تکلیف 5

مقدمه ای بر علم داده

TA(s):محمد جواد بشارتی

مدرسان: دکتر بهرک، دکتر یعقوب زاده

مهلت: سه شنبه،
18اردیبهشت ساعت 23:59

معرفی

در این تکلیف، از شما انتظار می‌رود تکنیک‌های مهندسی ویژگی‌ها را در مجموعه داده‌های مرتبط با فوتبال به کار ببرید تا احتمال به ثمر رساندن گل از طریق ضربه را تجزیه و تحلیل کنید. در مرحله بعد، با اجرای رگرسیون چند متغیره و اعتبارسنجی متقاطع k-foldاز ابتدا به مفاهیم رگرسیون و اعتبار متقابل بیشتر می پردازید و از آنها در مجموعه داده های از پیش پردازش شده مربوط به خودروها استفاده می کنید. در نهایت، شما نتایج خود را با نتایجی که با استفاده از کتابخانه های داخلی پایتون به دست آورده اید، مقایسه خواهید کرد. این کار درک شما از این مفاهیم و اجرای عملی آنها را تقویت می کند.

مجموعه داده مجموعه داده ای که برای قسمت پیش پردازش استفاده می کنید مربوط به داده های فوتبال است .(football.csv)این شامل اطلاعاتی در مورد ضربات مانند زمان، مکان (کرنر، پنالتی و غیره) و نتیجه ضربات (توسط دروازه بان مهار شده، مهار شده توسط مدافعان، از دست رفته و غیره) است. کشف حقایق جالب تر در مورد این مجموعه داده به شما بستگی دارد :)

برای بخش‌های پیاده‌سازی، از یک مجموعه داده کاملاً متفاوت و از پیش پردازش شده استفاده می‌کنید که حاوی اطلاعاتی درباره اتومبیل‌ها (cars.csv)است. شما باید از این مجموعه داده برای آموزش مدل‌های رگرسیون چند متغیره و k-foldسفارشی خود برای پیش‌بینی ستون‌های «قیمت به هزار» و «اسب‌بخار» استفاده کنید.

اشاره

مسئولیت تقسیم داده ها به مجموعه های آموزشی و آزمایشی با شماست، زیرا هیچ داده آزمایشی ارائه نشده است.

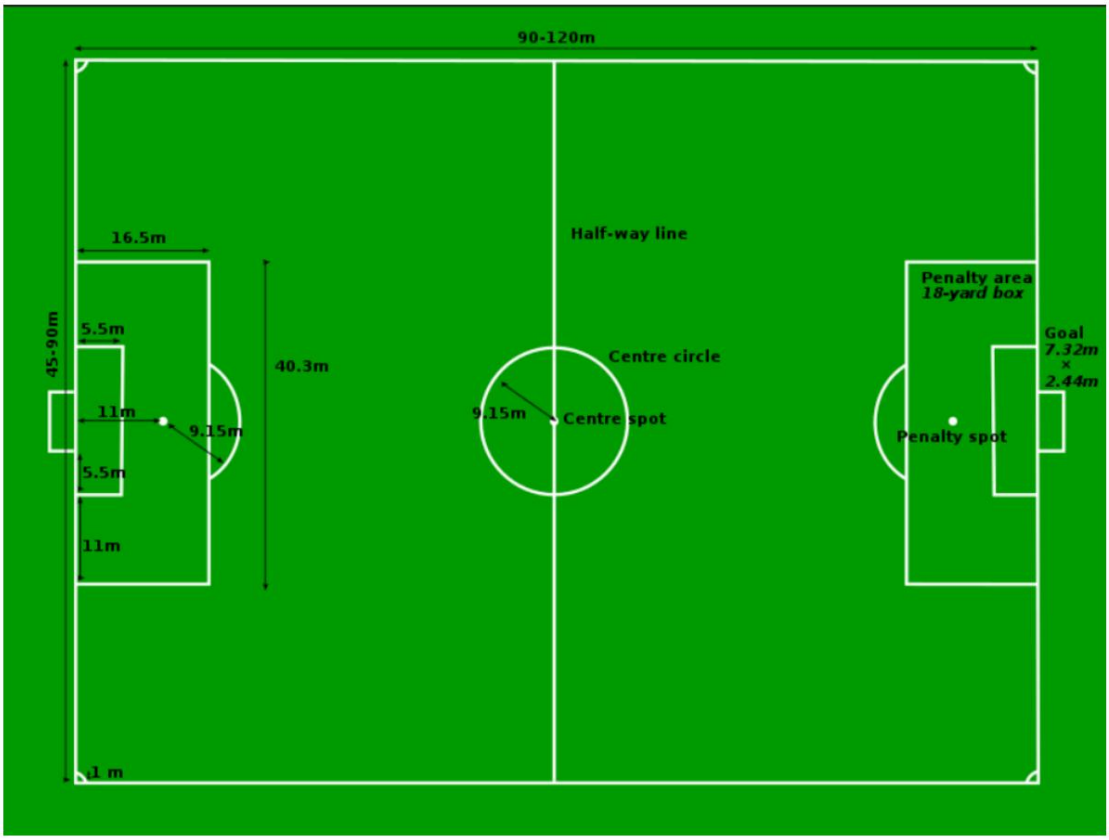
وظایف

.1پیش پردازش

شما وظیفه تمیز کردن و تجزیه و تحلیل مجموعه داده ها، برجسته کردن ویژگی های آماری و تجسم ویژگی های آن را دارید. هدف شما این است که ویژگی های مفید را شناسایی کنید و نتایج خود را به طور قانع کننده توجیه کنید. علاوه بر این، شما باید از تکنیک های مهندسی ویژگی برای اصلاح مجموعه داده استفاده کنید، چه با حذف یا جایگزین کردن ویژگی های کمتر مطلوب. برای به دست آوردن درک عمیق تر از مهندسی ویژگی، آموزش یک مدل دلخواه اما مناسب و ارزیابی نتایج قبل و بعد از پیش پردازش توصیه می شود. علاوه بر این، برای ارزیابی اهمیت هر ویژگی، از روش اطلاعات متقابل برای ایجاد یک چارچوب داده پاندا با دو ستون استفاده کنید: یکی برای ویژگی ها و دیگری برای اهمیت آنها. پس از آن، چارچوب داده را به ترتیب نزولی بر اساس اهمیت مرتب کنید و نتایج را نمایش دهید.

نکات

•برای به دست آوردن بینش بهتر در مورد زمین فوتبال، می توانید به موارد زیر مراجعه کنید تصویر:



•می توانید روش های مختلفی را که در طول دوره یاد گرفته اید برای پر کردن مقادیر از دست رفته و دستکاری ویژگی های طبقه بندی شده در داده های خود به کار ببرید. •می توانید ویژگی های مشابه را ادغام کنید. به عنوان مثال، می توانید "گل" و "گل به خودی" را یکسان در نظر بگیرید. •از انتخاب ویژگی برای حذف ویژگی های کمتر مهم استفاده کنید و در نتیجه کاهش دهید

ابعاد و کاهش هزینه های محاسباتی

•برای تجزیه و تحلیل دقیق تر، استخراج ویژگی های جدید و آموزنده تر از ویژگی های موجود را در نظر بگیرید. به عنوان مثال، فاصله و زاویه شات را با استفاده از فرمول های زیر محاسبه کنید و آنها را در تجزیه و تحلیل خود بگنجانید:

$$= \sqrt{\quad 2}$$

$$angle = \begin{cases} rad2\deg(arctan(\theta)) & \arctan(\theta) \geq 0 \\ rad2\deg(arctan(\theta + \pi)) & \arctan(\theta) < 0 \end{cases}, \theta = \frac{7.32x}{x^2 + y^2 - \left(\frac{7.32}{2}\right)^2}$$

2.پیاده سازی رگرسیون چند متغیره رگرسیون چند متغیره را از ابتدا پیاده سازی کنید و از الگوریتم نزول گرادیان برای به روز رسانی وزن ها استفاده کنید. اعتبار مدل رگرسیون را با ارائه یک مقایسه بصری بین مقادیر پیش‌بینی‌شده و واقعی برای «قیمت به هزار» و «اسب‌بخار» انجام دهید. علاوه بر این، برای تأیید قوی‌تر، دقت را در حالت‌های تصادفی مختلف ترسیم کنید. در نهایت، یک منحنی یادگیری برای نشان دادن پیشرفت فرآیند رگرسیون نمایش دهید.

3.اجرای K-Fold Cross Validation

اعتبار سنجی متقاطع K-Foldرا از ابتدا اجرا کنید. مانند قسمت قبل، از الگوریتم گرادیان نزول برای تنظیم وزن ها استفاده کنید. سپس، اجرای سفارشی K-Foldخود را با استفاده از معیارهای آماری تأیید کنید. در نهایت، پس از اتمام، یک منحنی یادگیری را نمایش دهید.

اشاره

می‌توانید از بخش‌هایی از کدی که در بخش قبل پیاده‌سازی کرده‌اید، در اینجا استفاده کنید.

4.مقایسه با کتابخانه های داخلی پایتون

اکنون، نتایج پیاده سازی های سفارشی خود را در بخش های 2و 3با نتایج به دست آمده با استفاده از کتابخانه های داخلی پایتون مقایسه کنید و یافته ها را گزارش دهید.

سوالات

.1استراتژی خود را برای پرداختن به چالش هایی مانند مدیریت ارزش های از دست رفته و ویژگی های طبقه بندی توصیف کنید.
آیا می توانید در مورد معیارهای انتخاب ویژگی خود نیز توضیح دهید و دلیل آن را توضیح دهید؟

.2چرا از رگرسیون برای پیش بینی اینکه آیا شوت منجر به گل می شود استفاده نکردیم؟

.3چگونه می خواهید صحت فرمول مورد استفاده را تأیید کنید
زاویه شات را در بخش پیش پردازش محاسبه کنید؟

.4در مورد مزایا و معایب اعتبارسنجی متقاطع k-foldبحث کنید. آیا می‌توانید انواع دیگری از روش‌های اعتبارسنجی متقاطع را نیز توضیح دهید که می‌توانند محدودیت‌ها و مسائل مرتبط با اعتبارسنجی متقاطع k-foldرا برطرف کنند؟

.5از چه معیارهایی برای ارزیابی پیاده سازی های دستی خود استفاده کردید
رگرسیون چند متغیره و اعتبارسنجی متقابل ،k-foldو چرا آنها را انتخاب کردید؟

یادداشت

•کار خود را به صورت فایل فشرده در این قالب در وب سایت آپلود کنید: .DS_CA5_[Std number].zip
اگر پروژه به صورت گروهی انجام می شود، تمام شماره دانشجویی اعضای گروه را در نام ذکر کنید.

•اگر پروژه به صورت گروهی انجام شود، فقط یک عضو باید اثر را آپلود کند. •ما کد شما را در حین تحویل پروژه اجرا خواهیم کرد، بنابراین مطمئن شوید که نتایج شما درست است

قابل تکرار