

فهرس المحتويات

2 ما هو التعلم الالى
2 ماهي استخدامات التعلم الالى
3 بعض الامثلة على التعلم الالى
3 أنواع التعلم الالى
4 كيفية الإشراف أثناء التدريب
5 القدرة على التعلم
6 طريقة العمل
6 التحديات الاساسية في التعلم الالى
8 كيفية اختيار النموذج
9 مشكلة عدم تطابق البيانات

الفصل الاول

ما هو التعلم الآلي (Machine Learning):

التعلم الآلي هو علم برمجة الحواسيب لتتعلم من البيانات يُعطى الحاسوب القدرة على التعلم دون الحاجة إلى برمجته صراحةً لكل مهمة على سبيل المثال، برنامج فلتر البريد العشوائي يتعلم كيفية تمييز بين الايميلات من خلال الأمثلة التي يتم توفيرها له. فالهدف منه هو تحسين أداء الحاسوب في مهمة معينة، بناءً على الخبرة والبيانات التي يتم تدريبه عليها.

ماهي استخدامات التعلم الآلي:

استخدام التعلم الآلي يمكن أن يكون أكثر فعالية من البرمجة التقليدية في مهام معينة مثل تصفية البريد العشوائي، في البرمجة التقليدية يجب عليك:

1. **فحص البريد العشوائي:** تحليل الرسائل لتحديد الكلمات أو العبارات الشائعة التي تظهر في البريد العشوائي.
2. **كتابة خوارزمية الكشف:** مجموعة قواعد للكشف عن هذه الأنماط وتمييز الرسائل كبريد عشوائي إذا تم اكتشاف عدد معين من هذه الأنماط.
3. **اختبار البرنامج:** تجربة البرنامج وتكرار الخطوات حتى يصبح جاهزاً للاستخدام.

بينما في التعلم الآلي، يتم تدريب النظام على مجموعة من البيانات (مجموعة التدريب) التي تحتوي على أمثلة للبريد العشوائي والبريد العادية، فالنظام يتعلم تلقائياً الأنماط التي تميز البريد العشوائي عن البريد العادي ويطور نموذجاً يمكنه من تصنيف الرسائل الجديدة بشكل صحيح، هذا يوفر الوقت والجهد المطلوب لتحديث القواعد يدوياً ويسمح للنظام بالتكيف مع الأنماط الجديدة للبريد العشوائي التي تظهر باستمرار.

فالتعلم الآلي مفيد بشكل خاص في مواجهة المشكلات المعقدة التي تتجاوز قدرات البرمجة التقليدية أو تلك التي لا يوجد لها خوارزميات معروفة.

بعض الأمثلة على التعلم الآلي:

التعلم الآلي يُستخدم في مجموعة واسعة من التطبيقات، وهنا بعض الأمثلة:

- تصنيف الصور: مثل تحليل صور المنتجات على خط الإنتاج.
- الكشف عن الأورام في الأشعة الدماغية: يُعرف بالتجزئة الدلالية للصور، حيث يتم تصنيف كل بكسل في الصورة.
- تصنيف النصوص: مثل تصنيف المقالات الإخبارية أو تحديد التعليقات المسيئة في المنتديات.
- تلخيص الوثائق الطويلة: يُعرف بتلخيص النص، ويستخدم نفس أدوات معالجة اللغات الطبيعية.
- إنشاء روبوت محادثة أو مساعد شخصي: يتضمن العديد من مكونات، بما في ذلك فهم اللغة الطبيعية.
- التنبؤ بإيرادات الشركة للعام القادم: مهمته تنبؤية، يمكن التعامل معها باستخدام نماذج مثل الانحدار الخطي أو الانحدار الكثير الحدود.....
- جعل التطبيق يستجيب للأوامر الصوتية: يُعرف بالتعرف على الكلام، ويتطلب معالجة عينات الصوت.

أنواع التعلم الآلي:

أنظمة التعلم الآلي يمكن تصنيفها إلى فئات واسعة بناءً على معايير معينة:

1. كيفية الإشراف أثناء التدريب:

- التعلم المشرف (Supervised Learning)
- التعلم غير المشرف (Unsupervised Learning)
- التعلم شبه المشرف (Semi-Supervised Learning)
- التعلم الذاتي (Self-Supervised Learning)

2. القدرة على التعلم:

- التعلم الدفعي (Batch Learning)
- التعلم الأونلاين (Online Learning)

3. طريقة العمل:

- التعلم المبني على الحالات (Instance-Based Learning)
- التعلم المبني على النماذج (Model-Based Learning)

يمكن دمج هذه المعايير بطرق مختلفة على سبيل المثال، قد يكون فلتر البريد العشوائي المتطور نظاماً للتعلم الآلي يعمل أونلاين ومبني على نموذج ويتم تدريبه بشكل مشرف باستخدام أمثلة بشرية للبريد العشوائي والعادي هذا يجعله نظام تعلم مشرف ومبني على النموذج ويعمل أونلاين.

❖ كيفية الإشراف أثناء التدريب:

1. التعلم المشرف (Supervised Learning) :

2. يتم تدريب النماذج على بيانات مصنفة (بيانات مميزة للحالات)، حيث يكون لكل مثال في البيانات التدريبية مُخرج معروف (مثل تصنيف أو قيمة) الهدف هو تعلم قاعدة تتنبأ بالمخرجات الصحيحة للبيانات الجديدة.

- مثال: تصنيف البريد الإلكتروني إلى "العادية" أو "عشوائي".
- التطبيق العملي: استخدام مجموعة من رسائل البريد الإلكتروني المصنف لتدريب نموذج يتعرف على السمات المميزة المهمل ويفرقها عن الرسائل العادية.

3. التعلم بدون إشراف (Unsupervised Learning) :

مُصنفة، والهدف هو اكتشاف أنماط وعلاقات في البيانات دون توجيه مسبق.

- مثال: تجميع العملاء حسب سلوك الشراء.
- التطبيق العملي: استخدام خوارزميات التجميع لتحديد مجموعات العملاء ذوي الخصائص المشتركة لتوجيه الحملات التسويقية.

4. التعلم شبه المشرف (Semi-Supervised Learning) :

إشراف، حيث يتم استخدام كمية صغيرة من البيانات المُصنفة مع كمية كبيرة من البيانات غير المُصنفة لتحسين أداء النموذج.

- مثال: تصنيف الصور عندما يكون لدينا عدد قليل من الصور المُصنفة.
- التطبيق العملي: استخدام الصور المُعلّمة لتحسين دقة تصنيف مجموعة كبيرة من الصور غير المُصنفة.

5. **التعلم الذاتي المشرف (Self-Supervised Learning)** : نوع من التعلم بدون إشراف حيث يقوم النموذج بإنشاء مهام تدريبية من البيانات نفسها، مما يسمح له بتعلم تمثيلات مفيدة دون الحاجة إلى بيانات مُصنفة.

- **مثال:** تعلم تمثيلات الكلمات في معالجة اللغة الطبيعية.
- **التطبيق العملي:** استخدام نماذج مثل BERT التي تتعلم من النصوص الكبيرة بإنشاء مهام تدريبية من النص نفسه، مثل إخفاء بعض الكلمات وتعلم التنبؤ بها.

6. **التعلم بالتعزيز (Reinforcement Learning)**: يتعلم النموذج كيفية اتخاذ قرارات عبر التفاعل مع بيئة معينة، بهدف تعظيم مكافأة معينة. يتم تدريب النموذج على اتخاذ سلسلة من القرارات التي تؤدي إلى أفضل نتيجة ممكنة.

- **مثال:** تدريب روبوت على المشي.
- **التطبيق العملي:** استخدام محاكاة حيث يتلقى الروبوت مكافآت عند اتخاذ خطوات صحيحة ويتعلم تجنب الحركات التي تؤدي إلى السقوط أو الفشل.

❖ القدرة على التعلم:

1. **التعلم الدفعي (Batch Learning)** :

- في التعلم الدفعي، يتم تدريب النموذج على مجموعة كاملة من البيانات دفعة واحدة.
 - يتطلب هذا النوع من التعلم موارد حسابية كبيرة لأنه يعالج كميات كبيرة من البيانات في وقت واحد.
 - بعد التدريب، يتم استخدام النموذج لإجراء التنبؤات دون تحديثه إلا عند توفر دفعة جديدة من البيانات.
 - لسوء الحظ، يميل أداء النموذج إلى التدهور ببطء مع مرور الوقت،
 - ببساطة لأن العالم يستمر في التطور بينما يبقى النموذج دون تغيير.
- مثال:** تدريب نموذج للتنبؤ بأسعار المنازل بناءً على مجموعة كبيرة من البيانات العقارية التي تم جمعها في الماضي.

2. التعلم الاونلاين(Online Learning) :

- في التعلم الاونلاين، يتم تحديث النموذج باستمرار مع كل نقطة بيانات جديدة تصل.
- يتميز بقدرته على التكيف مع البيانات الجديدة بسرعة، مما يجعله مناسباً للبيئات التي تتغير فيها الأنماط بسرعة.
- يستهلك موارد أقل مقارنة بالتعلم الدفعي لأنه يعالج البيانات واحدة تلو الأخرى بدلاً من معالجة مجموعات كبيرة.
- مثال :نظام توصية الأخبار الذي يتكيف مع اهتمامات المستخدمين بناءً على النقرات والتفاعلات.

❖ طريقة العمل:

1. **التعلم المبني على الحالات:** يعتمد هذا النوع من التعلم على مقارنة البيانات الجديدة مباشرةً بالحالات الموجودة في الذاكرة لإجراء التنبؤات أو القرارات، لا يتم إنشاء نموذج عام بل يتم استخدام البيانات الفعلية للتنبؤ بنتائج جديدة.
2. **التعلم المبني على النماذج:** يقوم بإنشاء نموذج رياضي أو معادلة تعبر عن البيانات، يتم استخدام هذا النموذج لإجراء التنبؤات حول بيانات جديدة، يتم تدريب النموذج على مجموعة من البيانات ويتم تحديثه بناءً على دقة التنبؤات!.

التحديات الأساسية في التعلم الآلي:

1. **كمية غير كافية من بيانات التدريب (Insufficient Quantity of Training Data) :** يحدث هذا عندما لا تكون هناك بيانات كافية لتدريب النموذج بشكل فعال، مما قد يؤدي إلى أداء ضعيف لأن النموذج لم يتعلم بما فيه الكفاية من الأمثلة المقدمة.
2. **بيانات تدريب غير ممثلة (Nonrepresentative Training Data) :** إذا لم تكن بيانات التدريب ممثلة للسكان أو الظواهر التي تحاول نمذجتها، فإن التنبؤات ستكون متحيزة ولن تعمم جيداً على السيناريوهات الواقعية.
3. **بيانات ذات جودة رديئة (Poor-Quality Data) :** تشير إلى البيانات التي قد تكون غير دقيقة أو غير كاملة أو غير متسقة، يمكن أن تضلل البيانات ذات الجودة الرديئة النموذج أثناء التدريب مما يؤدي إلى أداء ضعيف.

4. الميزات غير ذات الصلة (Irrelevant Features) :

هي الميزات (المتغيرات الداخلة) التي لا تمتلك أي علاقة بالمتغير الخارج، تضمين الميزات غير ذات الصلة يمكن أن يقلل من دقة النموذج لأنه يحاول إيجاد أنماط لا توجد.

5. التكيف الزائد مع بيانات التدريب (Overfitting the Training Data) :

يحدث هذا عندما يتعلم النموذج بيانات التدريب جيداً جداً، بما في ذلك الضوضاء والقيم الشاذة يؤدي النموذج بشكل استثنائي على بيانات التدريب ولكن بشكل ضعيف على بيانات جديدة غير مرئية لأنه لم يتعلم التعميم.

التكيف الناقص لبيانات التدريب (Underfitting the Training Data) :

يحدث التكيف الناقص عندما يكون النموذج بسيطاً جداً ولا يستطيع التقاط التعقيدات الموجودة في البيانات.

كيفية اختيار النموذج:

عند تقييم نماذج التعلم الآلي يمكن استخدام مجموعة اختبار لتحديد كيفية أداء النموذج على بيانات جديدة إذا كنت تتردد بين نوعين من النماذج، مثل النموذج الخطي والنموذج الكثير الحدود، يمكنك تدريب كلا النموذجين ومقارنة أدائهما باستخدام مجموعة الاختبار.

إذا كان النموذج الخطي يعمم بشكل أفضل ولكنك ترغب في تطبيق بعض التنظيم (regularization) لتجنب التكيف الزائد (overfitting)، يمكنك تدريب العديد من النماذج باستخدام قيم مختلفة للبارامترات الفائقة (hyperparameters) للتنظيم بعد اختيار أفضل قيمة تعطي أقل خطأ في التعميم، قد تجد أن النموذج لا يؤدي بشكل جيد عند إطلاقه في الإنتاج.

المشكلة هي أنك قمت بقياس الخطأ في التعميم عدة مرات على مجموعة الاختبار وقمت بتكييف النموذج والبارامترات الفائقة لإنتاج أفضل نموذج لتلك المجموعة بالتحديد، مما يعني أن النموذج قد لا يؤدي بشكل جيد على بيانات جديدة.

الحل الشائع لهذه المشكلة يسمى التحقق الصحيح (holdout validation)، حيث تحتفظ بجزء من مجموعة التدريب لتقييم عدة نماذج مرشحة واختيار الأفضل منها يُطلق على الجزء المحفوظ به اسم مجموعة التحقق (validation set) أو مجموعة التطوير (dev set) تقوم بتدريب نماذج متعددة ببارامترات فائقة مختلفة على مجموعة التدريب المخفضة (أي مجموعة التدريب الكاملة مطروحاً منها مجموعة التحقق)، وتختار النموذج الذي يؤدي بشكل أفضل على مجموعة التحقق بعد عملية التحقق الصحيح، تقوم بتدريب أفضل نموذج على مجموعة التدريب الكاملة (بما في ذلك مجموعة التحقق)، وهذا يعطيك النموذج النهائي. أخيراً، تقوم بتقييم هذا النموذج النهائي على مجموعة الاختبار للحصول على تقدير لخطأ التعميم.

التحقق المتقاطع المتكرر هو طريقة تساعدك على تقييم النموذج بشكل أدق بدلاً من استخدام مجموعة تحقق واحدة كبيرة، تقوم بتقسيم البيانات إلى العديد من مجموعات التحقق الصغيرة وتقييم النموذج على كل واحدة منها بشكل منفصل ثم تأخذ متوسط النتائج لتحصل على تقييم أكثر دقة لأداء النموذج ولكن، هذه الطريقة تتطلب وقتاً أطول لأنك تقوم بتدريب النموذج عدة مرات، مرة لكل مجموعة تحقق.

مشكلة عدم تطابق البيانات:

أن البيانات المستخدمة لتدريب النموذج قد لا تمثل بدقة البيانات التي سيتم استخدامها في الإنتاج الفعلي، على سبيل المثال، إذا أردت إنشاء تطبيق للهاتف المحمول لتصوير الزهور وتحديد أنواعها تلقائيًا، فقد تتمكن من تحميل ملايين الصور من الإنترنت، لكن هذه الصور قد لا تكون ممثلة بشكل كامل للصور التي ستلتقط فعليًا باستخدام التطبيق.

لذلك، من المهم أن تكون مجموعتنا التحقق والاختبار ممثلتين قدر الإمكان للبيانات المتوقعة استخدامها في الإنتاج، ويجب أن تتكونان حصريًا من الصور الممثلة، يمكن خلط هذه الصور وتقسيمها بين مجموعتي التحقق والاختبار، مع التأكد من عدم وجود صور مكررة في كلا المجموعتين.

بعد تدريب النموذج على الصور المأخوذة من الإنترنت، إذا كان أداء النموذج على مجموعة التحقق غير مرضٍ، فلن تعرف ما إذا كان ذلك بسبب تخصيص النموذج بشكل مفرط لمجموعة التدريب، أو بسبب عدم تطابق البيانات بين الصور المأخوذة من الإنترنت والصور التي ستلتقط باستخدام التطبيق.

الحل المقترح هو الاحتفاظ ببعض الصور المأخوذة من الإنترنت في مجموعة أخرى تُسمى مجموعة التدريب التطويرية (train-dev set)، والتي أطلق عليها (Andrew Ng) هذا الاسم، بعد تدريب النموذج على مجموعة التدريب (وليس على مجموعة التدريب التطويرية)، يمكن تقييمه على مجموعة التدريب التطويرية إذا كان الأداء ضعيفًا، فهذا يعني أن النموذج قد تخصص بشكل مفرط لمجموعة التدريب، ويجب عليك حينها تبسيط النموذج أو تنظيمه، والحصول على المزيد من البيانات للتدريب، ولكن إذا كان الأداء جيدًا على مجموعة التدريب التطويرية، فيمكنك تقييم النموذج على مجموعة التطوير إذا كان الأداء ضعيفًا، فإن المشكلة يجب أن تكون ناتجة عن عدم تطابق البيانات. يمكنك محاولة حل هذه المشكلة من خلال معالجة الصور المأخوذة من الإنترنت لجعلها تبدو أكثر شبهاً بالصور التي ستلتقط باستخدام التطبيق ثم إعادة تدريب النموذج.

بمجرد أن يكون لديك نموذج يؤدي بشكل جيد على كل من مجموعة التدريب التطويرية ومجموعة التطوير، يمكنك تقييمه للمرة الأخيرة على مجموعة الاختبار لمعرفة مدى احتمالية أدائه بشكل جيد في الإنتاج.

