بـــه نـــام خـــدا

دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم های خبره

پروژه اول (درخت تصمیم)

دکتر آرش عبدی

پاییز ۱۴۰۲

طراحان:

بهاره کاوسی نژاد

فرناز خوش دوست آزاد

- **در صورت وجود هرگونه ابهام در سوالات تنها به طراح آن سوال پیام دهید.**
- **با توجه به تنظیم شدن ددلاین تمارین توسط خود شما امکان تمدید وجود ندارد.**
- **خوانا و مرتب بنویسید.**
- **زبان برنامه نویسی دلخواه است.(پیشنهاد: پایتون)**
- **کل محتوای ارسالی را داخل فایل زیپ قرار داده و نام آن را شماره دانشجویی قراردهید.**
- **داک پروژه را واضح و مرتب بنویسید.**
- **انجام تمرین تک نفره است. لطفا به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر**
- **گرفته خواهد شد.**

آیدی تلگرام طراحان :

@iAmMafhoot

@HelenAzaad

@Bahareh_0281

پروژه اول هوش مصنوعی و سیستم های خبره: درخت تصمیم

هدف پروژه: پیاده سازی درخت تصمیم با استفاده از آنتروپی و Gini index برای تشخیص کلاه‌برداری‌های پرداخت‌های مجازی

شرح پروژه:

1. تجزیه و تحلیل مجموعه داده:

با تجزیه و تحلیل مجموعه داده ارائه شده که مربوط به تشخیص کلاه‌برداری‌های پرداخت آنلاین هستند، شروع کنید. featureهای موجود را گسسته سازی کنید، مقادیری که موجود نیستند را با مقداری مناسب جایگزین کنید (و یا می توانید آن سطر را حذف کنید)، مواردی که به صورت عدد نیستند را به عدد تبدیل کنید و ... .

برای گسسته‌سازی ورودی‌های از نوع پیوسته یا ورودی های دارای مقادیر خیلی زیاد بازه‌های عددی در نظر بگیرید. یک ایده آن است که بازه مینیمم تا ماکزیمم اعداد در مجموعه آموزشی را به تعدادی بازه مساوی تقسیم کنید و دو بازه اضافی هم برای مقادیر کمتر از مینیمم و بیشتر از ماکزیمم در نظر بگیرید. همچنین میتوانید ایده‌های دیگری را نیز برای گسسته‌سازی ورودی‌های پیوسته ارائه دهید و آنها را امتحان کنید.

2. پیاده‌سازی الگوریتم درخت تصمیم:

کلاس‌ها و الگوریتم درخت تصمیم مورد نظر خود را پیاده کنید. اگر در پیاده سازی این الگوریتم خلاقیت و نوآوری داشته باشید، نمره امتیازی خواهد داشت؛ مثلا الگوریتم شما به صورت تطبیقی (adaptive) باشد (با تشخیص الگوهای کلاه‌برداری جدید، خود را به روز رسانی کند) و ...

یک بار الگوریتم خود را با آنتروپی و بار دیگر با استفاده از Gini index پیاده‌سازی کنید.

3. آموزش درخت تصمیم (training):

درخت تصمیم خود را با استفاده از دیتاستی که در فایل onlinefraud.csv قرار گرفته train کنید. می‌توانید از 2000 داده اول برای این کار استفاده کنید. در این دیتاست:

- step: نمایانگر واحد زمانی است و هر step به معنای یک ساعت است.
- type: نوع تراکنش آنلاین
- amount: مقدار تراکنش
- nameOrig: کاربر شروع کننده تراکنش
- oldbalanceOrg: موجودی حساب قبل از تراکنش

- newbalanceOrig: موجودی حساب بعد از تراکنش
- nameDest: دریافت کننده در تراکنش
- oldbalanceDest: موجودی دریافت کننده قبل از تراکنش
- newbalanceDest: موجودی دریافت کننده بعد از تراکنش
- isFraud: آیا کلاه برداری است یا خیر

4. ارزیابی مدل:

مدل درخت تصمیم خود را با استفاده از تکنیک‌های cross-validation ارزیابی کنید و نتایج را نشان دهید. از میان داده‌هایی که برای test استفاده نکرده‌اید به صورت تصادفی برای test انتخاب کنید. دقت کنید که بهتر است به تعداد مساوی از نمونه‌های کلاه‌برداری و غیر کلاه‌برداری انتخاب کنید.

5. نمایش درخت تصمیم:

درخت تصمیم خود را نشان دهید. استفاده از کتابخانه‌های موجود در پایتون برای نمایش بهتر درخت تصمیم و همچنین روند پیشرفت و بهبود درخت در طول زمان، نمره امتیازی خواهد داشت.

6. تحلیل و بررسی نتایج:

درخت تصمیم بر اساس آنتروپی و Gini index را با یکدیگر مقایسه کنید. سعی کنید با استفاده از تکنیک‌های مختلف درخت تصمیم خود را بهبود بخشید و نتایج به دست آمده را تحلیل کنید. میزان دقت درخت خود را نمایش دهید.

اگر از ایده‌ی جدیدی در هر قسمت پروژه استفاده کردید (مثلا استفاده از روش جدیدی برای گسسته‌سازی داده‌ها)، نشان دهید که این کار چه تغییری در درخت تصمیم شما ایجاد کرده است و آیا باعث بهبود آن شده است یا خیر.

- ✓ خلاقیت شما برای افزایش دقت درخت مثل افزایش داده‌های آموزشی یا هر گونه انتخاب هوشمندانه از میان آن‌ها، روش‌های جدیدتر و حرفه‌ای تر گسسته سازی و یا حتی فعالیت‌های اضافه‌تر حرفه‌ای مانند تحلیل‌های آماری جداگانه از فیچر ها، Data cleaning یا Feature engineering و ... می‌تواند نمره امتیازی داشته باشد.
- ✓ در نظر داشته باشید برای پیاده سازی درخت تصمیم نباید از توابع آماده استفاده کنید. لذا فرمول آنتروپی، Gini index، تابع خود درخت تصمیم (همانند توابع بازگشتی و فرآیند درخت سازی و ...) را باید خودتان پیاده کنید.
- ✓ استفاده از توابع آماده تنها برای بخش‌های دیگر مانند خواندن اکسل، احیانا نمایش گرافیکی خروجی درخت (در صورت علاقه)، نمایش دقت خروجی و ... بلامانع است.

آنچه تحویل داده می‌شود:

1 . کداجرایی برنامه با توضیحات لازم برای اجرا

2 . درختی که پیدا کرده‌اید را به هر نحوی که می‌توانید و قابل فهم باشد باید نشان دهید (با هر پروتکلی که توضیح می‌دهید باید قابل فهم و توضیحات هر شاخه مشخص باشد).

3 . نشان دهید که در هر گره، کدام ویژگی تست می‌شود، مقدار دست‌آورد اطلاعات و آنتروپی در زیرشاخه‌ها چقدر است.

4 . گزارش مختصری از مسیر انجام کار و چالش‌هایی که مواجه شدید، اجراهای گرفته شده و روند پیشرفت پروژه، توضیحاتی در مورد تفاوت دو معیار آنتروپی و Gini index به صورت مختصر و همچنین توضیحاتی در مورد معیار و دقت خود در داده های آزمایشی ارائه دهید! آیا بیش برازش داشته اید؟ ایده‌ای برای افزایش دقت دارید؟ (حتی اگر پیاده نکرده باشید)

5 . هرگونه تحلیل اضافه مفید و خلاقیت☺ (می‌تواند نمره امتیازی داشته باشد)