بـــه نـــام خـــدا

دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم های خبره

تمرین اول (درخت تصمیم)

دکتر آرش عبدی

پاییز ۱۴۰۲

طراحان : متین محمود خانی – محمدعلی آژینی

- در صورت وجود هر گونه ابهام تنها به طراح پیام دهید.
- با توجه به تنظیم شدن ددلاین تمارین توسط خود شما امکان تمدید وجود ندارد.
- داک پروژه را واضح و مرتب بنویسید.
- مواردی که بعد از تاریخ فوق ارسال شوند قابل قبول نبوده و نمره ای نخواهد داشت.
- انجام تمرین تک نفره است. لطفا به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر گرفته خواهد شد.
- زبان برنامه نویسی دلخواه است.(پیشنهاد: پایتون)
- موارد ارسال شده در تاریخی که بعدا مشخص میشود به صورت آنلاین نیز تحویل گرفته خواهند شد (صرفا آنچه در کوئرا تحویل داده شده است بعدا به صورت آنلاین تست شده و توضیح داده میشود.)
- کل محتوای ارسالی را داخل فایل ریپ قرار داده و نام آن را شماره دانشجویی قرار دهید.

آیدی تلگرام طراحان :

@MATINN2001

@iAmMafhoot

پیاده سازی درخت تصمیم :

در این تمرین به پیاده سازی درخت تصمیم با استفاده از آنتروپی و Gini index میپردازیم.

در مرحله اول، درخت تصمیم را برای داده های گسسته ارائه شده در اسلایدهای کلاس پیاده سازی کنید. برای تست پیاده سازی صورت گرفته، داده های 12 گانه مثال رستوران را مورد آزمایش قرار دهید. سپس درختی که ایجاد کرده اید را با درختی که درون اسلاید ها می باشد مقایسه کنید و میزان دقت درخت پیاده سازی شده را بنویسید.

در مرحله دوم، ما میخواهیم ما میخواهیم بررسی کنیم که با توجه به داده هایی که از حدود 100 هزار پرواز داریم، مدل درخت تصمیمی طراحی کنیم که بتواند پیش‌بینی کند که آیا پرواز برای مسافر رضایت بخش بوده یا نه. برای مثال جنسیت مسافر، کلاس پرواز، مقدار تاخیر و ... در میان این داده ها موجود می باشند. این داده ها در قالب یک فایل airplane.csv تحویل شده داده شده است. در این پایگاه داده نمونه هایی با 23 ویژگی و یک خروجی راضی یا ناراضی وجود دارد. (برای راحتی کار مواردی که به صورت عدد نیستند را نیز میتوانید تبدیل به عدد کنید.) هدف آن است که با کمک این ورودی ها، رضایت یا عدم رضایت مسافر از پرواز تشخیص داده شود.

تعداد ورودی های پروژه زیاد است و بنابراین قبل از پیاده سازی درخت تصمیم شما باید 2 هزار نمونه اول را به عنوان داده های تست جدا کرده و سپس از میان باقی داده ها به صورت رندوم (حداقل 5 هزار نمونه) انتخاب کرده (دقت کنید که از مسافران راضی و ناراضی به مقدار مساوی نمونه تهیه کنید) سپس این نمونه ها را به عنوان داده های آموزشی استفاده کرده و به ادامه روند کار بپردازید. (بدیهی‌ست استفاده از مقدار بیشتر یا کل داده ها موردی ندارد و حتی توصیه نیز می‌شود)

برای گسسته سازی ورودی های از نوع پیوسته یا ورودی های دارای مقادیر خیلی زیاد بازه های 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 ایده های دیگری را نیز برای گسسته سازی ورودی های پیوسته ارائه دهید و آنها را امتحان کنید.

همچنین میتوانید ایده های جدید خود را با ایده اولیه مطرح شده در سوال مقایسه کنید و نتایج را ارائه دهید.

در صورت ارائه روش های پیوسته سازی جدید نمره امتیازی دریافت خواهید کرد اما درصورت اشتباه بودن روش، نمره را از دست خواهید داد. بنابراین، پیشنهاد میشود که ابتدا طبق روش ذکر شده در بالا عمل کرده تا نمره ای از دست ندهید و در ادامه در صورت علاقه به سراغ روش های جدید بروید.

در انتها نیز شما باید دقت درخت خود را با استفاده از نمونه های تست ارزیابی کرده و دقت خروجی داده های تستی درخت خود را گزارش کنید.

خلاقیت شما برای افزایش دقت درخت مثل افزایش داده های آموزشی یا هر گونه انتخاب هوشمندانه از میان آنها، روش های جدید تر و حرفه ای تر گسسته سازی و یا حتی فعالیت های اضافه تر حرفه ای مانند تحلیل های آماری جدا گانه 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 بخش های 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 برای مرحله اول و دوم پیدا کرده اید 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 زیرشاخه ها چقدر است.
4. گزارشی مختصری از مسیر انجام کار و چالشهایی که مواجه شدید ، اجراهای گرفته شده و روند پیشرفت پروژه و همچنین توضیحاتی در مورد تفاوت هر 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 ؟ ایده ای برای افزایش دقت دارید ؟ (حتی اگر پیاده نکرده باشید)
5. هرگونه تحلیل اضافه مفید و خلاقیت 😊 (می‌تواند نمره امتیازی داشته باشد)