تمرین سری ۲ واحد درسی داده کاوی مجازی

جناب آقای دکتر فراهانی
دستیار آموزشی : علی شریفی

۲۹ فروردین ۱۴۰۱

توجه کنید شما میتوانید بر روی ککگل یا کولب و یا کامپیوتر های شخصی خود کار کنید .
به جای دانلود و آپلود دیتاست در گوگل درایو برای استفاده در کولب میتوانید به شیوه زیر عمل کنید .
چگونه از دیتاست های ککگل در کولب استفاده کنیم ؟
ددلاین تمرین تا ۳۰ اردیبهشت ۱۴۰۱ می باشد.
نحوه تحویل پاسخ تمرین ها در ریپازیتوری متعلق به درس می باشد.

۱ تمرین

۱.۱ دیتاست شماره ۱

در این تمرین از دیتاست اطلاعات مشخصات و قیمت گوشی همراه استفاده شده است.
لینک دیتاست
از شما خواسته میشود به تسک های زیر را انجام دهید.

۱.۱.۱ تسک های اصلی

۱. روش انتخاب ویژگی Forward Selection را پیاده سازی کنید . برای معیار انتخاب فیچر جدید در هر مرحله از AUC استفاده کنید . استفاده از پکیج مجاز نمی باشد و باید این بخش را خودتان پیاده سازی کنید. برای سادگی این پیاده سازی موبایل ها را به دو گروه قیمت بالا (دو کلاس گران را ادغام کنید و یک کلاس در نظر بگیرید) و گروه با قیمت پایین تقسیم بندی کنید.
در روش انتخاب پیشرو ما از یک مجموعه تهی شروع کرده و در هرگام سعی داریم فیچر را به مجموعه فیچرهای انتخابی اضافه کنیم که AUC را افرایش دهد .

۲. با استفاده از کد پیاده سازی شده در بخش قبل به انتخاب ویژگی ها از فیچر ها بپردازید و سپس مدل لجستیک (با استفاده از پکیج) را بر روی فیچرهای انتخاب شده اجرا کنید و معیار های precision ، recall ، f1-score را گزارش کنید .

۳. با استفاده از الگوریتم PCA در حالتی که تعداد Component ها با تعداد فیچرهای انتخابی حاصل روش انتخاب ویژگی پیشرو برابر باشد (یعنی اگر در سوال ۱ شما با استفاده از انتخاب ویژگی پیشرو به طور مثال ۵ فیچر را انتخاب کردید در الگوریتم PCA هم به عنوان آرگومان ورودی تعداد Component را ۵ درج کنید .) دیتاست را تغییر دهید .

۴. با استفاده از دیتاست تغییر یافته در سوال قبلی و به کمک پکیج ها یک رگرسیون لجستیک را پیاده سازی کنید و معیار های precision ، recall ، f1-score را گزارش کنید .

۵.

۶. مهندسی ویژگی یکی از بخش های مهم در فرایندهای یادگیری ماشین میباشد . بر روی دیتاست موارد زیر را اجرا کنید .

(آ) بر روی فیچر battery power از روش binning استفاده کنید . (حداقل سه اندازه مختلف برای بین ها در نظر بگیرید و حتی سایز بین ها را نامساوی در نظر بگیرید .)

(ب) بر فیچرهای کتگوریکال در دیتاست one hot encoding را اعمال کنید . چرا ما باید به صورت کلی از این کدگذاری بهره ببریم .

(ج) بررسی کنید آیا استفاده از تبدیل هایی از قبیل log transform و یا تبدیل نمایی در اینجا کاربرد دارد . به صورت کلی چرا از این دست تبدیلات بهره میبریم . (در این بخش شما مجاز هستید اگر تبدیل دیگری را مناسب میدانید اعمال کنید این بخش نمره امتیازی برای شما خواهد داشت . حتما دلیل استفاده از تبدیل استفاده شده را بیان کنید .)

(د) یک فیچر جدید به نام مساحت یا حجم گوشی بسازید .

۷. برای هریک از حالت های سوال قبلی یک مدل رگرسیون لجستیک بسازید و بررسی کنید یکبار هم هر ۵ حالت را باهم اعمال کنید و مدل رگرسیون لجستیک روی آنها اجرا کنید . حاصل این مدل ها را گزارش کنید .

۸. Bootstrapping چیست و چه تفاوتی با Cross Validation دارد؟ در کجا ها از Bootstrapping استفاده میشود .

۹. 5x2 cross validation را در یک پاراگراف توضیح دهید سپس بیان کنید در چه جاهایی استفاده از این روش میتواند مفید باشد .

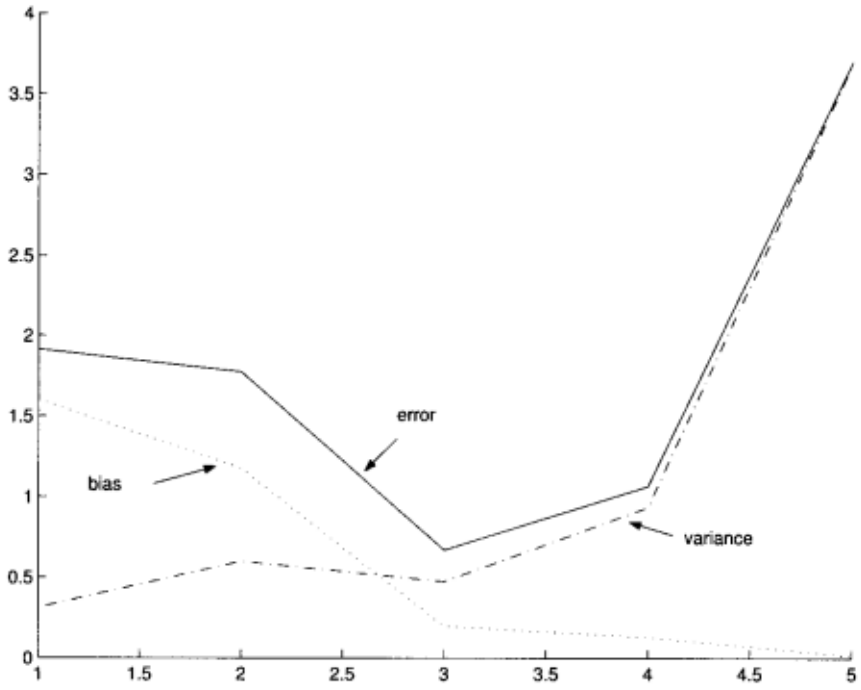
۱۰. در خصوص الگوریتم های مختلف ساخت درخت تصمیم (همانند ID3 ، CART و...) تحقیق کنید . به صورت کلی تفاوت الگوریتم های مختلف ساخت درخت تصمیم در چیست ؟

۱۱. به دلخواه با استفاده از پکیج ها بر روی دیتاست مطرح شده یک درخت تصمیم بسازید .

۱۲. برای درخت تصمیم پارامتر های مختلف مورد ارزیابی قرار دهید . آیا عمق درخت و تعداد نمونه های موجود در هر هر گره تاثیری در عملکرد درخت تصمیم دارد ؟

۱۳. در خصوص هرس کردن Pruning درخت تصمیم تحقیق کنید . چرا ما به بحث هرس کردن درخت تصمیم نیاز دارد و چه کمکی به ما میکند .

۱۴. آیا میتوان با استفاده از روش Elbow با استفاده نموداری مشابه نمودار زیر که نمایان گر بایاس ، واریانس و مرتبه مدل است . بهترین مرتبه مدل برای پیچیدگی مدل را یافت ؟ به طور مثال با استفاده از روش elbow میتوان در نظر گرفت که بر روی دیتاست ، مدلی از مرتبه ۳ جواب خوبی به ما میدهد . آیا همواره در تمامی مسائل و **نه صرفا بحث تحلیلی** میتوان اینگونه قضاوت کرد و مرتبه مناسب را به دست آورد ؟(راهنمایی برای پاسخ به این سوال توجه به مفهوم بایاس میتواند کمک کننده باشد.)
در شکل (۱) خطا از حاصل جمع توان بایاس و واریانس به دست می آید .



شکل ۱: نمودار بایاس و واریانس بنا بر مرتبه های مختلف مدل

۲.۱.۱ تسک های امتیازی

این تسک ها، فرای تسک های اصلی می باشد و پاسخ گویی به آنها دارای نمره امتیازی می باشد.

۱. روش انتخاب ویژگی Backward Selection را پیاده سازی کنید و با استفاده از فیچرهای انتخاب شده و کمک پکیج یک رگرسیون لجستیک را پیاده سازی کنید . معیار های precision ، recall ، f1-score را گزارش کنید و نتایج را با الگوریتم انتخاب ویژگی پیشرو در بخش تسک اصلی مقایسه کنید .

۲. چگونه میتوان با استفاده از statistical significance tests به مقایسه مدل ها پرداخت ؟ (توضیح کامل) [۱]

۳. معیار Matthews Correlation Coefficient(MCC) چیست و در چه جاهایی استفاده میشود .

۴. سعی کنید هرس کردن درخت ها که در تسک های اصلی در سوال ۱۳ مطرح شد را در مدل خود اجرا کنید و بررسی کنید آیا این هرس کردن در نتایج شما تاثیر داشته است .

۲.۱ دیتاست شماره ۲

این دیتاست نیز همان دیتاست تمرین ۱ شما یعنی آگهی خانه ها در کشور آلمان می باشد.
لینک دیتاست

۱.۲.۱ تسک های اصلی

۱. یک مدل رگرسیون تنها با سه تا فیچرها

- serviceCharge
- heatingType
- telekomUploadSpeed

از با تابع خطا MSE بدون استفاده ازپکیج پیاده سازی کنید.

۲. مدل پیاده سازی شده را با مدل با استفاده از پکیج sklearn یا statsmodels مقایسه کنید.

۳. با استفاده از پکیج ها بر روی دیتاست (انتخاب فیچرها بر عهده شما می باشد.) رگرسیون lasso ، ridge را بسازید و نتایج را با حالت های قبلی بررسی کنید.

[۱] راهنمایی : Statistical Significance Tests for Comparing Models را جستجو کنید .

۲.۲.۱ تسک های امتیازی

این تسک ها، فرای تسک های اصلی می باشد و پاسخ گویی به آنها دارای نمره امتیازی می باشد.

۱. مدل رگرسیون با یکی از تابع خطای زیر بدون استفاده از پکیج بر روی سه فیچر گفته شده سوال ۱ پیاده سازی کنید و نتایج بررسی کنید. ارور های معروف برای مسائل رگرسیون به صورت زیر میباشند .
تعداد نمونه ها N و r^t مقدار واقعی نمونه t ام و y^t مقدار پیش بینی شده برای نمونه t ام در نظر میگیریم .

- Absolute Error

$$\frac{1}{2}\sum_{t=1}^{N}|r^t - y^t| \qquad (۱)$$

- epsilon-sensitive Error

$$\sum_{t=1}^{N} 1\left(|r^t - y^t| > \epsilon\right)\left(|r^t - y^t| - \epsilon\right) \qquad (۲)$$

۲ نحوه ارسال

تمامی تمرین ها طبق نحوه بیان شده در کلاس های حل تمرین در داخل گیت هاب تحویل گرفته میشوند.
دانشجویان گرامی نهایتا تا ۳۰ اردیبهشت ماه فرصت دارند تا کارهای خود را موفق در گیت هاب قرار دهند.
با توجه به فرصت سه هفته ای قبلی جهت تمرین در کار کردن با گیت و بررسی ارسال ها در گیت هاب و بیان موارد موفق و غیر موفق، ارسال غیرموفق طبق دستورالعمل گفته به منزله عدم ارسال تمرین در نظر گرفته میشود.