دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

پروژه شاره پنج

کار بر روی دیتا بیس
Bank Account fraud

نگارش

عرفان نجفی

استاد

دکتر مهدی قطعی

استاد دوم

بهنام یوسفی مهر

اردیبهشت ۱۴۰۳

فهرست مطالب:

چکیده:

در این پروژه، از روش‌های مختلف یادگیری ماشین برای شناسایی تقلب استفاده شده است. ما یک دیتا ست داریم که یه سری اطلاعات از مشتریان یک بانک داده شده است . یک ستون داریم که نشان دهنده این است که مشتری دارد تقلب میکند یا نه و ما هدف این را داریم که یا بگیریم یک مشتری جدید که اطلاعاتش را میگیریم ایا تقلب میکند یا نه.

داده‌ها ابتدا پیش‌پردازش شده و سپس با استفاده از تکنیک‌های مختلفی همچون SMOTE، کاهش ابعاد و خوشه‌بندی K-Means تحلیل شده‌اند. همچنین، مدل‌های طبقه‌بندی مختلفی با استفاده از کتابخانه LazyPredict بررسی شده‌اند.

عکسی از پیاده سازی گذاشته نمیشود و در فایل کد بررسی میشود.

بخش اول: پیش پردازش

در پروژه های ماشین لرنینگ وقتی یک دیتاست به ما داده میشود ما باید دیتا ست را به نحوی تغییر دهیم تا برای ماشین قابل فهم باشد و در پردازش ما تاثیر منفی نگذارد به معنای دیگر داده های پرت را حذف کنیم.

کار هایی که باید انجام شود:

۱-پاک کردن خانه هایی که دیتای نال دارند.

۲-پاک کردن خانه هایی که مقدار آن ها بسیار پرت است.

۳-عددی کردن خانه هایی که تعریف غیر عددی دارند.

۴-درست کردن داده های دیتا و داده های تست.

۵-استاندارد سازی و ۵۰ ۵۰ کردن نسبت داده های تقلب به غیر تقلب قبل استفاده از الگوریتم های خوشه بندی(دلیل این است که تحلیل بهتری بتوانیم بکنیم).

حالا در این پروژه از کتابخونه predict lazy استفاده میکنیم.

بخش دوم : خوشه بندی

الگوریتم خوشه بندی یکی از پرکاربردترین الگوریتم‌های حوزه هوش مصنوعی می‌باشد که به صورت گسترده و به منظور گروه‌بندی داده‌ها بر اساس ویژگی‌های آنها به گروه‌های مشابه یا خوشه‌ها، مورد استفاده قرار می‌گیرد. یکی از پرکاربردترین، محبوب‌ترین و ساده‌ترین الگوریتم‌های خوشه بندی K-meansنام دارد. روش کار این الگوریتم به این شکل است که ابتدا تعداد خوشه‌ها به الگوریتم داده می شود، سپس داده‌های ورودی با توجه به فاصله آنها از مرکز خوشه به هر یک از خوشه‌ها اختصاص داده می‌شود.

در این مثال ما از الگوریتم K-means استفاده میکنیم و دیتا را به دو خوشه تبدیل میکنیم و آن را نمایش میدهیم.

اما قبل از این کار ما رو دیتای خود کاهش بعد انجام میدهیم تا بتوانیم سرعت پردازش بی شتر دا شته با شیم و بتوانیم آن را نمایش بدهیم.

پس از انجام این کار ها دقت خوشه بندی خود را حساب میکنیم که کاملا به پیش پردازش و به کاهش بعد ربط دارد.

بخش سوم : الگوریتم KNN

الگوریتم K نزدیک ترین همسایه یا KNNیکی از ساده‌ترین الگوریتم‌های یادگیری ماشین باناظر است که برای حل مسائل طبقه‌بندی و رگرسیون استفاده می‌شود.

این الگوریتم برای مسائل طبقه‌بندی k نزدیک ترین همسایه را پیدا و با اکثریت آرا نزدیک‌ترین همسایگان کلاس را پیش‌بینی می‌کند.

بیایید با هم مراحل این الگوریتم را بررسی کنیم:

۱. داده‌ها را بارگذاری می‌کنیم.
۲. مقدار K را تعیین می‌کنیم که همان تعداد نزدیک‌ترین همسایه‌ها هستند.
۳. برای هر نمونه داده :
    - فاصله‌ی میان نمونه داده‌ی جدید را با نمونه داده‌های موجود محاسبه می‌کنیم.
    - فاصله و شاخص هر نمونه را به یک فهرست وارد می‌کنیم.
۴. کل لیست را براساس فاصله‌ی نمونه داده‌ها، از کمترین به بیشترین فاصله، مرتب می‌کنیم.
۵. Kتا از اولین نمونه‌های فهرست مرتب‌شده را به‌عنوان K نزدیک‌ترین همسایه انتخاب می‌کنیم.
۶. برچسب این K نمونه را بررسی می‌کنیم.
۷. نمونه‌ی جدید هم همان برچسب K همسایه را خواهد داشت.

بخش چهارم: استفاده از predict lazy

پردیکت لیزی کاری که برای ما انجام میدهد این است که چهار پارامتر به عنوان ورودی میگیرد که که به x و y های داده های تست و ترین است و ۲۹ الگوریتم را روی آن ها پیاده سازی میکند همه آن هارا از نظر بهینگی و پارامتر های دیگر بررسی میکند.

برای اینکه این کار لود کاری بالایی دارد و هر سیستمی توانایی اجرا این را در زمان معقول ندارد ما دیتا ست را به صورت کوچک تر میدهیم برای پردازش.

بخش پنجم : cross validation

اعتبارسنجی متقاطع (CrossValidation) یک تکنیک در یادگیری ماشین است که برای ارزیابی عملکرد یک مدل بر روی داده‌های جدید استفاده می‌شود. این روش به ما کمک می‌کند تا از روی برازش بیش از حد (Overfitting)مدل بر روی داده‌های آموزشی جلوگیری کنیم.

در این روش، مجموعه داده اصلی به چندین زیرمجموعه یا بخش (Fold) تقسیم می‌شود. سپس مدل چندین بار آموزش و ارزیابی می‌شود. در هر مرحله، یکی از بخش‌ها به عنوان مجموعه آزمون در نظر گرفته می‌شود و بقیه بخش‌ها برای آموزش مدل استفاده می‌شوند. این فرآیند برای تمام بخش‌ها تکرار می‌شود، به این معنی که هر بخش یک بار به عنوان مجموعه آزمون استفاده می‌شود.

با استفاده از این روش، می‌توانیم عملکرد مدل را بر روی داده‌های جدید که در آموزش استفاده نشده‌اند، ارزیابی کنیم. این امر به ما کمک می‌کند تا برآوردی از عملکرد واقعی مدل بر روی داده‌های ناشناخته داشته باشیم.

اعتبارسنجی متقاطع انواع مختلفی دارد، مانند:

۱. :K-Fold Cross-Validation: در این روش، داده‌ها به K بخش تقسیم می‌شوند و مدل K بار آموزش و ارزیابی می‌شود.
۲. Leave-One-Out Cross-Validation (LOOCV) :در این روش، در هر مرحله، یک نمونه به عنوان مجموعه آزمون در نظر گرفته می‌شود و بقیه نمونه‌ها برای آموزش مدل استفاده می‌شوند. این فرآیند برای تمام نمونه‌ها تکرار می‌شود.
۳. Stratified K-Fold Cross-Validation :این روش مشابه K-Fold Cross-Validation است، با این تفاوت که در هر بخش، نسبت کلاس‌ها (برای مسائل طبقه‌بندی) یا توزیع داده‌ها (برای مسائل رگرسیون) مشابه مجموعه داده اصلی است.

اعتبارسنجی متقاطع به ما کمک می‌کند تا عملکرد مدل را به طور واقع‌بینانه‌تری ارزیابی کنیم و از برازش بیش از حد یا کم‌برازش (Underfitting) مدل جلوگیری کنیم. این روش همچنین به انتخاب بهترین هایپرپارامترها و مدل‌های بهینه کمک می‌کند.

برای زیاد نشدن لود کار باز قبل از شروع کار کاهش بعد میدهیم.