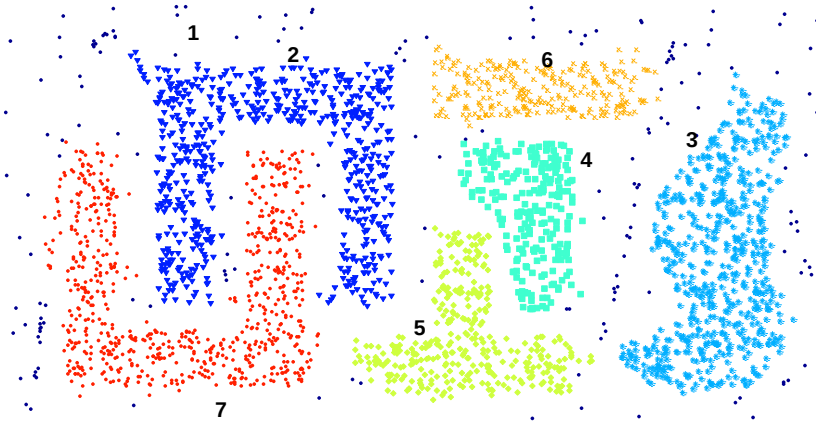
- اندازه‌گیری کیفیت خوشه‌بندی بدون اطلاعات خارجی.
- برای مقایسه نتایج دو خوشه‌بندی
- برای انتخاب تعداد خوشه‌ها هم موثر است.

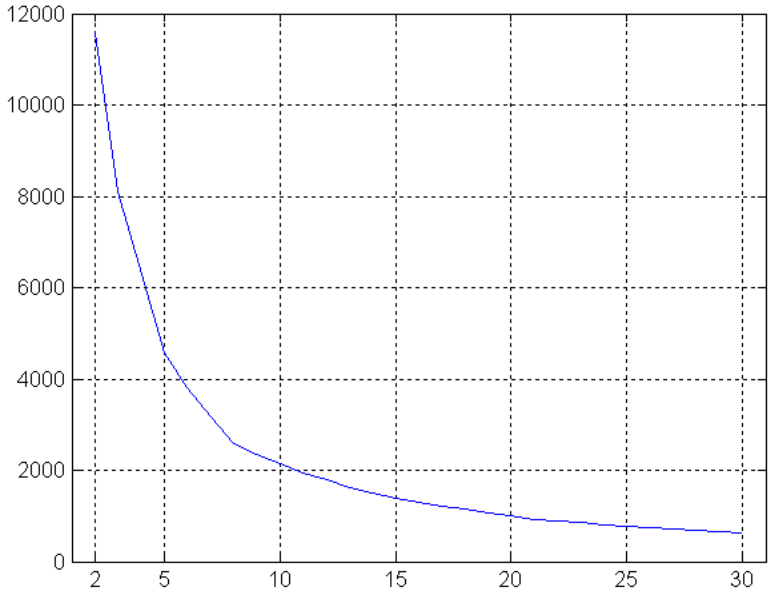




تعداد خوشه

معیار درونی SSE برای داده‌های پیچیده

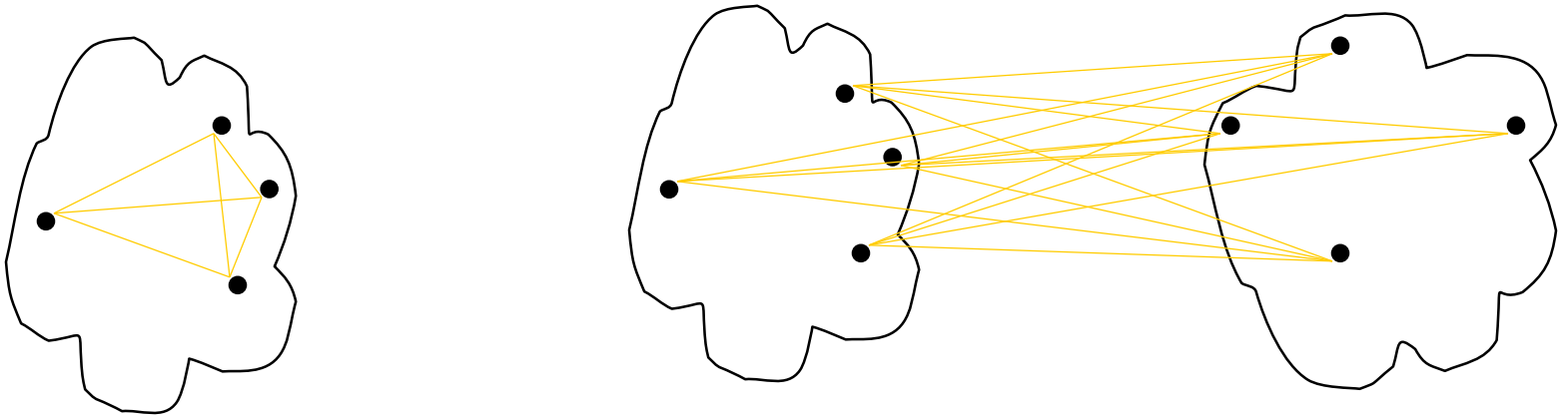




SSE of clusters found using K-means

معیار درونی SSE و جداپذیری انسجام :
COHESION AND SEPARATION

- انسجام: جمع‌وزن‌های درون خوشه
- جداپذیری: جمع وزن‌های بین خوشه‌ها



cohesion

separation

معیار بیرونی برای اعتبارسنجی خوشه‌بندی

- داده‌ها برچسب دارند
- به طور مثال: مستندات بر اساس عنوان، سناتورها بر اساس جمهوری‌خواه و دموکرات، افراد بر اساس گروه‌خونی
- خوشه‌ها بر اساس میزان همگن بودن اعتبارسنجی می‌شوند.
- هرچه داده‌های در یک خوشه از یک کلاس باشند، یا به عبارتی خوشه خالص‌تر باشد، اعتبار خوشه‌بندی بالاتر است.

- n = number of points
- m_i = points in cluster i
- c_j = points in class j
- m_{ij}= points in cluster i coming from class

	Class 1	Class 2	Class 3	
Cluster 1	m_{11}	m_{12}	m_{13}	m_1
Cluster 2	m_{21}	m_{22}	m_{23}	m_2
Cluster 3	m_{31}	m_{32}	m_{33}	m_3
	c_1	c_2	c_3	n

- Entropy:
  - Of a cluster i: $e_i = -\sum_{j=1}^{L} p_{ij} \log p_{ij}$
    - Highest when uniform, zero when single class
  - Of a clustering: $e = \sum_{i=1}^{K} \frac{m_i}{n} e_i$
- Purity:
  - Of a cluster i: $p_i = \max_j p_{ij}$
  - Of a clustering: $purity = \sum_{i=1}^{K} \frac{m_i}{n} p_i$

معیارها

- Precision:
  - Of cluster i with respect to class j: $Prec(i,j) = p_{ij}$
    - For the precision of a clustering you can take the maximum
- Recall:
  - Of cluster i with respect to class j: $Rec(i,j) = \frac{m_{ij}}{c_j}$
    - For the precision of a clustering you can take the maximum
- F-measure:
  - Harmonic Mean of Precision and Recall:

$$F(i,j) = \frac{2 * Prec(i,j) * Rec(i,j)}{Prec(i,j) + Rec(i,j)}$$

خوشه‌بندی بد و خوب

	Class 1	Class 2	Class 3	
Cluster 1	2	3	85	90
Cluster 2	90	12	8	110
Cluster 3	8	85	7	100
	100	100	100	300

Purity: (0.94, 0.81, 0.85) – overall 0.86
Precision: (0.94, 0.81, 0.85)
Recall: (0.85, 0.9, 0.85)

	Class 1	Class 2	Class 3	
Cluster 1	20	35	35	90
Cluster 2	30	42	38	110
Cluster 3	38	35	27	100
	100	100	100	300

Purity: (0.38, 0.38, 0.38) – overall 0.38
Precision: (0.38, 0.38, 0.38)
Recall: (0.35, 0.42, 0.38)