به نام خداوند درخت‌های سبز

دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم‌های خبره

ورکشاپ اول (درخت تصمیم)

دکتر آرش عبدی

زمستان 1403

طراح ورکشاپ: سید امیرحسین حسینی جبلی

- در صورت وجود هرگونه ابهام به طراح پیام دهید.
- باتوجه به وجود تاخیر 10 روزه، امکان تاخیر تحت هیچ شرایطی امکان پذیر نیست.
- داکیومنت خود را مرحله به مرحله و خودتان بنویسید(از chat gpt نگیرید)
- انجام ورکشاپ ها تک نفره می‌باشد.
- زبان برنامه نویسی پایتون است.
- موارد ارسال شده به صورت آنلاین تحویل گرفته خواهند شد.
- کل فایل محتوای ارسالی را داخل فایل زیپ قرار داده و نام آن را شماره دانشجویی خود قرار دهید.
- تاریخ ریلیز پروژه: 15 اسفند ماه
- تاریخ تحویل پروژه: 4 فروردین ماه
- آیدی طراح در تلگرام: @amir_jebbeli

1. لینک ورکشاپ آموزشی درخت تصمیم:

https://github.com/IUST-CE293/CE293-AI/blob/main/Problems/DT_ws1.ipynb

دقت کنید، 5 سوال در ورکشاپ بالا طراحی شده است که باید پاسخ دهید

2. پیاده سازی درخت تصمیم:

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

در مرحله اول، باید یک دیتاست انتخاب کنید و آن را دانلود کرده و در تمام مراحل این پروژه از همان دیتاست استفاده کنید، نکته مهم این است که دیتاست انتخاب شده باید حداقل 10 هزار دیتاسمپل و 20 فیچر داشته باشد.(اگر در این مرحله سوالی داشتید از طراح بپرسید).

از آنجاییکه تعداد ورودی ها زیاد است، بنابراین قبل از پیاده سازی درخت تصمیم شما باید 2 هزار دیتاسمپل را به عنوان داده های تست جدا کرده و سپس از میان باقی داده ها به صورت رندوم) حداقل 6 هزار دیتاسمپل (انتخاب کرده) (سپس این دیتاسمپل ها را به عنوان دیتا های train استفاده کرده و به ادامه روند کار بپردازید.) بدیهی ست استفاده از مقدار بیشتر یا کل دیتا ها موردی ندارد و حتی توصیه نیز می شود .

حتی میتوانید از روش held out که در ورکشاپ توضیح داده شده است، استفاده کنید و هایپرپارامترها را tune کنید و یا در هرس درخت از دیتاست validation استفاده کنید.

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 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 و ... می تواند 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 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. کد اجرایی برنامه با توضیحات لازم برای اجرا
2. دیتاست دانلود شده
3. پاسخ های ورکشاپ در یک فایل PDF یا پاسخ در همان فایل نوت بوک
4. درختی که به دست آورده 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)
5. نشان دهید که در هر گره، کدام ویژگی تست میشود، مقدار information gain و Gini index در زیرشاخه ها چقدر است.

6. گزارشی مختصری از مسیر انجام کار و چالشهایی که با آن مواجه شدید ، اجراهای گرفته شده و روند پیشرفت پروژه و همچنین توضیحاتی در مورد معیار و دقت خود در داده های تست ارائه دهید! آیا overfit داشته اید ؟

   ایده ای برای افزایش دقت دارید ؟ (حتی اگر پیاده نکرده باشید)

7. هرگونه تحلیل اضافه مفید و خلاقیت 😊 (می تواند نمره امتیازی داشته باشد)

نکته بسیار مهم: 7 مورد بالا را zip کرده و نام آن را شماره دانشجویی خود گذاشته و فقط در کوئرا ارسال کنید.