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.

هندسة وتحوير المدخلات

ما هي مرحلة هندسة المدخلات؟

هي مرحلة يتم فيها نقل البيانات من شكلها الخام إلى شكل يقدم معلومة أفضل ويخدم الهدف وتساهم هذه المرحلة في رفع دقة النتائج

ما هي مرحلة هندسة المدخلات؟

تجيب هذه المرحلة على الأسئلة التالية:

1. ما هو أفضل تمثيل للبيانات لاستنتاج حل للمشكلة التي نعمل عليها؟

3. كيف نستطيع تحويل المدخلات إلى صورة تستطيع فهمها الناتج المستهدفة؟

أهداف مرحلة هندسة المدخلات

تخدم هذه المرحلة هدفين رئيسيين:

1. تمثيل أفضل للبيانات
3. تحسين أداء النماذج

من أشكال هندسة المدخلات

إعادة ضبط المقياس Scaling

ومن أنواع إعادة ضبط المقياس:

1. Min-Max scaling
2. Standardization
3. Capping (Floor & Ceiling)
4. Quantile Transformers

Min-Max Scaling

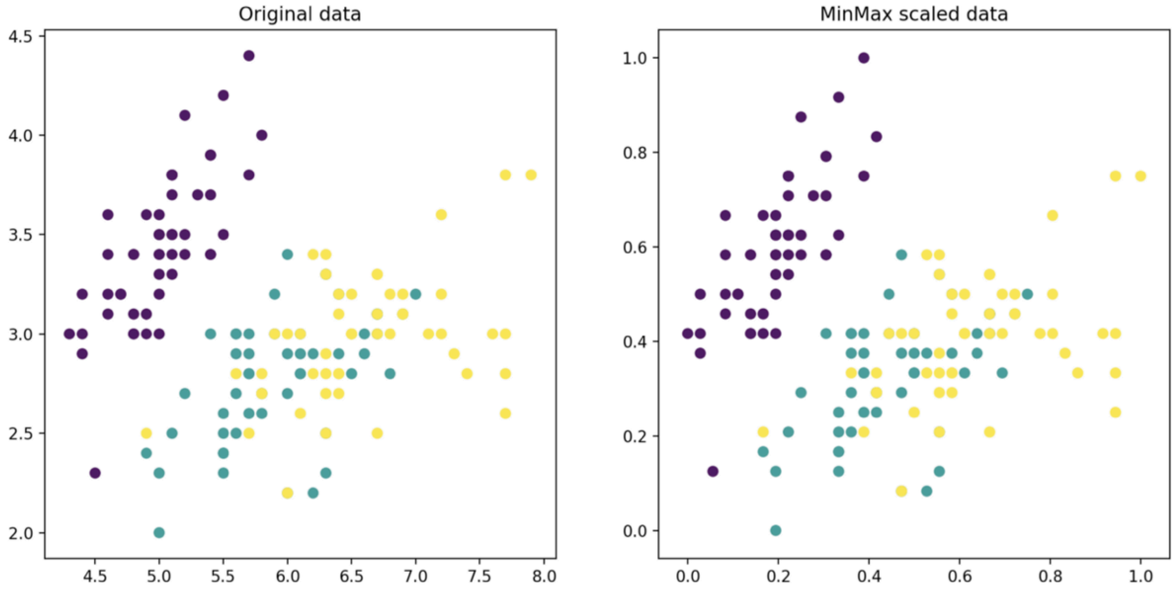
النماذج الهيكلية والشجرية قلّما تتأثر بالمقياس الموحد وما عداها بالغالب يتأثر مثل SVM & LDA

نتيجة الMin-Max Scalar:

- تحويل الخاصية العددية إلى أرقام بين 0 و 1
- تفادي التحيز عن طريق توحيد القياس للخواص العددية

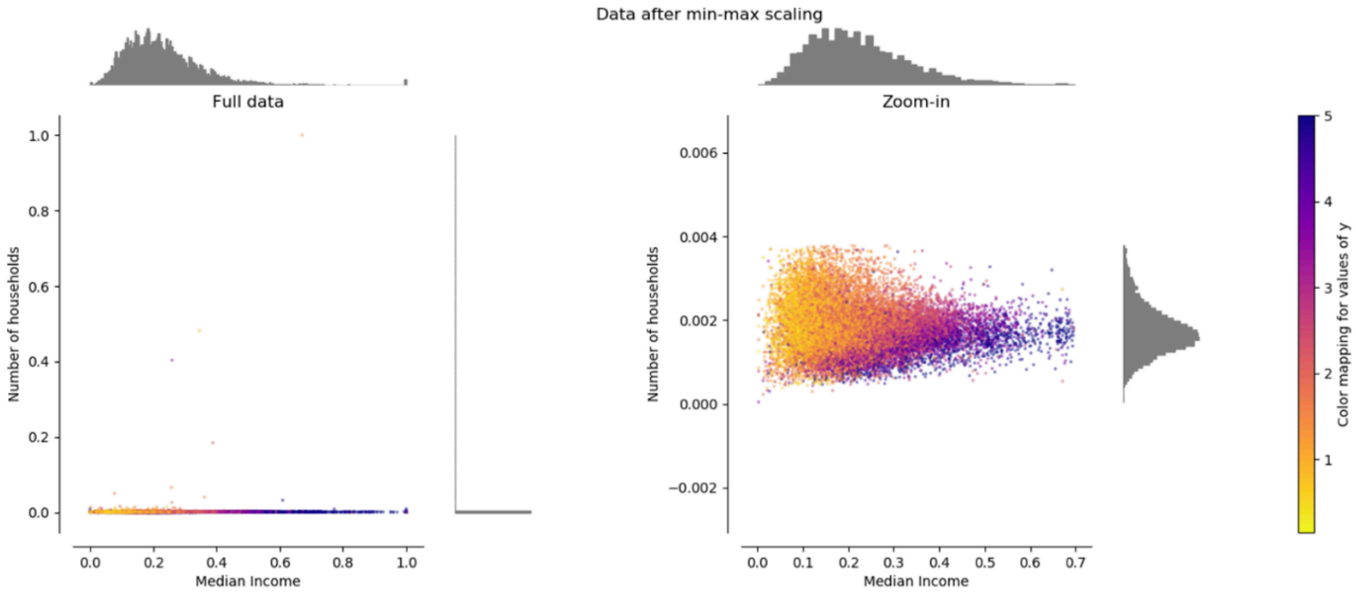
$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-Max Scaling



مثال لمتغير قبل وبعد تطبيق Min-Max Scalar

Min-Max Scaling



مثال يوضح الفائدة من هذه الخطوة لمتغير قبل وبعد تطبيق Min-Max Scalar

المصدر: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

Standardization

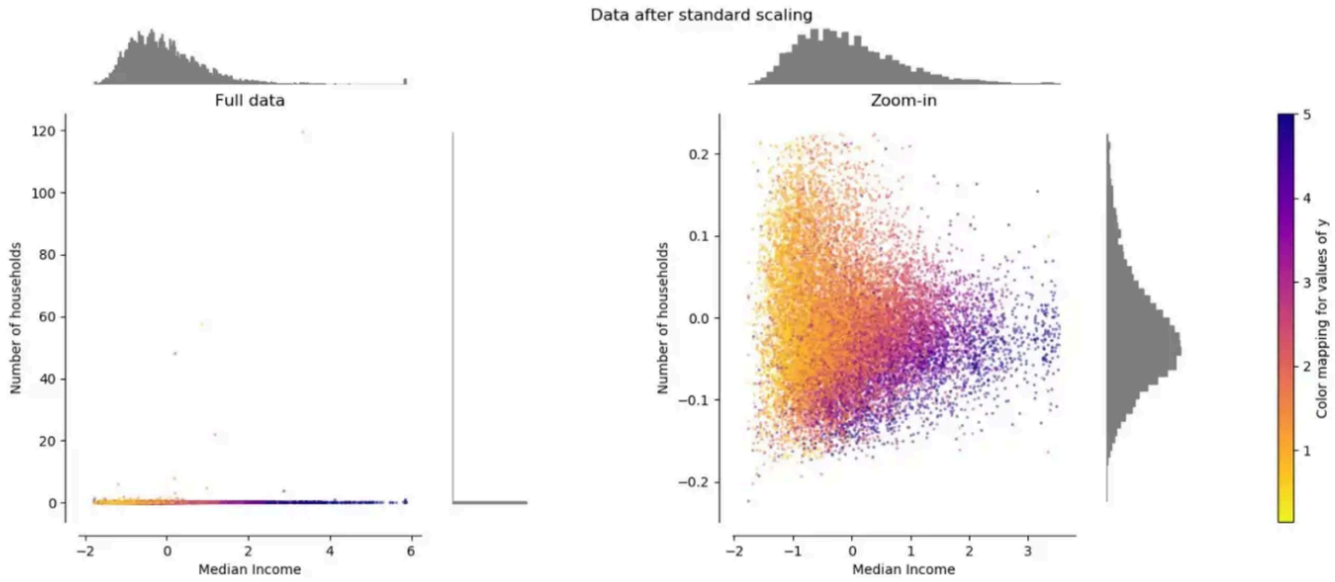
النماذج الهيكلية والشجرية قلّما تتأثر بالمقياس الموحد وما عداها بالغالب يتأثر مثل , & SVM ، LDA. كذلك يتأثر بالحالات الشاذة لأنه يتضمن تقدير المتوسط التجريبي والانحراف المعياري للخواص

نتيجة الStandarization:

- استبعاد المعدل ومن ثم ربطها بالتباين لقيم الخاصية التي ينمي لها
- تفادي التحيز عن طريق توحيد القياس للخواص العددية

$$z = \frac{x - \mu}{\sigma}$$

Standardization



مثال يوضح الفائدة من هذه الخطوة لمتغير قبل وبعد تطبيق Standarization

المصدر: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

التجميع والتقطيع Merging & Disretization

ومن أنواع التجميع والتقطيع:

1. Binning
2. Dimensionality Reduction
3. One-hot encoding
4. Aggregation Functions

Binning

هو ببساطة عملية تحويل الخواص العددية إلى تصنيفات

وتنقسم إلى:

- **تجزئة غير موجهة Unsupervised Binning-** تقطيع الأرقام إلى فئات متساوية
- **تجزئة موجهة Supervised Binning -** تقطيع مبني على الإنتروبيا أو القصور الحراري

Binning

ايش عكس الBinning؟

عكسها يسمى Encoding

Dimensionality Reduction

عملية تقليل المتغيرات العشوائية أو الخواص للحصول على المتغيرات الرئيسية فقط

نلجأ لتقنيات تقليل المتغيرات العشوائية لأن كثرة المتغيرات العشوائية يتسبب في ضعف ضعف أداء نماذج تعلم الآلة

Dimensionality Reduction

من تقنيات تقليل الأبعاد أو تقليل المتغيرات العشوائية:

- اختيار خواص محددة بسبب وجود علاقة قوية لها مع العنصر المُتوقع
- تحليل العنصر الرئيسي

تحليل العنصر الرئيسي PCA

عملية تحليل العناصر الرئيسية هي عملية تعلم غير موجهة تقوم بـ:

- تحسب العلاقة بين الخواص
- تحدد الخواص المتماثلة بالتأثير وتستهدفها للاستبعاد
- تتضمن تحويل مجموعة من المتغيرات العشوائية إلى مجموعات عشوائية جديدة تسمى العناصر الرئيسية Principal Components

تحليل العنصر الرئيسي PCA

افتراضات تحليل العنصر الرئيسية

- تفترض وجود علاقة خطية بين المتغيرات
- تفترض أن المتغير الرئيسي ذو التباين القليل هي متغيرات ضوضائية يتم الاستغناء عنها
- جميع المتغيرات لها نسب قياس متقاربة
- تفترض أنه تم استبعاد القيم الشاذة

مرجع لجميع أفكار هندسة الخواص لتجهيزها لنماذج تعلم الآلة

https://github.com/alicezheng/feature-engineering-book

للتسليم

أنواع الترميز:

1. Label encoding
2. Ordinal encoding
3. Frequency encoding
4. Binary encoding
5. One-hot encoding
6. Target Mean encoding

المقارنة بين أنواع الترميزات من حيث:

طريقة الترميز، مثال للترميز، فرضيات نوع الترميز، حالات لا يناسب فيها استخدام نوع الترميز

أكاديمية طويق
TUWAIQ ACADEMY

استفساراتكم؟ 🤔