أكاديمية طويق
TUWAIQ ACADEMY

معسكر علم البيانات و تعلم الآلة

16 -11 - 2022

أكاديمية طويق
TUWAIQ ACADEMY

نبذة عن المدرب

أكاديمية طويق
TUWAIQ ACADEMY

محتوى المعسكر

اليوم	الأسبوع الأول Getting Started	الأسبوع الثاني Data Analysis and Visualization	الأسبوع الثالث Machine Learning	الأسبوع الرابع EDA & FE in Action	الأسبوع الخامس Modeling Interpretation in Action	الأسبوع السادس Final Project
الأحد	Intro to DS	NumPy	Intro to ML	DS Knowledge Catalog	Models Families: Distance & Time Series	Final Project
الإثنين	Git & Github	Pandas	Supervised ML	EDA1: Univariate & Multivariate Analysis	Models Evaluation: Regression & Classification	Final Project
الثلاثاء	Python Review	Matplotlib	Supervised ML	EDA2: Association Analysis & Hypothesis Construction	Optimization Techniques	Final Project
الأربعاء	Python Review	Seaborn	Unsupervised ML	Features Engineering: Scaling, Merging & Discretization	NLP and Text Mining Basics	Final Project
الخميس	Python Review	Plotly	Unsupervised ML	**Models Families: Continuous & Categorical**	Neural Networks Basics	Presentation

**ملاحظة: قد تتغير المواضيع أو أوقات طرحها بناء على تقدم الطلاب.

عائلات النماذج

عوائل النماذج

الانحدار Continuous

تعتبر أنه توجد أنماط بين الخواص

* Logistic Regression
* Linear Regression
* Neural Networks

المسافة Distance

تعتبر وجود مسافة بين الخواص

* K-Means Clustering
* SVM
* DBScan

مصنِّفة Categorical

خواص تحتوي تصنيفات غير قابلة للترتيب (if statements)

* Naïve Bayes
* Decision Trees
* Random Forest

المتسلسلات الزمنية Time Series

تعتمد البيانات اللاحقة على البيانات السابقة

* ARIMA
* Prophet
* Markov

أكاديمية طويق
TUWAIQ ACADEMY

نماذج الانحدار
Continuous Models

Linear Regression

المعادلة:

$$y = a_1x_1 + a_2x_2 + \cdots + b$$

- a و b تمثل معاملات الخوارزمية hyperparameter
- مجال القيم التي يتخذها y تعتمد على مجال الخواص $x_n \ldots x_1$
- حساسة جدًّا للقيم الشاذة

Linear Regression

المعادلة:

$$y = a_1x_1 + a_2x_2 + \cdots + b$$

- **الافتراضات عند لاختيار الخوارزمية:**
  - Linearity
  - Homoscedasticity
  - Independence
  - Normality

Linear Regression

مميزاتها:

- قابلة للتعميم
- مستقرة نتائجها غالبًا
- سهلة التفسير

سلبياتها:

- بطيئة في التكيف مع التغيرات
- تضمن جميع البيانات في نمط واحد
- تعاني من كثرة أبعاد الخواص

مثال دارج لتطبيقاتها: توقع الأسعار

Logistic Regression

المعادلة:

$$y = \frac{1}{1 + e^{-(a1x1+a2x2+\cdots+b)}}$$

- a و b تمثل معاملات الخوارزمية hyperparameter
- متضمنة للانحدار الخطي
- مجال القيم التي يتخذها y تعتمد على مجال الخواص $x_1 \dots x_n$

Logistic Regression

المعادلة:

$$y = a_1x_1 + a_2x_2 + \cdots + b$$

- **الافتراضات عند لاختيار الخوارزمية:**
  - Linearity
  - Homoscedasticity
  - Independence
  - Normality

Logistic Regression

مميزاتها:

- قابلة للتعميم
- مستقرة نتائجها غالبًا
- سهلة التفسير

سلبياتها:

- بطيئة في التكيف مع التغيرات
- تضمن جميع البيانات في نمط واحد
- تعاني من كثرة أبعاد الخواص
- قد تعطي توقعات غير طبيعية

مثال دارج لتطبيقاتها: مخاطر ائتمان القروض

Neural Networks

المعادلة:

$$Z_1 = \sigma(A_1 X + B_1)$$
$$Z_2 = \sigma(A_2 Z_2 + B_2)$$
$$\dots$$
$$Z_n = \sigma(A_n Z_{n-1} + B_n)$$
$$y = f(Z_n)$$

Iterative application of linear regression followed by an “activation function” σ

Model parameters : Ai , Bi , and many hyperparameters

Neural Networks

المعادلة:

$$Z_1 = \sigma(A_1 X + B_1)$$
$$Z_2 = \sigma(A_2 Z_2 + B_2)$$
$$\dots$$
$$Z_n = \sigma(A_n Z_{n-1} + B_n)$$
$$y = f(Z_n)$$

- Universal approximator: can approximate any function F (given the same input)
- Multiple applications by combining different f and loss functions
- Regression? f = x, L = L2
- Classifications? f = softmax, L = crossentropy

Neural Networks

مميزاتها:

- تستطيع استنباط نماط معقدة غير خطية
- تغطي معظم أنواع المشاكل
- قليلة التأثر بالحالات الشاذة

سلبياتها:

- نتائج غير قابلة للتفسير
- نتائج غير مستقرة
- من الصعب ضبط معاملات الخوارزمية للوصول للنتيجة الأمثل
- تتطلب عدد ضخم من البيانات للتدريب

مثال دارج لتطبيقاتها: نماذج تحليل المشاعر

أكاديمية طويق
TUWAIQ ACADEMY

نماذج التصنيف
Classification Models

Naïve Bayes

المعادلة:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

$$P(C|X) = \frac{P(x_1|C)P(x_2|C) \dots P(x_n|C)P(C)}{P(X)}$$

- **Naïve** means the features are conditionally independent
  P(X|C)= P(x1|C) P(x|2C) … P(x|nC)
- P(X|C) is defined by user
- Suitable to use for unbalanced classes

Naïve Bayes

المعادلة:

$$y = a_1x_1 + a_2x_2 + \cdots + b$$

- **الافتراضات عند لاختيار الخوارزمية:**

- Features are independent
- Likelihood distribution

Naïve Bayes

مميزاتها:

- تستطيع استنباط أكثر من نمط
- تغطي معظم أنواع المشاكل
- سهلة التفسير

سلبياتها:

- يتأثر أداؤها عن التعديل في التوزيع
- تأخذ جميع البيانات لتضمنها بالنموذج

مثال دارج لتطبيقاتها: تصنيف الايميلات المزعجة

Decision Trees

المعادلة:

IF statements splitting the data into different dimensions based on class entropy

- Can split one feature multiple times
- No model parameters
- Hyperparameters are searchable

Random Forests combines multiple decision trees in order to bypass some of the disadvantages of decision trees

Decision Trees

- الافتراضات عند لاختيار الخوارزمية:

- Features are independent
- Assumes independence of features

Decision Trees

مميزاتها:

- تستطيع استنباط أكثر من نمط
- يرتب الخواص حسب الأهمية
- سهلة التفسير

سلبياتها:

- تتطلب عدد قليل من الخواص كمدخلات
- عرضة للOverfitting
- لا تعمم بطريقة ممتاز

مثال دارج لتطبيقاتها: تنبؤات الجو

للتسليم

من مشاريعكم، استعرضوا النماذج بعد التدريب واستخرجوا المعادلة النهائية بعد التعرف على معاملاتها coef_

المتوقع: شكل المعادلة لنموذجهم بعد التدريب (نموذج واحد فقط)

أكاديمية طويق
TUWAIQ ACADEMY

استفساراتكم؟ 🤔

أمثلة مشاريع

DS Projects

1. Great Firewall of China, gigantic censorship system of the Chinese government is a good example of what can be achieved with data science. It monitors millions of tweets, posts, links, pages, automatically block requests containing certain keywords, etc. Plus it does that at the scale of the whole Chinese Internet which is hundreds of millions of users and billions of strings of texts to process each minute.
2. Youtube recommender:
   - https://blog.hootsuite.com/how-the-youtube-algorithm-works/amp/
   - https://www.technologyreview.com/2022/09/20/1059709/youtube-algorithm-recommendations/amp/
3. Airbnb
4. http://www.cognitivetoday.com/2017/03/data-science-success-stories.html
5. أمثلة أخرى مع الدكتور حمود: https://twitter.com/dr_hmood/status/929463611961106432?s=46&t=ZZLOVGVoVr9IlpLjCKSsjQ