مبانی یادگیری ماشین - تکلیف سری دوم

مدرس: دکتر حامد ملک

پاییز ۱۴۰۱

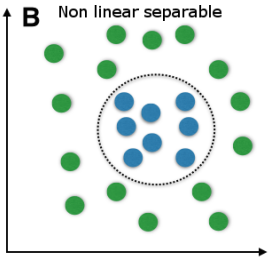
ددلاین: دهم آذر

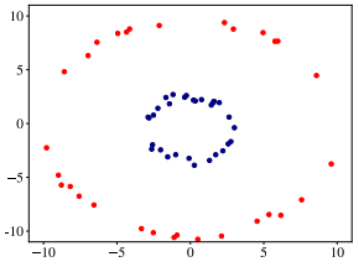
مسائل تحلیلی

۱. چگونه می‌توان اطمینان حاصل کرد که مدل رگرسیون خطی شما خوب تعمیم‌پذیر است و روی داده‌های جدید نیز خوب عمل می‌کند؟

۲. چگونه می‌توان از مدل‌های دسته‌بندی دودویی برای تسک طبقه‌بندی چندگانه استفاده کرد؟ دو روش متداول برای این شیوه استفاده را توضیح دهید. (راهنمایی: one-vs-one, one-vs-rest)

۳. چرا نمیتوان از الگوریتم های طبقه‌بندی خطی برای داده های موجود در تصویر زیر استفاده کرد؟





۴.برای ساخت یک مدل موثر در مواجهه با داده‌های نامتوازن چه راهکارهایی را پیشنهاد می‌دهید؟ چگونه می‌توان با این نوع داده‌ها بهترین عملکرد را از مدل خود حاصل کرد و از عملکرد آن اطمینان داشت؟

مسائل کدی

توجه: قسمت اعظمی از امتیاز این تمرین به حوزه مهندسی داده اختصاص یافته است، بنابراین توصیه می‌شود که با بهره‌گیری از روش‌های مناسب مانند کدبندی وان هات، binning، نرمالسازی داده‌ها، روش‌های هندل کردن داده‌های نامتوازن و انتخاب معیار ارزیابی مناسب برای مدلهای خود و دیگر رویکردهای مرتبط، در انجام این وظیفه پیشروی کنید.

۱.سکته مغزی، همچنین به عنوان حادثه عروق مغزی یا CVA نیز شناخته می‌شود، زمانی رخ می‌دهد که یک قسمت از مغز از تأمین خون خود محروم شده و قسمتی از بدن که سلول‌های مغزی محروم از خون آن را کنترل می‌کنند، کار خود را متوقف می‌کند. این از دست دادن تأمین خون می‌تواند به دلیل کمبود جریان خون یا به دلیل خونریزی در بافت مغزی باشد. سکته مغزی یک اورژانس پزشکی است زیرا ممکن است منجر به مرگ یا ناتوانی دائمی شود. امکاناتی برای درمان این نوع سکته‌ها وجود دارد، اما این درمان باید در چند ساعت اولیه پس از ظهور نشانه‌های سکته آغاز شود.

مجموعه داده strokes.csv شامل اطلاعات افراد و سابقه سکته مغزی آنان می‌باشد.در این مسئله مراحل زیر را انجام دهید:

الف) عملیات پیش پردازش را با توجه به هدف مسئله انجام دهید.

ب) داده ها را با نسبت مناسب داده آموزشی و داده تست تقسیم کنید.

پ) با استفاده از پیاده‌سازی آماده الگوریتم SVM در کتابخانه sklearn، مدلی مناسب را آموزش دهید.

۲. در صنعت بیمه سلامت، شرکت های بیمه اغلب در تعیین حق بیمه دقیق برای هر بیمه گذار با چالش هایی مواجه هستند. اشتباهات در ارزیابی ریسک های سلامتی بیمار، می تواند منجر به خسارات مالی قابل توجهی شود. بنابراین، تعیین دقیق حق بیمه سلامت برای حفظ ثبات مالی شرکت های بیمه و ارائه خدمات منصفانه به بیمه شدگان بسیار مهم است.

در این پروژه، ما از مجموعه داده ای حاوی اطلاعاتی در مورد بیمه شدگان بیمه درمانی، از جمله سن، جنسیت، شاخص توده بدنی (BMI)، تعداد فرزندان، عادات سیگار کشیدن، منطقه مسکونی و هزینه های پزشکی فردی که توسط بیمه ارائه می شود، استفاده خواهیم کرد. این مجموعه داده به عنوان یک منبع داده با ارزش برای توسعه مدل‌های پیش‌بینی است که می‌تواند به شرکت‌های بیمه سلامت در ارزیابی خطرات و تعیین حق بیمه دقیق‌تر کمک کند. این پروژه مستقیماً بر استراتژی

عملیاتی و تجاری شرکت‌های بیمه سلامت تأثیر می‌گذارد و در نهایت منافعی را هم برای شرکت‌ها و هم برای بیمه‌گذاران آن‌ها فراهم می‌کند.

اهداف این پروژه توسعه مدل های یادگیری ماشین است که می تواند به شرکت های بیمه سلامت در موارد زیر کمک کند:

- **تعیین دقیق حق بیمه:** از داده‌های بیمه‌گذار برای محاسبه دقیق‌تر حق بیمه بر اساس خطرات سلامتی که هر بیمه‌گر با آن مواجه است، استفاده کنید. در نتیجه، شرکت های بیمه می توانند زیان های مالی ناشی از حق بیمه های نادرست را به حداقل برسانند.
- **ارزیابی مخاطرات سلامتی:** عوامل خطری که بر هزینه‌های پزشکی فردی تأثیر می‌گذارند، مانند سن، BMI، تعداد فرزندان و عادت‌های سیگار کشیدن را شناسایی کنید. این می تواند به شرکت های بیمه کمک کند تا ریسک ها را به طور موثرتری ارزیابی و مدیریت کنند.

مجموعه داده insurance.csv شامل اطلاعات بیمه‌شوندگانی است در یکی از شرکت های بیمه سلامت جمع آوری شده است. در این مسئله مراحل زیر را انجام دهید:

الف) عملیات پیش پردازش را با توجه به هدف مسئله انجام دهید.

ب) داده ها را با نسبت مناسب داده آموزشی و داده تست تقسیم کنید. (دلیل انتخاب درصد نسبت داده آموزشی و تست را گزارش کنید)

پ) مدل رگرسیون خطی را با توجه به مباحث تدریس شده در کلاس درس از پایه پیاده سازی کنید و درصد دقت مدل را روی داده تست گزارش کنید.

ث) مدل قبلی خود را به حالت چندجمله‌ای تعمیم دهید و درصد دقت را گزارش کنید.

نکات تمرین

- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.
- استفاده از زبان غیر از پایتون مجاز نیست.
- فایل تکلیف را به صورت خواسته شده در سامانه کوئرا آپلود کنید.

موفق باشید