به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

هوش مصنوعی، ترم پاییز ۹۸-۹۹

پروژه چهارم، مهلت ارسال: جمعه ۱۵ آذر

طراحان پروژه: نازنین صبری، صدف صادقیان، محمدرضا یزدانی‌فر

مقدمه

هدف این پروژه آشنایی بیشتر شما درخت تصمیم، یادگیری گروهی، bootstrap aggregation یا همان bagging و همچنین جنگل‌های تصادفی است. به این منظور تحلیلی را بر روی داده‌های بیماران قلبی انجام می‌دهید.

شرح داده‌ی بیماران قلبی

این داده از اینجا قابل دریافت است. شرح ویژگی‌های داده را در همان سایت یا در ادامه می‌توانید ببینید:

age: سن افراد به سال

sex: جنسیت (یک برای آقایان و صفر برای بانوان)

cp: نوع سینه درد (مقداری حسابی بین صفر تا سه)

trestbps: فشار خون در حالت استراحت (میلی‌لیتر جیوه)

chol: کلسترول سرم (میلی‌گرم بر دسی‌لیتر mg/dl)

fbs: اینکه قند خون ناشتا از ۱۲۰ میلی‌گرم بر دسی‌لیتر بیشتر باشد. این مقدار یک است اگر قند خون ناشتا از ۱۲۰ بیشتر باشد و در غیر این‌صورت صفر است)

restecg: نتایج الکتروکاردیوگرافی در حالت استراحت (مقداری حسابی بین صفر و دو)

thalach: بیشینه نرخ ضربان قلب

eang: این ویژگی که به Exercise Induced Angina موسوم است و به معنای بروز سینه درد هنگام ورزش کردن است. (مقدار صفر به معنای عدم وجود این مساله و یک به معنای وجود این مساله است)

oldpeak: مقداری اعشاری که بیانگر ST Depression در حالت ورزش نسبت به حالت استراحت است.

slope: شیب ST Segment در اوج ورزش

ca: تعداد عروق اصلی رنگی که توسط فلوروسپی رنگ آمیزی شده اند. این مقدار بین صفر تا سه است.

thal: ۳ = عادی؛ ۶ = نقص ثابت؛ ۷ = نقص برگشت‌پذیر

target: صفر یا یک است و به معنای ابتلا یا عدم ابتلای فرد به بیماری قلبی است.

پیاده سازی

در قسمت‌های مختلف این بخش باید مقدار target را پیش‌بینی کنید.

1) به کمک کتابخانه‌ی sklearn الگوریتم درخت تصمیم را پیاده‌سازی کنید و دقت مدل حاصل را بدست آورید.

2) مراحل زیر را جهت ساخت یک جنگل تصادفی انجام دهید: (در این قسمت مجاز یه استفاده از کلاس جنگل تصادفی کتابخانه‌ی sklearn نیستید.)

1) داده‌های موجود را به صورت تصادفی به پنج دسته‌ی ۱۵۰تایی تقسیم کنید. (امکان تکرار داده در دسته‌های مختلف وجود دارد.)

2) با کمک گرفتن از بخش قبل روش bagging را پیاده‌سازی کنید و دقت مدل حاصل را بدست آورید.

3) با شروع از تمام ویژگی های موجود هر بار یک ویژگی را حذف کنید و دقت مدل حاصل را بدست آورید حذف کدام ویژگی افت کمتری را نتیجه می دهد؟

4) به صورت تصادفی از میان ویژگی‌های موجود، پنج ویژگی را انتخاب کرده و درخت تصمیم را درست کنید.

5) .با کمک گرفتن از بخش قبل مدل جنگل تصادفی را پیاده سازی کنید.

گزارش کار

در گزارش خود ابتدا به صورت کوتاه تمامی کاری که در پروژه انجام داده‌اید را بیان کنید. گزارش در عین خلاصه بودن باید جامع باشد. هم‌چنین در گزارش خود به توضیح موارد زیر نیز بپردازید:

- bootstrapping چیست و چه تاثیری برروی واریانس و انحراف معیار استاندارد دارد؟
- overfitting چیست و چرا درخت تصمیم به آن حساس است؟ bagging سعی دارد چه مساله‌ای را حل کند؟
- ارتباط random forest با bagging چیست؟ random forest سعی دارد چه مساله‌ای را حل کند؟
- با مقایسه دقت ثبت شده در قسمت‌های ۱، ۲,۲ و ۵,۲ و با توجه به پاسخ سوالات بالا چه نتیجه‌ای می‌گیرید؟

نکات پایانی

- هدف از تمرین یادگیری شما است لطفا تمرین را خودتان انجام دهید
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند، در غیر این صورت ایمیل بزنید یا حضوری از یکی از طراح‌های پروژه بپرسید.