تکلیف شماره ۱-۲ یادگیری ماشین: یادگیری درخت تصمیم- پیاده‌سازی عملی

در این بخش با دادگان داده شده باید الگوریتم ID3 را برای ساختن درخت تصمیم پیاده‌سازی کنید. برای پیاده‌سازی می‌توانید از نرم‌افزار MATLAB یا پایتون استفاده کنید.

دادگان داده شده با نام Adult از مجموعه UCI استخراج شده است. هدف آن یک طبقه‌بندی باینری برای این پیش‌بینی است که آیا فردی بالای 50K در سال درآمد دارد یا خیر. این دادگان در فرم اولیه شامل ۱۴ ویژگی است که ۸ تای آن گسسته categorical و ۶ تای آن پیوسته می‌باشند، و علاوه بر آن یک ویژگی برچسب کلاس دارد. برای ساده‌سازی و امکان پیاده‌سازی ID3، ویژگیهای پیوسته را از مجموعه دادگان حذف کرده‌ایم. همچنین تعدادی از مثالها، شامل ویژگیهای با مقدار نامعلوم (missing) بودند که آنها نیز حذف شده‌اند. درنتیجه دادگان موجود، دادگان تمیز از ۸ ویژگی categorical می‌باشد که به ۲ مجموعه train و test که در هر کدام 10000 مثال وجود دارد تقسیم شده است. اولین ویژگی هر مثال، برچسب داده را مشخص می‌کند و ۸ ویژگی دیگر شامل مقادیر زیر هستند:

1) workclass (8 values)
2) education (16 values)
3) marital-status (7 values)
4) occupation (14 values)
5) relationship (6 values)
6) race (5 values)
7) sex (2 values)
8) native-country (41 values)

حال مراحل زیر را انجام دهید:

۱- با الگوریتم ID3 و مجموعه دادگان train داده شده یک درخت تصمیم بسازید. همانطور که در درس اشاره شد، برای انتخاب بهترین ویژگی در هر مرحله باید از information gain استفاده کنید. میزان صحت طبقه‌بندی را روی مجموعه‌های train و test محاسبه کنید. (راهنمایی: درصورتی که پس از استفاده از تمام ویژگیها، کلاس مشخص نمی‌شود، همانطور که در درس اشاره شد باید از رای‌گیری بین داده‌ها برای برچسب‌زنی انتهای درخت استفاده کنید)

برای آموزش از چند حالت زیر استفاده کنید:

a) ۲۵ درصد دادگان آموزش را بصورت تصادفی انتخاب کنید و درخت را با آن آموزش دهید. سپس روی کل دادگان تست آن را تست کنید و درنهایت صحت طبقه‌بندی را روی دادگان آموزش و تست به همراه سایز درخت گزارش کنید. این عمل تقسیم تصادفی داده‌ها برای عمل آموزش را ۵ بار تکرار کنید و در هر بار مقادیر صحت را مشخص کنید، و درنهایت هم میزان متوسط صحت را بدست آورید.

b) اندازه‌گیری تاثیر میزان سایز دادگان آموزش: علاوه بر انتخاب تصادفی ۲۵ درصد دادگان آموزش که در مرحله قبل انجام دادید، مقادیر ۳۵ درصد، ۴۵ درصد، ۵۵ درصد، ۶۵ درصد، ۷۵ درصد دادگان آموزش را هم بصورت تصادفی انتخاب کنید و دقیقا همان نتایجی که در مرحله قبل از شما خواسته شده را با ۵ بار اجرای

مختلف برنامه بدست آورید. در آخر هم کل دادگان آموزش را استفاده کنید و نتایج را گزارش کنید. در گزارش، بحث کنید که میزان دادگان آموزش چه تاثیری در صحت طبقه‌بندی روی دادگان تست و سایز درخت تصمیم داشته است.

۲- انجام post-pruning و کاهش تعداد گره‌های درخت: باید عمل هرس کردن درخت به روش reduced error pruning روی داده‌ها به ترتیب زیر انجام شود:

a) ۷۵ درصد دادگان آموزش را بصورت تصادفی بعنوان دادگان آموزش انتخاب کنید و ۲۵ درصد باقیمانده را بعنوان دادگان validation. سپس به کمک دادگان آموزش با ID3 درخت تصمیم را بسازید و با دادگان validation عمل هرس کردن را انجام دهید. منحنی تغییر خطای طبقه‌بندی روی هر ۳ مجموعه train، validationو test را به ازاء تعداد گره‌های مختلف درخت بکشید.

c) این بار کل دادگان آموزش را بعنوان دادگان آموزش انتخاب کنید و ۲۵ درصد از دادگان تست را بصورت تصادفی بعنوان دادگان validation. سپس به کمک دادگان آموزش با ID3 درخت تصمیم را بسازید و با دادگان validation عمل هرس کردن را انجام دهید. منحنی تغییر خطای طبقه‌بندی روی هر ۳ مجموعه train، validation و test را به ازاء تعداد گره‌های مختلف درخت بکشید.

راهنمایی: برای راحتی کار در مرحله هرس کردن، می‌توانید از یک آستانه برای تصمیم‌گیری در مورد هر گره استفاده کنید، مثلا اگر در مورد هر گره حداقل 0.5% بهبود عملکرد طبقه‌بندی درخت تصمیم نتیجه روی مجموعه validation داشتید، می‌توانید آن گره را هرس کنید.

گزارش: برنامه‌های نوشته شده به همراه نتایج و شکل‌ها و آنالیزهای خواسته شده در هر مرحله را ضمیمه کنید. همچنین یک بحث و بررسی در مورد نتایج و اثر overfitting به دادگان آموزشی و تاثیر pruning انجام دهید. گزارش شامل تحلیل و مقایسات و توضیحات کافی در مورد نحوه پیاده‌سازی‌ها، لیست برنامه‌ها و پارامترهای بکار رفته می‌باشد. لطفا از نوشتن گزارش‌های طولانی اجتناب کنید.

توجه: کل تمرین و گزارش باید بصورت انفرادی نوشته شود، و به برنامه‌ها و گزارش‌های کپی شده نمره‌ای تعلق نمی‌گیرد.

توجه: از کتابخانه‌های آماده درخت تصمیم نمی‌توانید استفاده کنید.

نمره مثبت اضافی: تحلیل کامل نتایج، هرس کردن درخت، چندین بار اجرای برنامه با انتخاب داده‌های تصادفی مشخص شده در مرحله هرس کردن و پیدا کردن مقدار متوسط نتایج.

موعد تحویل: ۲۱ اردیبهشت ۱۳۹۹

روش تحویل: آپلود در courses

موفق باشید

سیدین