تمرین سری ۱ واحد درسی یادگیری ماشین

جناب آقای دکتر فراهانی
دستیار آموزشی : علی شریفی

۲۸ اسفند ۱۴۰۰

پیشاپیش سال نو بر شما مبارک باد. با امید سالی سرشار از سلامتی و برکت و شادی.
یا مقلب القلوب و الابصار، یا مدبر اللیل و النهار، یا محول الحول و الاحوال، حول حالنا الی احسن الحال
ای تغییر دهنده دلها و دیده ها، ای مدبر شب و روز، ای گرداننده سال و حالت ها، بگردان حال ما را به نیکوترین حال

توجه کنید شما میتوانید بر روی کگل یا کولب و یا کامپیوتر های شخصی خود کار کنید .
به جای دانلود و آپلود دیتاست در گوگل درایو برای استفاده در کولب میتوانید به شیوه زیر عمل کنید .
چگونه از دیتاست های کگل در کولب استفاده کنیم ؟
ددلاین تمرین تا ۲۰ فروردین ۱۴۰۱ می باشد.
نحوه تحویل پاسخ تمرین ها در ریپازیتوری متعلق به درس می باشد.

۱ تمرین

۱.۱ دیتاست شماره ۱

در این تمرین از دیتاست اطلاعات مشخصات و قیمت گوشی همراه استفاده شده است. لینک دیتاست از شما خواسته میشود به تسک های زیر را انجام دهید.

۱.۱.۱ تسک های اصلی

۱. پاکسازی داده از قبیل بررسی داده های از غیرموجود و یافتن داده های پرت.

۲. ارایه اطلاعات کلی در حالت تجمیعی در خصوص بررسی فیچرها و اطلاعات آماری مربوط به آنها . لازم است حتما در این بخش از بحث مصورسازی داده ها استفاده کنید و این اشکال ایجاد شده را تفسیر کنید.

۳. مطرح کردن ۵ آزمون فرض دلخواه در داده ها و پاسخ گویی و تفسیر آنها (حداقل از ۳ آزمون فرض متفاوت استفاده کنید.)

۴. ارایه مدل های کلاسیفایر برای پیش بینی بازه قیمتی (کلاس قیمت) گوشی های همراه برای فیچرهای موجود. لازم است که برای مدل های مختلف حداقل سه مجموعه پارامتر مختلف بررسی شود. بررسی کنید مدل مورد استفاده شده از استراتژی OVO استفاده کرده است و یا استراتژی OVA .

۵. بررسی نتایج کلاس بندی مدل با استفاده از confusion matrix . آیا در همه کلاس ها به میزان یکسانی مدل کلاسیفایر شما عمل کرده است ؟ اگر خیر لطفا بررسی کنید چرا ؟

۶. بررسی کنید آیا داده ها متوازن می باشند ؟ در صورتی که داده ها متوازن نباشند سه راه حل مطرح کنید ؟

۷. برروی داده ها از scaling ها مختلف استفاده کنید آیا متد های مختلف scaling تاثیری در نتایج دارند ؟ مقایسه انجام دهید.

۸. داده ها را به نسبت ۸۰ به ۲۰ برای تست جدا کنید و نتایج را روی داده های تست گزارش دهید .

۹. با استفاده از روش PCA با pov های مختلف 0.75،0.8،0.9،0.95،0.99 تعداد فیچرها را کاهش دهید. آیا این روش باعث بهبود نتایج شده است . اگر نه چرا از این روش استفاده میکنیم ؟

۱۰. لیبل کلاس های ۲و۳ و۴ را تغییر دهید و همه رکورد های مربوط به این کلاس ها را تبدیل به کلاس ۵ کنید. داده ها نامتوازن میشوند . آیا این عدم توازن تاثیری در مدل ها داشت ؟یکی از راه حل های ارایه شده در پرسش ۶ را به دلخواه انتخاب کرده و بر روی داده ها اعمال کنید و مجددا نتایج مدل را بررسی کنید.

۲.۱.۱ تسک های امتیازی

این تسک ها، فرای تسک های اصلی می باشد و پاسخ گویی به آنها دارای نمره امتیازی می باشد.

۱. استفاده از dask یا pyspark در بخش پیش پردازش داده ها و بخش آموزش مدل ها.

۲.۱ دیتاست شماره ۲

در این تمرین از دیتاست از آگهی های استخراج شده از یکی از بزرگترین پلتفرم های املاک کشور آلمان استفاده میشود. لینک دیتاست

۱.۲.۱ تسک های اصلی

۱. پاکسازی داده از قبیل بررسی داده های از غیرموجود و یافتن داده های پرت . بررسی موارد مشابهت [۱]

۲. ارایه اطلاعات تجمیعی از تعداد آگهی و پارامترهای مختلف آگهی از قبیل تعداد آگهی ها در مناطق مختلف جغرافیایی ، اطلاعات در خصوص تعداد انواع خانه ها ، بررسی قیمت در مناطق مختلف جغرافیایی و ... . لازم است حتما در این بخش از بحث مصورسازی داده ها استفاده کنید و این اشکال ایجاد شده را تفسیر کنید.

۳. تلاش جهت مدل سازی قیمت ها بر اساس پارامتر های مختلف آگهی.

۴. استفاده از بحث multiprocessing در بخش پاکسازی و پیش پردازش داده ها و و بررسی runtime فرایندها.

۵. استفاده از dask و pyspark در بخش پاکسازی و پیش پردازش داده ها و بررسی runtime فرایندها.

۲.۲.۱ تسک های امتیازی

۱. استفاده از مهندسی ویژگی در جهت بهبود مدل ها .

۲. استفاده از dask و pyspark در بخش مدلسازی داده ها و مقایسه با حالت عدم استفاده از آنها.

[1] duplicate

۲ نحوه ارسال

تمامی تمرین ها طبق نحوه بیان شده در کلاس های حل تمرین در داخل گیت هاب تحویل گرفته میشوند.
دانشجویان گرامی نهایتا تا ۲۱ فروردین ماه فرصت دارند تا کارهای خود را موفق در گیت هاب قرار دهند.
با توجه به فرصت سه هفته ای قبلی جهت تمرین در کار کردن با گیت و بررسی ارسال ها در گیت هاب و بیان موارد موفق و غیر موفق، ارسال غیرموفق طبق دستورالعمل گفته به منزله عدم ارسال تمرین در نظر گرفته میشود.