



دانشکده مهندسی سامانه‌های هوشمند

گروه علوم داده

یادگیری ماشین


تمرین دوم

استاد درس

دکتر سامان هراتی‌زاده


زمان تحویل: 1403/8/18

1403/8/8

یادگیری ماشین – تمرین «2»		
 <p>دولاین: 1403/8/18</p>	<p>دستیاران آموزشی <u>سجاد دشتی</u></p>	<p>دکتر سامان هراتی زاده دانشگاه تهران - دانشکده سامانه‌های هوشمند نیمسال اول ۱۴۰۳-۱۴۰۴</p>

مقدمه و نکات

- در این تمرین، شما از الگوریتم‌های جنگل تصادفی (Random Forest) و آدا بوست (AdaBoost) برای طبقه بندی استفاده خواهید کرد و تاثیر پارامترهای مختلف را بر عملکرد مدل‌ها بررسی می‌کنید. همچنین، آنها را با یک مدل درخت تصمیم (Decision Tree) مقایسه می‌کنید و نتایج آنها را تحلیل خواهید کرد.
- مجموعه داده این تمرین Breast Cancer از کتابخانه scikit-learn می باشد.
- می بایست روش adaboost را خودتان پیاده سازی نمایید. می توانید از کلاس Decision tree و Random Forest از کتابخانه scikit-learn استفاده نمایید.
- آماده سازی داده ها:
مجموعه داده انتخابی خود را به سه بخش:
60% آموزش (train)
15% اعتبارسنجی (validation)
25% آزمون (test) تقسیم کنید.
- در صورت نیاز، داده ها را پیش پردازش کنید.
- تمامی اجزا کد و نتایج می بایست توسط مصححین، عیناً تکرار پذیر باشند (حتی برای بخش بندی دادگان). می‌توانید از random_seed برای تکرار پذیری استفاده کنید.
- تحلیل نتایج حائز اهمیت است. لذا در تمامی قسمت ها نتایج به دست آمده از تمرین باید حتماً در گزارش درج و تحلیل شوند.
- در حل این تمرین، شما مجاز به استفاده از مدل های زبانی بزرگ (LLM) برای کمک به نوشتن کد یا حل مسائل هستید. با این حال، شما باید به طور کامل به تمامی کدی که تحویل می‌دهید تسلط داشته باشید و قادر به توضیح عملکرد و نیز تغییر کد باشید. استفاده از ابزارها و مدل های کمکی به این معنا نیست که بتوانید بدون درک کافی از کد آن را ارائه دهید؛ هدف این است که دانش و درک عمیقی از مفاهیم و راه حل ها داشته باشید.
- توضیح مختصری از نحوه عملکرد اجزای کلیدی کد ضروری است.
- لطفاً گزارش را در فرمت مشخص شده بنویسید.
- برای سوالات خود می توانید از طریق ایمیل s.dashti.k@gmail.com و نیز گروه تلگرامی اقدام بفرمایید.

یادگیری ماشین – تمرین «2»		
 <p>دولاین: 1403/8/18</p>	<p>دستیاران آموزشی سجاد دشتی</p>	<p>دکتر سامان هراتی زاده دانشگاه تهران - دانشکده سامانه‌های هوشمند نیمسال اول ۱۴۰۳-۱۴۰۴</p>

1. مدل جنگل تصادفی (Random Forest):

مدل Random Forest را می‌بایست به ازای تغییر 3 هاپیرپارامتر زیر آموزش دهید تا بهترین ترکیب هاپیرپارامتر های گفته شده را بدست آورید (از شاخص Gini استفاده نمایید) و برای هر ترکیب با 10 random seed متفاوت آموزش دهید و میانگین accuracy بدست آمده روی داده های ولیدیشن را به عنوان accuracy مربوط به آن ترکیب از هاپیرپارامتر ها در نظر بگیرید.

- به ازای مقادیر از 5 تا 155 برای تعداد درخت‌ها (می‌توانید حداکثر از گام های 10 تایی برای افزایش تعداد درخت ها استفاده کنید)

- عمق درختان از 1 تا 26 (می‌توانید حداکثر از گام های 5 تایی استفاده کنید)

- تعداد ویژگی های مورد استفاده در هر split (یا همان max_features) از مقادیر 1 تا 26 (می‌توانید حداکثر از گام های 5 تایی استفاده کنید)

از میان مدل های ساخته شده، مدلی که بهترین accuracy را بر روی داده‌های ولیدیشن دارد، انتخاب کنید.

عمق و تعداد درختان و تعداد ویژگی های مورد استفاده در split مربوط به مدل منتخب را گزارش کنید. (هاپیرپارامتر های بهینه)

مدل منتخب را بر روی داده‌های آزمون ارزیابی کنید و موارد زیر را بر اساس داده های آزمون گزارش دهید:

Precision، Recall، F1-Score برای هر دو کلاس و نیز Accuracy و confusion matrix

با استفاده از accuracy های بدست آمده، سه نمودار به شرح زیر رسم نمایید. در هر نمودار، accuracy مربوط به دادگان آموزش و ولیدیشن نمایش داده شود:

1- به ازای تعداد درختان بهینه و تعداد بهینه ویژگی های مورد استفاده، نموداری رسم کنید که محور افقی آن عمق درختان را نشان می‌دهد و محور عمودی accuracy را نشان می‌دهد. برداشت خود را از این نمودار بیان نمایید.

2- به ازای تعداد درختان بهینه و عمق بهینه، نموداری رسم کنید که محور افقی آن تعداد ویژگی های مورد استفاده را نشان می‌دهد و محور عمودی accuracy را نشان می‌دهد. برداشت خود را از این نمودار بیان نمایید.

یادگیری ماشین – تمرین «2»




ددلاین: 1403/8/18

دستیاران آموزشی
سجاد دشتی

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده سامانه‌های هوشمند
نیمسال اول ۱۴۰۳-۱۴۰۴

3- به ازای تعداد بهینه ویژگی‌های مورد استفاده و عمق بهینه، نموداری رسم کنید که محور افقی آن تعداد درخت‌ها را نشان می‌دهد و محور عمودی accuracy را نشان می‌دهد. برداشت خود را از این نمودار بیان نمایید.

یادگیری ماشین – تمرین «2»		
 <p>دولاین: 1403/8/18</p>	<p>دستیاران آموزشی سجاد دشتی</p>	<p>دکتر سامان هراتی زاده دانشگاه تهران - دانشکده سامانه های هوشمند نیمسال اول ۱۴۰۳-۱۴۰۴</p>

2. مدل آدا بوست (AdaBoost):

الگوریتم Adaboost را پیاده سازی نمایید بطوریکه از Decision Tree به عنوان طبقه بند های پایه استفاده شود:

از روش adaboost کتاب witten استفاده نمایید. می توانید کتاب را از این [لینک](#) دانلود نمایید.

از مدل درخت تصمیم، به عنوان طبقه بند پایه برای آدا بوست استفاده کنید. مدل را به ازای تغییر 2 هاپیرپارامتر زیر آموزش دهید تا بهترین ترکیب هاپیرپارامتر های گفته شده را بدست آورید (از شاخص Gini استفاده نمایید) و برای هر ترکیب با 10 random seed متفاوت آموزش دهید و میانگین accuracy بدست آمده روی داده های ولیدیشن را به عنوان accuracy مربوط به آن ترکیب از هاپیرپارامتر ها در نظر بگیرید.

- به ازای مقادیر از 5 تا 155 برای تعداد درخت ها (می توانید حداکثر از گام های 10 تایی برای افزایش تعداد درخت ها استفاده کنید)

- عمق درختان از 1 تا 26 (می توانید حداکثر از گام های 5 تایی استفاده کنید)

از میان مدل های ساخته شده، مدلی که بهترین accuracy را بر روی داده های ولیدیشن دارد، انتخاب کنید.

عمق و تعداد درختان مربوط به مدل منتخب را گزارش کنید. (هاپیرپارامتر های بهینه) مدل منتخب را بر روی داده های آزمون ارزیابی کنید و موارد زیر را بر اساس داده های آزمون گزارش دهید:

Precision، Recall، F1-Score برای هر دو کلاس و نیز Accuracy و confusion matrix

با استفاده از accuracy های بدست آمده، دئومودار به شرح زیر رسم نمایید. در هر نمودار، accuracy مربوط به دادگان آموزش و ولیدیشن نمایش داده شود:

1- به ازای تعداد درختان بهینه، نموداری رسم کنید که محور افقی آن عمق درختان را نشان می دهد و محور عمودی accuracy را نشان می دهد. برداشت خود را از این نمودار بیان نمایید.

2- به ازای عمق بهینه، نموداری رسم کنید که محور افقی آن تعداد درخت ها را نشان می دهد و محور عمودی accuracy را نشان می دهد. برداشت خود را از این نمودار بیان نمایید.

یادگیری ماشین – تمرین «2»



ددلاین: 1403/8/18

دستیاران آموزشی
سجاد دشتی

دکتر سامان هراتی زاده
دانشگاه تهران - دانشکده سامانه‌های هوشمند
نیمسال اول ۱۴۰۳-۱۴۰۴

مقایسه و تحلیل:

دو مقایسه زیر را انجام دهید و با استفاده از آنها، عملکرد Random Forest و Adaboost را مقایسه نمایید.

-نمودار های شماره 1 از سوال 1 و نمودار شماره 1 از سوال 2، برای دادگان ولیدیشن را مقایسه نمایید. (این دو نمودار را روی یک نمودار نشان دهید)

-نمودار های شماره 2 از سوال 1 و نمودار شماره 2 از سوال 2، برای دادگان ولیدیشن را مقایسه نمایید. (این دو نمودار را روی یک نمودار نشان دهید)