محاضرة تنقيب البيانات

التجميع وخوارزمية K-Means

تنقيب البيانات (Data Mining) — IT401 | ملزمة | جامعة الحكمة — كلية العلوم والهندسة

١. ما هو التجميع؟

التجميع (Clustering) تعلّم غير موجَّه: لا توجد فئات معروفة مسبقاً، والمطلوب تقسيم البيانات إلى مجموعات بحيث تتشابه عناصر المجموعة الواحدة وتختلف عن عناصر المجموعات الأخرى.

	التصنيف	التجميع
نوع التعلّم	موجَّه	غير موجَّه
الفئات	معروفة مسبقاً	تُكتشف من البيانات
مثال	هل الرسالة مزعجة؟	تقسيم العملاء إلى شرائح

٢. قياس التشابه

أشهر المقاييس المسافة الإقليدية بين نقطتين p و q:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

إذا اختلفت مقاييس الخصائص (مثل العمر بالسنوات والدخل بالآلاف) فيجب توحيدها أولاً، وإلا طغت الخاصية ذات الأرقام الكبيرة.

٣. خطوات K-Means

1. اختر عدد المجموعات k.
2. اختر k نقطة كمراكز ابتدائية.
3. أسند كل نقطة إلى أقرب مركز.
4. أعد حساب كل مركز كمتوسط لنقاط مجموعته.
5. كرر الخطوتين ٣ و٤ حتى لا تتغير المجموعات.

٤. مثال محلول

النقاط الأربع: A(1,1) و B(2,1) و C(4,3) و D(5,4)، وعدد المجموعات $k = 2$، والمركزان الابتدائيان هما النقطتان A و B.

النقطة	المسافة إلى (1,1)	المسافة إلى (2,1)	المجموعة
A(1,1)	0	1	١
B(2,1)	1	0	٢
C(4,3)	3.61	2.83	٢
D(5,4)	5	4.24	٢

المراكز الجديدة: الأولى (1, 1)، والثانية (3.67, 2.67) وهي متوسط النقاط B و C و D.

في الدورة الثانية تنتقل B إلى المجموعة الأولى (المسافة ١ مقابل ٢٫٣٦)، فتصير المراكز (1.5, 1) و (4.5, 3.5). في الدورة الثالثة لا يتغير شيء، فتتوقف الخوارزمية.

النتيجة: المجموعة الأولى {A, B} والمجموعة الثانية {C, D}.

٥. المزايا والعيوب

المزايا	العيوب
بسيطة وسريعة مع البيانات الكبيرة	يجب تحديد k مسبقاً
سهلة التفسير	النتيجة تتأثر بالمراكز الابتدائية
تتقارب دائماً	حساسة للقيم الشاذة ولا تكتشف الأشكال غير الكروية

لاختيار k تُستخدم طريقة الكوع (Elbow): ارسم مجموع مربعات المسافات مقابل k، واختر النقطة التي يتباطأ عندها الانخفاض.