ملخص خوارزميات التصنيف

مراجعة قبل الاختبار النصفي

١. خطوات التصنيف

1. **بناء النموذج:** من بيانات التدريب المعروفة فئاتها.
2. **اختبار النموذج:** على بيانات لم يرها أثناء التدريب.
3. **الاستخدام:** تصنيف البيانات الجديدة.

٢. شجرة القرار

تُبنى من الجذر إلى الأوراق. في كل عقدة نختار الخاصية التي تعطي أفضل تقسيم.

$$\text{Entropy}(S) = -\sum p_i \log_2(p_i)$$

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum (|S_v| / |S|) \times \text{Entropy}(S_v)$$

مثال: ١٤ سجلاً، منها ٩ «نعم» و٥ «لا». الإنتروبي = ٠٫٩٤٠. نحسب الكسب لكل خاصية ونختار الأعلى جذراً للشجرة.

> خوارزمية ID3 تستخدم كسب المعلومات، و C4.5 تستخدم نسبة الكسب، و CART تستخدم مؤشر جيني.

٣. بايز البسيط (Naive Bayes)

$$P(C \mid X) = P(X \mid C) \times P(C) / P(X)$$

يفترض أن الخصائص مستقلة عن بعضها بمعرفة الفئة. إذا كان أحد الاحتمالات صفراً نستخدم تصحيح لابلاس (نضيف ١ لكل عدّاد).

٤. أقرب الجيران (KNN)

- لا يبني نموذجاً: يخزّن بيانات التدريب ويقارن عند التصنيف.
- يحسب المسافة بين السجل الجديد وكل سجلات التدريب، ويأخذ أقرب k سجلاً، والفئة بالأغلبية.
- يُفضَّل أن يكون k فردياً لتجنب التعادل، ويجب تطبيع الخصائص قبل حساب المسافة.

٥. مقارنة سريعة

الخوارزمية	المزايا	العيوب
شجرة القرار	سهلة التفسير	تميل إلى فرط التخصيص
بايز البسيط	سريعة وتعمل مع بيانات قليلة	تفترض استقلال الخصائص
KNN	بسيطة بلا تدريب	بطيئة عند التصنيف

٦. تقييم المصنِّف

	متوقع: موجب	متوقع: سالب
فعلي: موجب	TP	FN
فعلي: سالب	FP	TN

- الدقة الكلية Accuracy = (TP + TN) ÷ الكل
- الضبط Precision = TP ÷ (TP + FP)
- الاستدعاء Recall = TP ÷ (TP + FN)
- مقياس F1 = 2 × الضبط × الاستدعاء ÷ (الضبط + الاستدعاء)

عند عدم توازن الفئات لا تكفي الدقة الكلية: مصنِّف يقول «سالب» دائماً يحقق ٩٩٪ إن كانت نسبة الموجب ١٪.