

Subject:

بنای خطا

Date:

دانشگاه علم و صنعت
آموزشگاه مهندسی

(Data Cleaning)

تمرینات بینویسی

نام و نام خانوادگی: زهرا در خوار / صدیق رفai زاده / پریم سعیدی

واحد درس: زیبایی داد و پرداز

رست: مهندس کامپیوترا

مرتب: آقای حسن احمدزاده

۱۴۰۳ قمری

Subject :

Date :

1- جلسة دروس Data Cleaning

بيانات دروس Data Cleaning

نتائج تقييم نباتات مائية حسب ما درس في البان

بيانات بسترة سرعة وكمية الماء الماء بالطبع

بيانات الارض بتنمية بسترة حفاظ

بيانات كثافة جذور وعمقها واستهلاكها للفحص والقياس

Subject:

Date:

Missing values - ۱

دریج مقدار نهاده:

۱- حذف: سطوح ارزش رکنده دارند یا استون های با مقادیر کمتره زیاد خنثی شوند.

۲- جایزیت:

مقدار نهاده: جایزیت باشد و معتبر باشد.

میانگین/ میانه / حد: جایزیت با میانگین / میانه باشد استو.

۳- تخفیض مقادیر: جایزیت نهاده از مقادیر موجود:

۴- تغییر مدل:

هر کسیوں / نزدیک ترین مسایعه: استفاده از مدل برای پنهان کردن مقادیر کمتره.

۵- الگوریتم مارکینه: MICE

۶- رویکرد خاص الگوریتم: برخی الگوریتم ها خود مقدار کمتره را بایزیز می کنند.

نکات عمده:

۷- مدل آنالوگیک مدل (MNAR, MCAR, MAR, MAR)

۸- مسازه رویکرد استفاده شده:

۹- ارزیابی تأثیر بر نتایج تحلیل.

Subject:

Date:

Outliers

داده های خارجی (Outlier) مقادیری هستند که باقی داده ها متفاوت هستند.

چند روش برای تشخیص این داده های خارجی وجود دارد. روش های تشخیص شون عبارتند از:

واماده (نمونه قابل قبول است) و غایب از حوزه که در داده های پرداخته شده بروز نموده است!

Subject:

Date:

دیتا ترنسفورمیشن (Data Transformation)

تبدیل دادهها (Data Transformation) به دلایل از زیر کاربرد دارد.

ابهبود کیفیت دادهها:

- از این دادهها نافع باشند.

- رفع ناسازگاری های منابع متنابع.

۲- آسانسازی برای تعلم و درآمد سازی:

- افزایش دقت معلمات ابزاری الگوریتم های مقیاس پذیر.

- تسريع هکثر از درآمد الگوریتم های دیکشناریستی.

۳- سهولت تفسیر و تجییم:

- عملیات دادهها در فرم های آسان برای فهمتر.

- بیبود تجییم دادهها برای تحلیل بصری.

۴- حفاظت از حریم خصوصی:

- بنهایت سازی اطلاعات حساس

- عماقیت خطا رئیسی افراد.

در یک تبدیل دادهها که در کنترل کیفیت و ارزش دادهها افزایش یابد و تغییل های منابع است

منیر شوند

Subject:

Date:

جیف نقاوی مارینا (Label Encoding), Encoding Techniques (one-Hot Encoding - d)

: Label Encoding:

- بہترین طریقہ کوڈنگ (category) یہ کوڈنگ ہے۔
- بہترین طریقہ کوڈنگ (label) "خیل رافٹ"، "رافٹ"، "نا رافٹ" فوریہ۔
- مسادی میں اعلیٰ سطح مکانیکی الگوریتم ہا مرکوز کرنے و ترتیب انتباہ رہیا دیکھیں۔

: one-Hot Encoding

- بہترین طریقہ کوڈنگ (label) جدید درسیں کیا کرو۔
- بہترین طریقہ کوڈنگ (label) قدرتی "آئی اے سیز" فوریہ۔
- از کلیج سوچنے والے الگوریتم ہا جاوا سیرکل ہے اور وہ مکانیکی تعداد سمعنے ھا زیاد بسی۔

: انتباہ

ماں اپنے ترتیب دے سکتے ہیں : Label Encoding
 ماں اپنے ترتیب دے سکتے ہیں : one-Hot Encoding

Subject:

Date:

Model-building and feature selection

انتخاب ورک (Feature Selection)

مدل ریاضی برای این بروای با خفت و پر (افزایش).

مدل ریاضی که در روش کنده (مدل ساده تر) است.

روکم کنده (مدل پترید) overfitting (معکوس).

مسرع آزمون رو زیاد کنند (سریع پترید) (کم).

قابلیت تعمیم برای این بروای (روج داده های جدید هم خوب باشند).

تفصیل نیزه رو زیاد کنند (معکوس کرد) ورک (همچو).

چنان که فواد (متضمن) برای (متضمن) (متضمن).

از "تفصیل ابعاد" جلوگیری کنند.

برای مدل های سریع تر قابل غنیمت سازیم.

Subject:

Date:

Duplicate Data - حکم داده های تکراری در پایگاه داده ها

برای حذف داده های تکراری در پایگاه داده:

1) $\text{ROW_NUMBER} \text{ BY }$, $\text{HAVING } \text{SELECT } \text{DISTINCT } \text{GROUP}$: این روش برای حذف داده های تکراری استفاده کن.

2) حذف داده های تکراری در SQL Server: DELETE با ROW_NUMBER .

روز دیگر: به جدول جدید سازی داده های غیر تکراری و مبتدا کن. جدول اصلی را حذف کن.
اعمال جدید جدید را عوض کن.

وضعیت انتها که در داده جدید کن (در MySQL) تابع IGNORE مانند است.

نتایج:

دقیقی بعد از بارگذاری

استون های ناید و درست مسح شوند کن.

بهره از این کن

مانع جبری حجم داده های بیشین روی را اختاب کن.

Subject:

Date:

Artificial Machine Learning Classifications Irrelevant Data ^

داده های آنرا در مدل آن را از داده های آن خارج کرده و مدل را با آن داده های خارج شده آزموده ایم. این روش Overfitting را در مدل آنرا کاهش می دهد.

دیگر مکانات آموزشی که اندیشه کنندگان در آنها می‌سینند

خلاقه ایش اینه کرداده های بدر نخور کلا پیوژه رو خواهی عکس لندز.

Lesson 10 / Missing values and Data Imputation 9

فیل از الگوریتم های ادیده ای ناقص کار نماین. جنف داده های ناقص همچنان تایج و خطا

نحوه مجموعه داده IMPUTATION Data از پایگاه داده های آزمایشی

لِيَقْرَأُونَ الْكِتَابَ مَنْ يَشَاءُ وَمَا يَنْهَا إِلَّا أَنْفُسُهُمْ وَاللَّهُ عَزَّ ذِيْلَهُ عَلَىٰ كُلِّ شَيْءٍ

Subject:

Date:

وایکلستون میتواند نرمالیتی Normality را در داده های انتشاری بررسی کند.

برای بررسی نرمال بودن داده های عددی می توان از زنگولارها (میتوکنام Q-Q Plot) استفاده کن.

که نوزیم و نسون (Nemenyi-Nemenyi) که بسته میکنند داده های آماری (منفی شناسی و میانگین) را بررسی میکنند.

حد مرزی نرمال بودن تردیدی است که آن را بازیابی نرمال باشند و P-value را تصریح نمایند.

اگر باشد، جواب داشته باشند آن را تصریح میکنند.