دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

تمرین عملی اول درس یادگیری ماشین
پیاده‌سازی درخت تصمیم

استاد درس

دکتر سیدین

نیم سال اول ۱۴۰۲

به موارد زیر توجه کنید:

- به همراه صورت پروژه، دو فایل نوتبوک پایتونی(برای هر یک از دو فاز) قرار داده شده است. نوتبوک فاز اول با نام Phase1.ipynb و نوتبوک فاز دوم با نام Phase2.ipynb قرار داده شده است برای تکمیل این پروژه، باید از همین نوتبوک استفاده کنید و پیاده‌سازی خود برای هر قسمت را به آن اضافه کنید. بنابراین پس از پایان تکمیل نوتبوک، حتما یکبار تمام سلولهای نوتبوک را به ترتیب اجرا کرده تا از صحت عملکرد اجرای کد و تعریف متغیرها اطمینان حاصل کنید.
- مجموعه داده‌های آموزش و تست برای فاز اول به ترتیب PlayTennis_train.csv و PlayTennis_train.csv هستند.
- مجموعه داده برای فاز دوم Decision_Tree_Dataset.csv است.
- مجموعه داده قسمت هرس کردن فاز دوم(امتیازی) mushrooms.zip است.
- در فایل گزارشکار باید یک توضیح کلی از پیاده سازی که هر قسمت انجام داده‌اید ارائه دهید و همچنین به سوالات پاسخ دهید.
- این پروژه به صورت انفرادی است. همفکری و مشورت در رابطه با سوالات پروژه مشکلی ندارد با این حال در صورت تشخیص کپی بودن پاسخهای افراد، نمره صفر برای آنها در نظر گرفته خواهد شد.
- آخرین مهلت تحویل پروژه تا ساعت ۲۳:۵۹ روز شنبه ۱۸ آذر میباشد. برای هر تمرین عملی تا ۵ روز تاخیر بدون کسر نمره مجاز می‌باشد و پس از این زمان، تاخیر با کسر نمره خواهد بود.
- در این پروژه استفاده از کتابخانه‌های Pandas، Numpy و Scikit-Learn محدودیتی ندارد و میتوانید بدون مشکل از آنها استفاده کنید.
- در صورت وجود ابهام در رابطه با پروژه میتوانید سوالات خود را از گروه متصل به کانال مطرح کنید.
- برای آپلود دو فایل نوتبوک که تکمیل کردید به همراه گزارشکار در قالب pdf به صورت یک فایل زیپ شده با فرمت Proj1_[studentnumber].zip آپلود کنید.

فاز اول: پیاده‌سازی الگوریتم درخت تصمیم ID3

هدف این فاز، به دست آوردن درک عملی از الگوریتم درخت تصمیم ID3 با پیاده سازی آن از scratch است. با پیاده‌سازی این قسمت، با مفاهیم بنیادی درختان تصمیم، information gain و ساختار درخت بازگشتی بهتر آشنا می‌شوید.

- **کتابخانه‌های مورد نیاز:**

ابتدا لازم است دو کتابخانه Numpy و Pandas در این بخش را اضافه کنید.

- **مجموعه داده:**

مجموعه داده در نظر گرفته شده برای این تمرین PlayTennis می‌باشد که ابتدا می‌بایست آن را بارگزاری کنید و در ادامه اطلاعات زیر مربوط به داده را با کدهای مناسب گزارش دهید:
تعداد سطرها و ستون‌ها / ستون‌های ویژگی / برچسب / تعداد کلاس

* فایل‌های csv مجموعه داده قرار داده شده است.

- **پیاده‌سازی الگوریتم:**

برای پیاده سازی الگوریتم باید توابع خواسته شده را تکمیل نمایید. توجه کنید که فقط قسمت‌های که با #complete مشخص شده را در همان بخش پیاده‌سازی کنید.

این توابع شامل:

- محاسبه آنتروپی

یک تابع برای محاسبه آنتروپی، یک متریک که ناخالصی یا بی نظمی در مجموعه داده را نشان می‌دهد، ایجاد کنید.

سوال: چگونه آنتروپی برای یک مجموعه داده مشخص محاسبه می شود؟

- محاسبه information gain

یک تابع برای محاسبه بهره اطلاعاتی برای یک ویژگی معین و مقادیر آن پیاده سازی کنید.

سوال: چگونه information gain برای یک ویژگی خاص محاسبه می شود؟

- پیدا کردن ویژگی با بیشترین بهره اطلاعاتی

مانند ویژگی Outlook، ما باید بهره اطلاعاتی را برای هر ویژگی در مجموعه داده محاسبه کنیم. سپس باید ویژگی با بالاترین بهره اطلاعاتی را انتخاب کنیم.

سپس دو تابع مربوط به الگوریتم اصلی ID3 را تکمیل می‌کنیم:

تابع اصلی نشان دهنده الگوریتم اصلی ID3 باشد. این تابع به صورت بازگشتی درخت تصمیم را با انتخاب ویژگی با بالاترین بهره اطلاعاتی در هر مرحله می‌سازد. موارد پایه باید تعریف شوند تا در صورت تحقق شرایط خاص، بازگشت مجدد متوقف شود.

- **نمایش درخت:**

تابع ID3 نمای کلی درخت را نمایش می‌دهد.

سوال: درخت تصمیم ایجاد شده را رسم کنید و اینکه هر شاخه توسط کدام ویژگی تعیین شده است.

- **پیشبینی:**

تابع پیش بینی در الگوریتم درخت تصمیم، برچسب کلاس یک نمونه ورودی را با پیمایش بازگشتی درخت ساخته شده پیش بینی می کند. بررسی می کند که آیا گره فعلی یک برگ است یا خیر و در این صورت مقدار آن را برمی‌گرداند. برای گره‌های داخلی، ویژگی مورد استفاده برای تقسیم را استخراج می‌کند، درخت را بر اساس مقدار ویژگی نمونه حرکت می‌کند، و فرآیند را تا رسیدن به یک گره برگ تکرار می‌کند و برچسب کلاس پیش‌بینی‌شده را ارائه می‌کند. اگر مقدار ویژگی در زیردرخت یافت نشد، تابع با برگرداندن None به خوبی این کار را انجام می دهد.

- **ارزیابی:**

برای ارزیابی مدل یعنی درخت به یک مجموعه داده آزمایشی برچسب دار نیاز داریم. سپس پس از پیش‌بینی می‌توان اختلاف مقدار واقعی و پیش‌بینی‌شده را بر حسب درصد محاسبه کرد.

* ارزیابی را روی داده‌های تست انجام دهید.

فاز دوم: طبقه‌بندی با استفاده از درخت تصمیم scikit-learn

هدف این فاز، به کارگیری دانش به دست آمده از پیاده سازی الگوریتم ID3 برای انجام طبقه بندی با استفاده از پیاده سازی درخت تصمیم scikit-learn است.

- **آشنایی با scikit-learn:**

درک کلی از scikit-learn، یک کتابخانه یادگیری ماشینی پرکاربرد در پایتون به دست آورید. اسناد و ویژگی‌های مربوط به طبقه بندی درخت تصمیم را بررسی کنید.

- **اضافه کردن کتابخانه‌های ضروری:**

کتابخانه‌های مورد نیاز را در اسکریپت یا دفترچه یادداشت خود قرار دهید. برای استفاده از درخت تصمیم، DecisionTreeClassifier را وارد کنید.

- **آماده‌سازی داده‌ها:**

مجموعه داده در نظر گرفته شده برای این قسمت را بارگذاری کنید سپس اطلاعات مربوط به مجموعه داده را که خواسته شده با کد مناسب تکمیل نمایید.

این مجموعه داده مربوط به وام دادن/ندادن برای خرید خانه می‌باشد.

سوال: چرا لازم است مجموعه داده به مجموعه های آموزش و تست تقسیم شود؟

* فایل csv مجموعه داده قرار داده شده است.

- **تجسم داده‌ها(امتیازی):**

تجسم داده‌ها جزء مهمی از تجزیه و تحلیل داده‌ها است که هدف آن ارائه اطلاعات پیچیده به شکل بصری است که در دسترس و به راحتی قابل درک باشد. در این بخش به کمک کتابخانه‌های مربوط به تجسم داده‌ها مانند seaborn، matplotlib و ... میتوانید ویژگی‌های داده را تصویرسازی کنید.

- **ساختن مدل:**
- **تقسیم داده‌ها:**

  مجموعه داده‌های خود را به مجموعه‌های آموزش و تست تقسیم کنید. مجموعه داده آموزش برای آموزش مدل استفاده می شود، در حالی که مجموعه تست عملکرد مدل را ارزیابی می کند.
- **نمونه سازی طبقه بندی درخت تصمیم:**

  یک نمونه از کلاس DecisionTreeClassifier ایجاد کنید. پارامترهای آن، مانند حداکثر عمق یا حداقل نمونه‌ها در هر برگ را مطابق با مجموعه داده و اهداف خود تعیین کنید.
- **آموزش مدل:**

  طبقه بندی درخت تصمیم را با استفاده از داده‌های آموزش آموزش دهید. این شامل راه‌اندازی مدل با ویژگی‌ها و برچسب‌های هدف مربوط به آنها است.

سوال: چرا fit کردن مدل با داده های آموزشی ضروری است؟

- **پیشبینی:**

از مدل آموزش دیده برای پیش بینی در مجموعه تست استفاده کنید. این مرحله میزان تعمیم مدل را به داده‌های دیده نشده ارزیابی می کند.

سوال: خروجی پیش‌بینی‌شده در طبقه‌بندی چه چیزی را نشان می‌دهد؟

- **ارزیابی مدل:**

عملکرد مدل را با استفاده از معیارهای مختلف مانند دقت، صحت، بازیابی(recall) و F1-score ارزیابی کنید. این مرحله بینش‌هایی را در مورد اینکه مدل در تکلیف طبقه بندی داده شده چقدر خوب عمل می کند، ارائه می دهد.

سوال: معیارهای استفاده شده برای ارزیابی عملکرد طبقه‌بندی چگونه تفسیر می شوند؟

- **بهبود مدل:**

مدل را با تنظیم هایپرپارامترها بهبود بخشید و به ازای چند حالت مختلف نتایج را بررسی کنید. هدف این فرآیند بهینه سازی عملکرد طبقه‌بندی کننده است.

- **هرس کردن(امتیازی):**

هرس تکنیکی است که برای جلوگیری از پیچیده شدن بیش از حد درختان تصمیم و تطبیق بیش از حد داده های آموزشی مورد استفاده قرار می گیرد. در حالی که ID3 ذاتاً برای هرس طراحی نشده است، می توانید یک الگوریتم هرس را برای درخت تصمیم ID3 پیاده سازی کنید. یکی از رویکردهای رایج این است که یک درخت کامل را پرورش دهید و سپس آن را هرس کنید.

در این بخش با استفاده از کدی که در فاز اول پیاده سازی کردید برای مجموعه داده mushrooms یک تابع تعریف کنید که درخت ایجاد شده را هرس نماید. به این نکته باید توجه کرد که هر نود به شرطی هرس بشه که خطا روی داده ولید کمتر شود و حداقل ۱ درصد بهبود یابد.