پروژه دوم مبانی هوش محاسباتی

مقدمه

دیتاستی مرتبط با ماشین ها در اختبار شما قرار گرفته است. در این پروژه از شما می‌خواهیم تا درخت‌تصمیمی بسازید، که بتواند داده های این دیتاست را بر اساس مدل ماشین طبقه‌بندی کند.

۱ فاز اول

در ابتدا برای لود کردن داده های این دیتاست، می‌توانید از کتابخانه Pandas استفاده کنید. سپس دیتاست را به دو بخش train و test تقسیم کنید. همانطور که در درس خوانده اید، از الگوریتم ID3 برای ساخت درخت تصمیم استفاده می‌شود؛ با استفاده از این الگوریتم درخت تصمیم مناسبی بسازید که با محاسبه Entropy و Information Gain در هر مرحله Attribute مناسب را انتخاب کند. برای این منظور نیاز به پیاده‌سازی کلاس Node دارید؛ همچنین توابع Entropy و Information Gain را نیز خودتان باید پیاده سازی کنید. در این فاز موارد زیر را بررسی و گزارش کنید.

- نحوه تشکیل درخت و انتخاب فیچرها را به طور کامل بررسی کنید.
- راه حل خود را برای فیچرهای پیوسته توضیح دهید.
- بررسی کنید که چه زمانی Entropy و Information Gain به مقدار حداکثر خود می‌رسند.

۲ فاز دوم

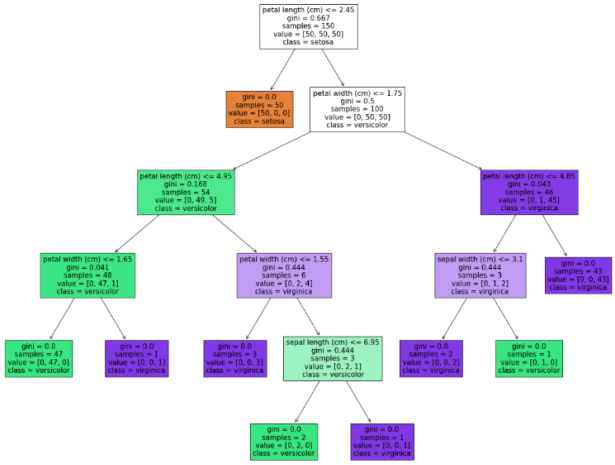
دیتاستی که در اختیار شما قرار گرفته است شامل تعدادی Missing data می‌باشد. با توجه به روش های مطرح شده در کلاس درس، این مورد را باید هندل کنید. نتایج حاصل از هر کدام از روش ها را بررسی و تحلیل کنید؛ همچنین نتایج را با هم مقایسه کنید. علاوه بر این وجود Noise و Outlier در دیتاست را هم بررسی کنید.

۳ فاز سوم

یکی از مشکلات درخت تصمیم، Overfit شدن آن است؛ یعنی درخت داده های آموزشی را حفظ کرده باشد و عملکرد مناسبی برای داده های تست نداشته باشد. برای حل این مشکل راه حلی ارائه کنید و آن را پیاده سازی نمائید. به عنوان یک راه Random Forest) و Bagging را پیاده سازی کنید، علاوه بر این حداقل یک روش دیگر را بررسی کنید.

۴ فاز چهارم

در نهایت درخت ساخته شده را visualize کنید (مشابه شکل ۱). در هر Node درخت، نام Attribute، ها Entropy و Information Gain مربوطه را ذکر کنید. برای این منظور می‌توانید از کتابخانه scikit-learn استفاده کنید.



شکل ۱: درخت تصمیم

توضیحات تکمیلی

- انجام پروژه می‌تواند در قالب گروه های دو نفره و یا به صورت انفرادی صورت گیرد.
- علاوه بر سورس کد پروژه، فایل مستندات نیز باید آپلود شود.
- در فایل مستندات پروژه نام هر دو عضو گروه را ذکر کنید و آپلود فایل‌ها همین که توسط یکی از اعضای گروه انجام شود کافی است.
- هر گونه شباهت نامتعارف بین کد شما و کد سایر گروه‌ها و یا کدهای موجود بر روی اینترنت تقلب محسوب می‌شود و نمره‌ای برای این پروژه دریافت نخواهید کرد.
- سورس کد پروژه را کامنت گذاری کنید و از هرگونه استفاده از اسامی بی معنی برای توابع و متغیر ها پرهیز نمائید.
- در صورت نوشتن داکیومنت تمیز (برای مثال با LaTeX) نمره اضافه برای شما در نظر گرفته خواهد شد.
- استفاده از هرگونه روش خلاقانه نمره اضافی خواهد داشت.
- استفاده از کتابخانه ها و فریم ورک‌های آماده به جز مواردی که درصورت پروژه از شما خواسته شده تا پیاده سازی کنید، بلامانع است.
- فایل شامل سورس کد پروژه و مستندات را در قالب فایل zip و با نام شماره دانشجویی خود ذخیره و ارسال نمائید.
- در صورت داشتن هرگونه سوال میتوانید با kourosh_hsz و یا mhmdrzrs در ارتباط باشید.

با احترام - تیم حل تمرین