بسمه تعالی

گزارش تمرین درخت تصمیم

درس یادگیری ماشین دکتر سیدین

تهیه‌کنندگان:

حمیدرضا دشت‌آبادی ۹۸۲۳۰۳۲

سید جلال طباطبایی ۹۸۲۳۰۵۹

خلاصه :

در این پروژه یک کلاس برای تعریف گره ها (Node) که شامل ویژگی و شاخه هاوبچه ها(childs) تعریف شده است .در ادامه یک کلاس دیگر برای تعریف درخت تصمیم که ریشه آن شیئی از جنس کلاس Node می‌باشد و هر درخت تصمیم یک ریشه یا rootدارد .تابع fit در هرمرحله ویژگی root درخت تصمیم فعلی را با محاسبه بهترین ویژگی به لحاظ information gain تعیین می‌کند .بهترین ویژگی با تابع getbestfeat بدست می‌آید که از تابع getIG برای محاسبه informationgain بهره می‌برد که آن‌هم از تابع getentropy یا getgini به عنوان مقیاس اندازه گیری استفاده می-کند.همچنین در کلاس Nodeیک متغیر به اسم branch_childs نیز برای تعریف شاخه و گره مربوط به آن تعریف شده است. لذا برای هر شاخه آن (key)، root درخت سطح بعدی نسبت داده می‌شود که خود از جنس Node است (شیءگرایی).تابع fitدر ادامه یک object موقت از نوع treeبا دیتای مربوط به آن branchو عمق منهای یک وchild آن شاخه می‌سازد.در ادامه به صورت بازگشتی تابع fit روی درخت موقتی ایجاد شده فراخوانده می‌شود تا root آن را بیابد و درون child آن شاخه بریزد.در ادامه هنگامی که درخت ساخته شد می‌توان تابع predictرا با پاس دادن داده بدون لیبل فراخوانی کرد که در شاخه-های درخت جلو می‌رود (به صورت بازگشتی با ورودی‌های بخشی از داده ای که مربوط به آن شاخه است و child آن شاخه فراخوانی می‌شود)تا به انتهای درخت برسد (یعنی leaf).حال ویژگی آن را به y_predict متناظر با سطرهای باقی‌مانده داده تست ،نسبت (assign) می‌دهیم.تابع accuracyنیز داده y-predict با y-testمقایسه می کند.

بخش الف : در این بخش درخت با عمق ۸ آموزش داده شد و روش entropy روی ۸۰ درصد داده ها به صورت تصادفی که حدود ۳۰ ثانیه طول کشید و صحت آن تقریبا برابر ۱۰۰ درصد به دست آمد

بخش ب : در این بخش ۸ حالت ممکن بررسی شد .در حالت ۵۰ درصد عمق ۶ و۸ و روش های آنتروپی و جینی صحت بدست آمده تقریبا یکسان است . در حالت ۷۵ درصد عمق ۸ صحت داده ها با روش جینی ۱۰۰درصد و آنتروپی برخلاف انتظار ۹۵ درصد به دست آمد و درخت overfit شده است.