ملزمة تنقيب البيانات

الفصل الأول: أساسيات تنقيب البيانات وتجهيز البيانات

١. ما هو تنقيب البيانات؟

تنقيب البيانات هو استخراج أنماط ومعرفة مفيدة وغير بديهية من كميات كبيرة من البيانات. هو خطوة ضمن عملية أوسع تسمى اكتشاف المعرفة في قواعد البيانات (KDD).

٢. مراحل اكتشاف المعرفة

1. **تنظيف البيانات:** معالجة القيم المفقودة والشاذة والمتعارضة.
2. **تكامل البيانات:** دمج البيانات من مصادر متعددة.
3. **اختيار البيانات:** أخذ ما يتعلق بالمهمة فقط.
4. **تحويل البيانات:** وضعها في صيغة تناسب الخوارزمية.
5. **التنقيب:** تطبيق الخوارزميات لاستخراج الأنماط.
6. **تقييم الأنماط:** الإبقاء على المفيد منها.
7. **عرض المعرفة:** تقديم النتائج بشكل يفهمه متخذ القرار.

٣. مهام تنقيب البيانات

المهمة	النوع	مثال
التصنيف	تنبؤية	هل سيسدد العميل القرض؟
التنبؤ العددي	تنبؤية	توقع مبيعات الشهر القادم
التجميع	وصفية	تقسيم العملاء إلى شرائح
قواعد الارتباط	وصفية	المنتجات التي تُشترى معاً
كشف الشذوذ	وصفية	عمليات بطاقات مشبوهة

٤. أنواع الخصائص

النوع	الوصف	مثال
اسمية	أسماء بلا ترتيب	لون العين، التخصص
ثنائية	قيمتان فقط	ناجح / راسب
ترتيبية	لها ترتيب بلا فرق محدد	التقدير: مقبول، جيد، ممتاز
عددية	قيم قابلة للقياس	العمر، المعدل، الدخل

٥. لماذا نجهّز البيانات؟

البيانات الحقيقية ناقصة ومشوشة وغير متسقة، ونتيجة التنقيب لا تكون أفضل من جودة البيانات الداخلة. يأخذ التجهيز غالباً أكثر من نصف وقت المشروع.

٦. تنظيف البيانات

القيم المفقودة

- حذف السجل: مقبول إن كانت السجلات الناقصة قليلة.
- التعويض بالمتوسط أو الوسيط للخصائص العددية، وبالأكثر تكراراً للاسمية.
- التعويض بمتوسط الفئة نفسها: أدق من المتوسط العام.

القيم الشاذة والضوضاء

- التقسيم إلى فترات (Binning) ثم التنعيم بمتوسط كل فترة.
- كشف القيم الشاذة بالمدى الربيعي: ما يقل عن $Q1 - 1.5\times IQR$ أو يزيد على $Q3 + 1.5\times IQR$.

٧. تحويل البيانات

التطبيع Min-Max

$$v' = (v - \min) / (\max - \min)$$

مثال: الدخل بين ١٢٬٠٠٠ و٩٨٬٠٠٠. القيمة ٧٣٬٦٠٠ تصير (73600 − 12000) ÷ (98000 − 12000) = ٠٫٧١٦.

التطبيع Z-Score

$$v' = (v - \text{mean}) / \text{std}$$

مثال: المتوسط ٥٤٬٠٠٠ والانحراف المعياري ١٦٬٠٠٠. القيمة ٧٣٬٦٠٠ تصير ١٫٢٢٥.

> يُفضَّل Z-Score عند وجود قيم شاذة، لأن Min-Max يتأثر بالحدّين الأدنى والأعلى.

٨. تقليص البيانات

- اختيار الخصائص: حذف الخصائص غير المؤثرة أو المكررة.
- تحليل المكونات الرئيسية (PCA): تمثيل البيانات بعدد أقل من الأبعاد.
- أخذ العينات: العمل على عينة ممثلة بدل البيانات كلها.

٩. تمارين

1. البيانات: 4، 8، 9، 15، 21، 21، 24، 25، 26، 28، 29، 34. قسّمها إلى ثلاث فترات متساوية العدد، ثم نعّمها بمتوسط الفترة.
2. طبّع القيمة 21 بطريقة Min-Max إلى المجال [0, 1].
3. صنّف الخصائص التالية حسب نوعها: الرقم الجامعي، المعدل التراكمي، الجنس، المستوى الدراسي.