

«بے نام خدا»

دریں: میجا جنت و پیرہ

اسٹار: محمد احمدزادہ

رسنے: مہندس ناظمیوٹر

دانشجو: زهراء بیگم امل - زهراء عالمی

دانشجو: ملک عمارت آفروزشہ مہینا

۱۲ بھونج ماه

صفات

«فهرست»

«لیست»

53

پر کامولڈم دار جیز A Data Cleaning

53, 54

چیزیں کو شووندے ہیں Missing Value B

54, 55

چیزیں ویکوئے فورانیں کو اقتصر کر سکتے ہیں Outliers C

55

چیزیں کا باریک دار جیز Data Transformation D

56

? (One-hot Encoding Techniques) one-Hot Encoding E

57

? پر کامولڈم دار جیز Missing Values, Feature selection F

58, 59

? پر کامولڈم دار جیز Model building, Irrelevant Data H

58

چیزیں دریافت کاں کو شووندے ہیں Duplicate Data G

59, 60

? پر کامولڈم دار جیز Missing Values, Data Imputation I

60

? پر کامولڈم دار جیز Mandegar, Normality ijtibai' J

« Data PreProcessing » « پیش‌پردازی داده »

A چرا Data cleaning در علم داده اهمیت دارد؟ پایان‌سازی Data cleaning

داده‌ها این از مدخل حیاتی دارند که علم داده بتواند فوایدی داشته باشد.

وقت و صفت تابعی: داده‌های نادرست را ناقص می‌نامند و متناسب با توابع مخلوط است، اینها ممکن است مدل‌های تحلیل ویژگی‌ها را از داشت.

***غیرهمogen**: فنیک شونده پایان‌سازی داده‌هاست که مدل‌های تحلیل ویژگی‌ها را بر این اندیشه می‌بینند.

***ویژگی تحلیل دقیق‌تری**: فناوری شونده اینها را در مدل‌های تحلیل ویژگی‌ها از داشت.

***کاهش هزینه**: داده‌های نادرست را تولید نمی‌کنند و هزینه‌های انتقال و حفظ را کاهش می‌نمایند.

شوند پایان‌سازی داده‌ها را مرتباً این هزینه کاهش دارند.

***تعیین دریجی و تحلیل**: داده‌های تمیز و مرتب تجزیه و تحلیل و استخراج نیشتند.

ساده‌ترین نتیجه این توالی سیاست بعتری اتخاذ شود.

B Missing Values چگونه صدایت و می‌شوند؟ مدیریت مقادیر ناشی از جالتش

های عدم درکیل را داده‌ها و یادگیری مانند است. درین حالی فنیک شونده پایان‌سازی داده‌ها

مقادیر ناشی از وجود داده‌ها بسته به نوع را در ویژگی خالص پیروزه را تواند اینها را شوند.

حذف داده‌ها:

***حذف ریفی:** این تعداد ریفی داده‌ای است که مقادیر ناشی از عدم باشد و توالی این را ندارند.

های از حذف نمود.

حذف اینترس ها: آنری دیترس (سترن) پیشتر از حد معمول مقاییر نسبت داشته باشد و قول آن را اینترس را لازم بگوید در اینجا همان حذف شده است
میان / میان / حد: منوال از میان میان و حد برای پیرکل مقاییر نسبت داشته باشد

چایزین با مقادیر ثابت: منوال با مقاییر تابع برای مقاییر نسبت در فلکن میان
میان های پیش زدن: منوال از مدل های یادبری مانند برای پیش زدن مقاییر نسبت
 براساس مایه ویژه ها استفاده شود.

کوتلر Outliers چیست و چونیه فر تولید آن را تشخیص دهید؟
 کوتلر این اتفاقات میشود که بطور قابل توجه از سایر مقادیر داده دری دیگر را در فلکن دارند
 این مقادیر ممکن است به دلایل مختلف ابعاد شوند از جمله خطا در پیغام آوری داده
 ها تغییرات طبیعی در داده ها یا وجود پریره های طبیعی در داده ها یا تأثیر بر داده ها
*** نمودار های بصیری ...**

نمودار جعبه ای این نمودار فر تولید راهنمای مقادیر پیرت را تاسیس نمود. مقادیر خارج از
 محدوده های جعبه ای تحریر دارند و به عنوان مقادیر پیرت در فلکن نسبت فرموده شوند
نمودار پراپرتر: این نمودار فر تولید روابط بین دو متغیر را نشان می دهد و مقادیر پیرت را
 رابطه و نویح مشخص کند.

میان های یادبری مانند: بفرض از الگوریتم های یادبری مانند DBSCAN از تحلیل پیرت به نارونز.

تغییر زمانی سری: در راههای زمانی متوال با استفاده از راههای پیش‌بینی و
و متأخره از افادت غیر محدود، تفکل بین راستا مسیر تردد
تشیعی و مدیریت تفکل بین متوالیهای تغییراتی که در

چیزیابی در داده‌ها \rightarrow کیفیت مداخل تبدیل در فرآیند تغییرات
تغییراتی دارمهای یادگیری ماشین است این فرآیند شامل تغییراتی تغییراتی دارمهای به شکل
مناسب برای استفاده در الگوریتم‌های مختلف است.

۱) آنالیز با الگوریتم‌ها:

سیاری از الگوریتم‌های یادگیری ماشین مانند روشی محاسبه شبکه‌های عصبی پنهان
دلیل وروی های عددی نیاز دارند به عنوان دلیل دارمهای متغیر یادگیری باشند فرم
کردنی سریل شوند

۲) پیروزی دقت: با تغییر دارمهای به شکل مناسب متوال به بودجه و موارد
محلی های یادگیری ماشین سریل برقراری می‌نمایند اطلاعات مقیدی از آن
دهند.

۳) ناهمش ابعاد: بین این دو دسته ناهمش ابعاد دارمهای سریل به
نوبه خود باعث ناهمش زوایا می‌گردند و می‌توانند مدل مناسب شوند.

۴) محدودیت راههای تغییرات: در فرآیند \rightarrow transform متوال به نوعه مدیریت دارمهای
نمتشه پیروزی داشت و آن‌ها را به شکل مناسب جایگزین کرد.

چه Label Encoding و Encoding Techniques (one-Hot Encoding) هستند؟

Label Encoding: هر دسته بینی عددی مفهوم به فرد تبدیل می‌شوند. اینها

مثال آنر دیتاشیت ای شامل دسته "قهوه" "سبز" پیشی باشد. من توانم همانها را تبدیل به ۰ و ۱ و ۲ نیز می‌کنم.

مزایا: ساده و سریع است. در برآوردهای آنرا را ساده‌سازی کرد.

مکانی: این روش در تولید نتایج به اینجاد رابطه‌های ناگزین می‌پرسد. دسته‌ها شرکت‌زیرا مرل (همی‌خانه‌بری) همچنان است. نتیجه این است که در مجموع ای ترتیب و فاصله میان اندیادهای مختلف این روش را نمی‌توان به اینجا زد.

one-Hot Encoding: هر دسته بینی بیانی برای تبدیل می‌شود. مثلاً این

بیانی با تعداد دسته‌ها است. فقط یکی مفهوم در این بیانی ۱ است و بقیه ۰ هستند.

مزایا: این روش از اینجاد رابطه‌های ناگزین جلوگیری می‌کند و برای مرل (همی‌خانه‌بری) آنرا ترتیب عددی حساس نمی‌سازد.

مکانی: من توانم به فناوری ستری نیازداشت باشند. هنوز هم این تعداد دسته‌ها زیاد باشند، ممکن است صنایع مشکلات سوزنده

2.1.1 Causal Model building \rightarrow Feature selection \rightarrow F

در فراغت انتخاب و پیروزی در دارووله بر مختلف بار این
اهیت وجود دارد.

نامه ایجاد با انتخاب و تیر هار مناس سوتان ایجاد داده مار آنهاش
دار این خارج ناهمز ز هان حسابات و مبالغ مورد نیاز ببار آوردهش علا
عکس و نمود.

ببور وقت امیر: با انتخاب او نیز هار مرتبه و حنف و نیز هار غیر خود
و توان صدر را افزایش داد و نیز هار غیر خود یا سارر متن اینست باشد و نیز
در راه ها سونه وقت ایش بین راماهش رهند.

تسلیل تفسیر عدل: انتخاب و پرور هار کسب و کار را از دید به تفسیر پیش عدال
اعضای هنر و فرهنگ نهاده و پرور ها همچنان مجاز است (علیل و تفسیر نتایج آسان) تدریس
خواهد بود.

جیونه در پایه داده حذف از شود Duplicated Data

حذف داده هارس ار در پایه داده های از وظایف هم معتبر ندارد.
بر عین داده ها سایر رسان از روشن هار مختار استفاده کنند در این به عنوان
روش حذف اشاره می شوند.

نیازمند سازر قیمت هار بگیر : بر این جا میگیر از ورد داده هارس ار به پایه داده
مرتفع از نفس هار بگیر این میان این داده های داده هار خود را از
وورد داده هارس ار میگیر می شوند.

استفاده از اسپیت هار برنامه نویس در بفر صادر حصل اس نیاز باش از
تبیان هار برنامه نویس مانند java، Python استفاده نمی تاره هارس ار
را استایل و خذف نماین روشن در پروژه هار بگیر ترک نیاز
بیدار نمی شوند هار بجهیه تر داشت فیلم باشد.

جیه میله های را دریش بین های Machine learning - H الجیاد فیلم

نامه دقت دارد : داده هار نامه بعده از تواند باعث کاهش دقت دارد
نمی شوند. این داده ها از تواند بین این تغییر دارد نمی شوند و باعث کاهش دارد
الجیاد را افزایش داده هار را میگیرد.

افزارهای زمان آموزش ز داده هار نا صریع ملود توانند زمان لازم برای آموزش
عمل را افزایش دهنند زیرا بعد از داده هار افزایش نیز پرداخت
نمی شوند.

مسئله در انتخاب وینتیر: انتخاب وینتیر هار نا مناسب بایان اهم برای اهم
توانیده باشد زیرا انتخاب وینتیر آسیب بر ساندویچ معتبر به انتخاب وینتیر
هار نا مناسب بایان اهم برای اهم شود.

۱- جراحتی برای پردازش missing values Data Imputation

کی تئینی داده هار یا دیگر عوامل است که به دلایل زیر تاکه بردار

جلوی پردازبروز خواهد: بیمار را از آنکه رسم هار یاد نمی شود مانند
داده هار ناقص کار نمی کند این قابلیت پردازبروز خواهد بود هار قابل با
داده هار ناقص جلوی پرداز خواهد.

۱

نهایت بیتر: در تحلیل راه های مقادیر گمشده مرتباً از نتایج تحلیل راهنمای تأثیر
قرار دهد هنر با پردازین این مقادیر مرتباً نتایج بینزین را درست آور در تحلیل
هار دقیق تر اخبار دارد.

بی ملور کار بجهت پیشگیری از این و راهنمایان داده شده فرستاده ها را ناقص
بقدرت کار نمود و به نتایج دقیق تر رسیده است یا نباید.

لذ: چونکه می توانیم Normality را در طبقه های عده ببررسی کنیم

برای بررسی مالیتی داده ها در مورد تراویث از روین ها فنی از استادهای زیر
در زیر بفرز از این روشن ها اسما را فرموده ام:

صفحه اول: بارسونی: مستوی از داده های توزیع داده های افکارهای
که در آن داده های بدل تسلیم شده و مقایسه حواله میانهای توزیع شده با انتظاهات
نرمال هستند.

صفحه دوم: این صفحه ایجاد شده تا ناهنجاریها و توزیع داده ها
را بررسی کرد در صورت وجود داده های دورافتاده یا جبع از این داده ها
ظرفیت این داده های محدود می شوند است از این داده های محدود