أكاديمية طويق
TUWAIQ ACADEMY

معسكر علم البيانات و تعلم الآلة

20 -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

نماذج للمسافات
Distance Models

K-Means

المعادلة:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

- الجزء الأول يمثل عدد الكتل clusters
- الجزء الثاني يمثل عدد الحالات
- الجز الأخير يمثل معادلة حساب المتغيرات والتي قدد تتغير باختلاف الخوارزمية المتبعة، هنا نرى خوارزمية المسافة الإقليدية Squared Euclidean Distance

أمثلة أخرى لحساب المسافات؟

K-Means

مميزاتها:

- سهلة الفهم والتطبيق
- تعمل بشكل جيد مع البيانات الضخمة
- Unsupervised

سلبياتها:

- غير قابلة للتفسير
- حساسة للتغيرات التي تطرأ على الـ Scales
- حساسة للحالات الشاذة
- حساسة لكثرة الأبعاد

مثال دارج لتطبيقاتها: تقسيم العملاء / تصنيف المستندات

K-Means

طرق أخرى لحساب المسافات:

- Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

- Manhattan Distance (Taxicab or City Block)

$$\sum_{i=1}^{k} |x_i - y_i|$$

- Minkowski Distance

$$\left(\sum_{i=1}^{k} \left(|x_i - y_i| \right)^q \right)^{1/q}$$

خوارزميات أخرى تتضمن حساب المسافات

- KNN
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)

DBSCAN

فكرتها:

نقطة من البيانات تنتمي إلى الكتلة فقط إذا كانت قريبة من عدد كبير من النقاط المتواجدة في تلك الكتلة.

المتغيرات فيها:

- Eps: وهو المسافة اللي تحدد النقاط المجاورة

- minPts: أقل عدد من النقاط ليتم اعتبار مجموعة من هذه النقاط تعتبر كتلة

DBSCAN

فكرتها:

بالنظر للمتغيرات السابقة التي تعتمد عليها DBSCAN فإننا نستنتج أنها تعتمد على أنواع مختلفة من نقاط البيانات

أنواع النقاط:

1. نقطة أساسية Core Point
2. نقطة الحد Boarder Point
3. نقاط شاذة Outliers

DBSCAN

النقطة الأساسية Core Point

هي النقطة التي:

1. لها عدد minPts من النقاط المجاورة لها مع النقطة ذاتها
2. نصف القطر منها يساوي eps

DBSCAN

نقطة الحد Boarder Point

هي النقطة التي:

1. لها عدد أقل من minPts من النقاط المجاورة

2. قابلة للوصول من النقطة الأساسية

DBSCAN

نقاط شاذة Outliers

هي النقطة التي:

1. لا تُعد نقطة أساسية

2. غير قابلة للوصول من أي من النقاط الأساسية

DBSCAN

مميزاتها:

- لا تحتاج لتحديد عدد الكتل مسبقًا
- ذات أداء ممتاز مع التجمعات عشوائية التمثيل
- متكيفة مع النقاط العشوائية وتتعامل منعها على حدة

سلبياتها:

- في معظم الأحيان، تحديد متغير المسافة eps لا يُعد عملية سهلة ويحتاج معرفة بالأعمال التي يُبنى من أجلها
- إذا كانت الكتل أو التجمعات التي نهدف للوصول إليها لا تعتمد بشكل أساسي على الكثافة والتكدس، فهي ليست بالاختيار الأمثل

مثال دارج لتطبيقاتها: التسويق وتطبيقات نماذج التوصية

Support Vector Machine - SVM

المعادلة:

$$C = \alpha Kx + \beta > \epsilon$$

- المعاملات: α, β
- معاملات الخوارزمية: kernel function K
- تقسم أبعاد البيانات المدخلة إلى تصنيفين
- تفترض وجود ترابط كبير بين المدخلات Binary Classes

ما هي الخوارزمية (من اليوم السابق) التي تفترض وجود تصنيفين فقط لتعمل بشكل صحيح؟

Support Vector Machine - SVM

مميزاتها:

- لا تتأثر بشكل توزيع البيانات
- قليلة التأثر بالحالات الشاذة
- تعمل بشكل جيد مع البيانات الضخمة
- تتواءم مع الأنماط المعقدة

سلبياتها:

- بطيئة في التحسن والاقتراب من النتائج الفعلية
- عرضة للOverfitting

مثال دارج لتطبيقاتها: تصنيف النصوص

أكاديمية طويق
TUWAIQ ACADEMY

المسلسلات الزمنية
Time Series

ARIMA & SARIMA

المعادلة:

$$y_n = a_1 y_{n-1} + a_2 y_{n-2} + \dots$$
$$b_1 y_{n-s} + b_2 y_{n-2s} + \dots$$
$$\mu_1 w_1 + \mu_2 w_2 + \dots$$

- Model parameters: ai , bi , ci

  ARIMA can be summarized in three hyperparameters (p,q,r): p for number of previous terms, q for number of moving average terms, and r for differencing order

ARIMA & SARIMA

المعادلة:

$$y_n = a_1 y_{n-1} + a_2 y_{n-2} + \dots$$
$$b_1 y_{n-s} + b_2 y_{n-2s} + \dots$$
$$\mu_1 w_1 + \mu_2 w_2 + \dots$$

- wi is independent normally distributed parameters Differencing may be applied to ensure stationarity
- $zi \rightarrow (yi - yi-1)$ then applying ARIMA on $\{zn\}$
- Seasonal ARIMA adds seasonality terms with hyper parameter (s). SARIMA is summarized by (p,q,r) + s(p,q,r)

ARIMA & SARIMA

مميزاتها:

- قابلة للتعميم
- سرعة في التقارب من النتائج الفعلية
- سهلة التفسير

سلبياتها:

- تتأثر بالحالات الشاذة
- تضمن جميع البيانات في نمط واحد
- تتأثر بالتذبذبات العالية

مثال دارج لتطبيقاتها: توقع حجم الطلبات وتوقعات الأسهم

Markov Chains

المعادلة:

$$P(y_n \mid y_{n-1}, y_{n-1}, \dots) = P(y_n \mid y_{n-1}) = D$$

- احتمالية الانتقال بين حالتين تقاس بالتوزيع D
- تتوقع الحالة الحالية باعتماد على آخر حالة

الافتراضات:

- تغير التوزيع بطيء (لا يحدث بكثرة تغير في النمط)
- الحالة المتوقعة تتأثر فقط بالحالة التي تسبقها مباشرة

Markov Chains

مميزاتها:

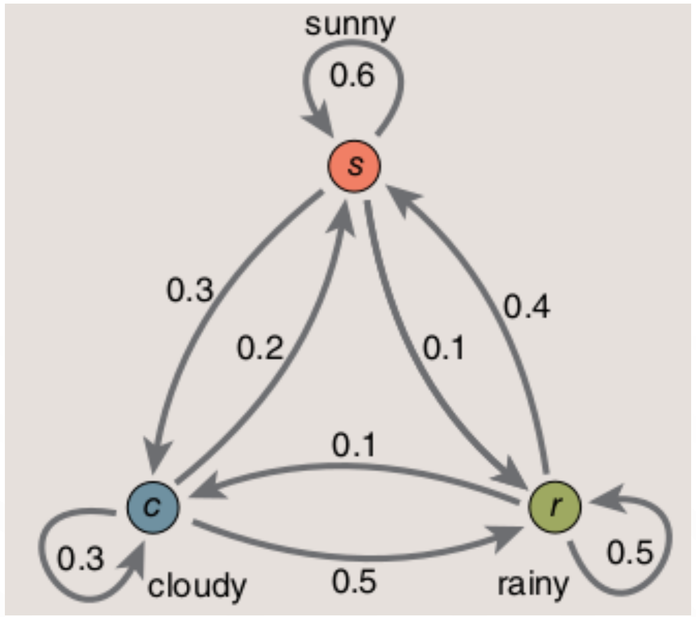
- قابلة للتعميم
- تستخدم في الكثير من الإجراءات المتسلسلة كالألعاب، الأنشطة الاقتصادية ... الخ

سلبياتها:

- بطيئة في استيعاب التغيرات
- تتضمن نمط واحد للحالة السابقة
-

مثال دارج لتطبيقاتها: تحديد حالة الأسواق المالية، التحكم بحركة المرور

Markov Chains



المصدر: https://www.linkedin.com/pulse/weather-forecast-markov-chain-carlos-serra-traynor

للتسليم

مثال لاستخدام Markov Chains مع شرح مبسط للانتقال بين الحالات

أكاديمية طويق
TUWAIQ ACADEMY

استفساراتكم؟ 🤔