تمرین سوم درس داده‌کاوی

دسته‌بندی

پاییز ۹۸

۱ تمرین‌های تئوری

سوال ۱

boosting چیست و چگونه باعث افزایش دقت می‌شود؟ یکی از روش‌هایی که از ایده boosting استفاده می‌کند، gradient boosting می‌باشد که برای حل مسئله رگرسیون مجموعه‌ای از درخت‌های تصمیم را نتیجه می‌دهد. در مورد این روش تحقیق کنید و در حد یک پاراگراف توضیح دهید.

سوال ۲

نشان دهید که accuracy تابعی از precision و recall است.

سوال ۳

جدول زیر را در نظر بگیرید که در آن A ، B و C ویژگی‌ها هستند و Y هدف است.

A	B	C	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	0	0
1	1	1	0

۱. information gain را برای ریشه‌ی درخت حساب کنید و مشخص کنید که در این مرحله الگوریتم ID3 کدام ویژگی را انتخاب می‌کند.

۲. درخت تصمیم نهایی را با استفاده از الگوریتم ID3 نمایش دهید. (در این‌جا بیش‌برازش اهمیتی ندارد پس لازم نیست درخت را هرس کنید.)

۳. یکی از راه‌های جلوگیری از بیش‌برازش، پیش‌هرس کردن است، به این معنا که هرگاه مقدار informa-tion gain از یک آستانه کمتر شود دیگر اجازه رشد به درخت نمی‌دهیم. اگر این حد آستانه برابر با ۰.۰۰۰۱ انتخاب شود، درخت را رسم کنید.

۴. رویکرد دیگر پس‌هرس کردن است که در آن، بعد از تشکیل دادن کامل درخت تصمیم، از طرف برگ‌ها به سمت ریشه شروع کرده و با در نظر گرفتن حد آستانه برای information gain ، هرس را انجام می‌دهیم. با همان مقدار آستانه در بخش قبل، درخت تصمیم به دست آمده با استفاده از روش پس‌هرس کردن را نمایش دهید.

۵. دو روش هرس کردن را که در بخش‌های قبل مطرح شد، با هم مقایسه کنید و مزایا و معایب هر کدام را ذکر کنید.

سوال ۴

آیا ID3 تضمین می‌کند که به جوابی برسد که globally optimum باشد؟ توضیح دهید.

سوال ۵

به طور کلی، نرمال کردن داده‌ها به چه منظور صورت می‌گیرد؟

۲ تمرین‌های عملی

سوال ۱

در این مسئله باید با استفاده از روش رگرسیون خطی داده‌های مربوط به بیماری دیابت را دسته‌بندی کنید. برای کسب اطلاعات بیشتر می‌توانید به این لینک مراجعه کنید. توجه شود که در این سوال باید رگرسیون را به طور کامل پیاده سازی کنید و از کتابخانه‌های آماده استفاده نکنید. داده‌های مورد نیاز برای این سوال در پوشه ۱ قرار گرفته. در انتها لازم است میزان خطا و دقت مدل خود را گزارش کنید.

سوال ۲

در این مسئله باید به کمک روش naive bayes پیش‌بینی کنید که یک سوال صادقانه است یا خیر. به سوالی غیرصادقانه گفته می‌شود که قصد آن بیشتر بیان یک حکم باشد و فرد پرسشگر به دنبال پاسخ مناسبی نباشد. سوالات غیرصادقانه بعضا می‌خواهند یک گزاره را درباره‌ی گروهی از مردم (مانند یک قومیت یا جنسیت مشخص) به صورت ضمنی به کاربر القا کنند. به عنوان مثال سوال‌هایی از قبیل:

- آیا آمارها نشان می‌دهند که تصادف‌های رانندگی در بین رانندگان خانم رایج‌تر از رانندگان آقاست؟
- آیا این که ایالت x با بیشترین جمعیت سیاه‌پوست، دارای بالاترین میزان نزاع خیابانی ثبت شده می‌باشد، اتفاقی است؟

غیر صادقانه هستند.

مجموعه داده‌ای که برای این سوال در نظر گرفته شده، دارای پرسش‌هایی است که در سایت quora مطرح شده‌اند. داده‌های آموزش طوری برچسب خورده‌اند که سوالات غیرصادقانه با ۱ مشخص شده است و در غیر این صورت برچسب آن‌ها ۰ است که البته این داده‌ها ممکن است خطا نیز داشته باشند. لازم است بخشی از داده‌های دارای برچسب را برای آموزش و باقی آن را برای validation در نظر بگیرید و معیار f1 مدل خود را برای هر یک از این مجموعه‌ها گزارش کنید. پیش‌پردازش مناسب می‌تواند تاثیر زیادی روی دقت بگذارد. پاسخ شما باید طبق فرمت مورد نیاز برای بارگذاری در سایت kaggle باشد. برای دسترسی به داده‌های این سوال یک لینک در پوشه ۲ نیز قرار داده شده است.

سوال ۳

در این مسئله با استفاده از مجموعه داده تایتانیک به حل یک مسئله داده‌کاوی خواهید پرداخت.
از طریق این لینک می‌توانید به صفحه‌ی مسابقه دسترسی داشته باشید و توضیحات دقیق درباره‌ی مسئله را بیابید.
برای شروع می‌توانید از این لینک‌ها نیز استفاده کنید:

- A Journey through Titanic
- How to score 0.8134 in Titanic Kaggle Challenge

هدف از این مسئله آموزش مدلی برای پیش‌بینی زنده ماندن یا کشته شدن مسافران کشتی تایتانیک از روی داده‌های موجود است. انتظار می‌رود با کمک لینک‌های اشاره شده، کد مورد نیاز برای پیش‌بینی نتایج روی داده‌های تست و ساخت خروجی طبق فرمت مورد نیاز برای بارگذاری در سایت kaggle را پیاده‌سازی نمایید.

- از کامنت‌های مناسب برای بیان قسمت‌های مختلف کد استفاده کنید.
- در فایل بارگذاری شده کد و نتایج مربوط به بهترین عملکرد خود را گزارش کنید.
- در مورد مدل‌ها، ویژگی‌ها و پارامترهای استفاده شده، توضیحات و تحلیل‌های مد نظرتان را گزارش کنید.
- رعایت کردن تمامی مراحل پیش پردازش داده ها، مهندسی فیچرها، انتخاب بهترین فیچر، ساخت مدل های مختلف بر اساس تناسب آن ها با داده ی موردنظر و مقایسه ی آن ها، پس پردازش داده ها و مراحل مورد نیاز دیگر، الزامی است و از آنجا که راه حل های مختلفی از این مساله در خود سایت هم موجود است، نمره ی اصلی این تمرین مربوط به گزارش شما از روند حل مساله تان خواهد بود.
- داده‌های مورد نیاز برای این سوال در پوشه ۳ قرار داده شده.

نکات کلی

- برای حل سوالات عملی ترجیحا از زبان پایتون استفاده کنید.
- مهلت تحویل تمرین ۱۳ دی ۹۸ می‌باشد.
- به ازای هر روز تاخیر ۱۰ درصد از نمره کاسته خواهد شد.
- پاسخ بخش تئوری و گزارش بخش عملی را در قالب pdf به نام "ID_CLS.pdf" به همراه فایل‌های کد در قالب یک فایل zip. به نام "ID_CLS.zip" آپلود کنید. مثلا اگر شماره دانشجویی شما ۹۶۱۳۱۹۰ می‌باشد، فایل را 9613190_CLS نام‌گذاری کنید.
- در صورت وجود سوال یا ابهام با ایمیل‌های yasaman.m.1997@gmail.com و heidarymm@yahoo.com در ارتباط باشید.