ملخص محاضرات هندسة العناصر الدفعة الثانية - محاضرة رقم (1)

- سيتم التعامل في هذا الكورس مع البيانات الـ structured فقط .
- عملية بناء النموذج هي عملية لا تأخذ أكثر من 10٪ من وقت مهندس الـ ML يمكن القيام بها بسهولة عبر المكتبات، فالجزء الأصعب هي هندسة العناصر، لأنه ليس له خطوات واضحة.
- عملية هندسة العناصر هي قريبة أو مرادفة لعملية الـ Data Processing.

1- ما هي هندسة العناصر/السمات Feature Engineering ؟

- هي فن تحضير البيانات والتعامل مع الـ Features لتكون جاهزة لاستخدامها في الـ ML Model وتساعد في تحقيق أفضل دقة ممكنة.
- أصعب شيء في هندسة العناصر هو عدم وجود خطوات واضحة تنفع مع كل البيانات، فكل بيانات لها خصائصها وأسلوب تعامل خاص بها وحتى في الشركة الواحدة ومع العميل الواحد قد تختلف طرق تحضير البيانات بناء على تغيير طلبات العميل واحتياجاته. مثالـــ على ذلك، إذا لديك عمود يحتوي على تصنيف الطلاب حسب الديانة (مسلم، مسيحي، يهودي،)، إذا أردت بناء نموذج لتوقع احتمالية القبول في الجامعة فستكون كارثة كبيرة لأنه سيكون نموذج به تحيز (Bias) كبير لذا يجب أن تزيل هذا العمود وألا تدرب الـــ Model على بياناته. ولكن في حالة أردنا اســتخدام نفس البيانات لتوقع احتمالية القبول في رحلة عمرة أو حج، فهنا يجب اســتخدام هذا العمود لأنه من غير الممكن أن يفوز مسيحي برحلة حج.
- من لديه خبرة جيدة في هندسة العناصر يعرف كيف ينمي نفسه، وكيف يتعامل مع الكود بطريقة معينة لتحقيق أفضل Accuracy فمثلاً عند النظر إلى Code لشخص محترف في Kaggle ستجد أن قرابة الـ 80% أو أكثر من الكود يركز على الـ

  Data Processing or Feature Engineering.

- ولو دققت في كود شخص مبتدئ ستجد أن كل تركيزه على خطوات بناء الـ ML Model

  (Splitting--> Training--> Predict --> Accuracy)

  ولو قام بشيء سيقوم بإزالة الـ Nulls

2- ما هي المتطلبات القبلية لدراسة هندسة العناصر؟

1- معرفة فوق المتوسط بلغة بايثون.
2- أساسيات تعلم الآلة.

3- ما هي أهمية هندسة العناصر؟

1- معرفة كيفية التعامل مع البيانات المعقدة كالتي تكون بها الكثير من الـ (Rubbish &Noise) و البيانات التي يكون الارتباط (Correlation) بين الـ Feature & Output ضعيف.

2- أكثر من 60٪ من مهام عمل مهندس تعلم الآلة يتركز في مرحلة تحضـــير البيانات وهي هندسـة العناصـر. فالشـخص الذي يقوم بهندسـة العناصـر هو أكثر شـخص يتخذ قرارات في عالم الذكاء الاصطناعي.

3- قبل أكثر من عشرين سنة كان المصدر الأساسي للبيانات هي الاستبيانات ومجموعات التركيز، ومع الانتشـار الضـخم للأنترنت والبرمجيات والهواتف الذكية أصـبح هناك كمية هائلة من البيانات. فمهمة مهندس الذكاء الاصـطناعي هي تحديد مدى نظافة هذه البيانات، والى أي درجة يمكن الاعتماد عليها؟ وكلما كثرت البيانات، كلما كان التحكم فيها أصـعب، وكلما كانت إمكانية عمل Checking لها تكون أصـعب. وهناك مبدأ مهم(Rubbish in Rubbish out) أي إذا دخلت بيانات غير صـحيحة في الـ Model ستؤدي إلى نموذج قراراته قد تكون كارثية .

4- مهمة هندسة العناصر هي كيف يمكن من بيانات بها Rubbish & noise كثيرة تستطيع أن تخرج بأفضل البيانات وأفضل الـ Models للتعامل معها. فالأفكار كثيرة، ولكن المعيار هو كم شخص يستطيع تطبيقها بنجاح على بياناته؟ وما هي درجة الاستدامة (Sustainability) -أي مدى نجاح النموذج في التطبيقات العملية-.

5- في حالات قد تكون الـ Accuracy عالية ولكن في المقابل يكون هناك Bias or Overfitting، وفي بعض الأحيان 1% من الخطأ يمكن أن يؤدي إلى مشاكل كبيرة. لذا كيف يمكن اجراء عملية hypernating للتفريق بين الأشياء بدقة هو المهم .

4- ما هي عملية تنظيف البيانات (Data Cleaning)؟

- تتضمن عملية تنظيف البيانات إزالة كل العناصر التي لا تتلاءم مع البيانات التي لدي، مثل الأخطاء البشرية، القيم الفارغة Null أو الاضافات التي تنشأ من نوع البيانات مثل HTML Tags التي نحصل عليها من عملية الـ scrabbling لمواقع الانترنت والتعامل معها هو من نوع الـ (text cleaning).
- تنظيف البيانات هي أول خطوة في الـ ML ويدخل فيها عدد كبير من العوامل، فمثلاً عملية مثل إزالة الـ Nulls أو البيانات المفقودة والتي تتلخص في إحدى ثلاثة خيارات
  1- Dropping row
  2- Dropping column
  3- Substitution

- اختيار أي منها يعتمد على عوامل عديدة:
  - نوع البيانات،
  - عدد العناصر features ،
  - هل يتم عمل Classification or Clustering،
  - وإذا كانت Classification هل تكون Binary or Multi؟،
  - وما هو المخرج المطلوب؟،
  - وما هي درجة المضاهاة بين الـ Features & Output،
  - وما الذي يحتاجه العميل؟ وعشرات الأسئلة الأخرى.

- إذا سال أحدهم أي طرق التنظيف أفضل؟ فهذا يعني أنه ليس له خبرة كبيرة في الذكاء الاصطناعي لأنه يوجد عشرات الأسئلة التي تحتاج للجواب عليها قبل التحديد .
- البيانات الـ Null تكون واضحة ويمكن تمييزها، ولكن هناك نوع من الـ Rubbish يكون غير واضح. فمثلاً قد تكون البيانات خاطئة، ومن مهام تنظيف البيانات هي تحديد ما إذا كانت البيانات خاطئة أم لا؟

5- ما هي أفضل دقة Accuracy يمكن الوصول إليها في الـ ML MODEL؟

- الدقة المقبولة في الـ ML Model هي الدقة التي تضاهي سقف الدقة البشرية

(Human Ceiling Accuracy).

- تكون الدقة مثلاً بين (70-80%) أو ما يوازيها مقبولة في حالة عدم تمكن الإنسان الطبيعي من توقع الـ output من خلال الـ features الموضوعة وذلك لضعف العلاقة بين الـ Features & Output.
- ربما تكون دقة (99%) غير مقبولة في حالة أنه من السهل على الإنسان الطبيعي معرفة العلاقة بين الـ Features & Output ومثال لهذا هي خوارزمية الفيسبوك التي صنفت فيديو لرجل أسود على أنه كائنات بدائية (Primates) برغم دقتها العالية مما أدى إلى ضجة إعلامية كبيرة .

6- ما هو دور مهندس الـ ML؟

- دوره هو استلام البيانات من العميل، والقيام بتهيئتها، ومن ثم القيام ببناء النموذج المناسب .
- هناك بعض الشركات الصغيرة قد تطلب أشياء مثل تصميم الـ Front End, GUI, Development وغيرها، وكلها ليس لها علاقة بعمل مهندس تعلم الآلة. وينصح بعدم التشتت ودراسة هذه الأشياء لأنها ستأخذ من وقتك.
- ولكن إذا كان هناك شيء واحد خارج حزمة تعلم الآلة يمكنك دراسته هو الـ Cloud Computing نسبة لأنها أصبحت مهمة جداً مع بدء معظم الشركات في التحول للعمل Remotely والأفضل منها هو الـ Amazon AWS، ولكن ركز على فهم الجزء الخاص بالذكاء الاصطناعي وليس التعمق في الـ Cloud.

- في بعض الحالات قد لا يكون مهندس الـ ML مسؤول عن مهمة تحضير البيانات، ولكن يجب عليه أن يفهم الطريقة التي تم بها تحضير البيانات وما إذا كان الشخص الذي قام بها يمتلك المهارة الجيدة أم لا، وما إذا كانت البيانات تحتاج إلى معالجة أكثر، لأن أي خطأ في الـ Model ستكون مسؤول عنه أنت أولاً حتى إذا لم يكون خطأك.
- لذا من الضروري على مهندس الـ ML معرفة الخطوات التي تسبق عمله (Feature Engineering) وايضًا الخطوات التي تأتي بعد بناء الـ Model تتمثل في عملية الـ (Evaluation)

7- ما هو الفرق بين الـ Machine Learning والـ Deep Learning؟

- الـ Machine Learning تتعامل مع الـ Structured Data مثل الأرقام.
- الـ Deep Learning تتعامل مع الـ Unstructured Data مثل الصور والنصوص والفيديوهات...
- لكن هذا الفرق ليس شيء ثابت، لأن هذه العلوم حديثة نسبياً وقد يصادفك مشاريع من الصعب التفريق بينها.

8- ما هي خطة الدراسة المفضلة في الـ ML؟

إذا كنت مبتدأ تحتاج إلى البدء في دراسة:

- أساسيات الـ Machine Learning مثل الـ Regression, Classification, Feature Engineering, Svm, Scikit Learn....etc

- دراسة أحد لغات البرمجة الخاصة بالـ AI
والأفضل بايثون، استخدام لغة برمجة غير خاصة بالـ AI تعتبر مشكلة لنقص الـ Resources And Communities

- بعد الانتهاء من دراسة اساسيات تعلم الآلة، وبايثون للذكاء الاصطناعي، قم بالاطلاع بشكل خفيف على هذه التخصصات واختر إحداها فقط (لا تحاول التعمق في جميعها) :
    1. Computer Vision
    2. Reinforcement Learning
    3. Natural Language Processing

يمكنك دراسة كل قسم حوالي شهرين، ومن اختيار أحدهم للتخصص فيه.

تلخيص: مصطفى بوش تنسيق: أحمد زكي
