دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

تمرین دوم درس یادگیری ماشین

عنوان تمرین : Classification بخش اول

استاد درس :

دکتر اکبری

تاریخ انتشار : 12 آبان

تاریخ برگزاری کلاس رفع اشکال : 20 آبان

مهلت تحویل : 26 آبان

تاریخ برگزاری کارگاه : 4 آذر

بخش اول : سؤالات تشریحی

۱) جدول زیر را در نظر بگیرید.

سرطان	سابقه خانوادگی سرطان	سن	
بله	بله	جوان	۱
بله	خیر	پیر	۲
بله	بله	پیر	۳
بله	خیر	جوان	۴
بله	خیر	میانسال	۵
خیر	بله	پیر	۶
خیر	بله	میانسال	۷
خیر	خیر	میانسال	۸
بله	بله	جوان	۹
خیر	بله	میانسال	۱۰
خیر	بله	پیر	۱۱
بله	خیر	میانسال	۱۲
خیر	بله	پیر	۱۳
بله	خیر	پیر	۱۴
خیر	خیر	میانسال	۱۵
خیر	بله	میانسال	۱۶
بله	بله	پیر	۱۷
بله	بله	جوان	۱۸
؟	خیر	جوان	۱۹
؟	خیر	پیر	۲۰

الف) افراد ۱ تا ۱۳ را دادهٔ آموزش و افراد ۱۴ تا ۱۸ را داده تست در نظر بگیرید. درخت تصمیم را برای این جدول رسم کنید. در هنگام رسم، بهرهٔ اطلاعات را در هر مرحله، برای هر ویژگی محاسبه کنید. در انتها به کمک درخت تصمیم ایجاد شده، وضعیت سرطان افراد ۱۹ و ۲۰ را پیش‌بینی کنید.

ب) مقدار صحت (precision) و دقت (accuracy) را برای داده‌های آزمون محاسبه کنید.

۲) به موارد زیر در مورد درخت تصمیم پاسخ دهید.

الف) یکی از مشکلات درخت تصمیم، بالا بودن خطای واریانس آن است. توضیح دهید جنگل تصادفی چگونه این مشکل را حل می‌کند.

ب) آیا ساخت درخت تصمیم به طور حریصانه و یا با کمک‌گرفتن از معیارهایی همچون بهره اطلاعاتی، همیشه بهترین درخت را به ما می‌دهد؟ توضیح دهید.

۳) سؤالات زیر را در مورد درخت تصادفی پاسخ دهید.

الف) برای یک مسئله دسته‌بندی از Random Forest استفاده کردیم، اما جواب خوبی به دست نیامد. برای بهبود این روش دو پیشنهاد داریم: ۱) افزایش عمق درخت‌ها. ۲) افزایش تعداد درخت‌ها. این دو روش برای بهبود را با یکدیگر مقایسه کنید و بیان کنید هرکدام در چه شرایطی می‌توانند مفید باشند؟

ب) توضیح دهید چرا Random Forest این امکان را به ما می‌دهد تا با ثابت ماندن عمق درخت‌ها، عملکرد بهتری در دسته‌بندی داشته باشیم.

بخش دوم : Reading Assignment

در این بخش باید درباره یکی از موضوعات زیر که در کلاس تدریس نشده، تحقیق و مطالعه کنید. هدف از این تمرین آشنایی با روند یادگیری مطالب جدید و همین‌طور ارائه این مطالب است. گزارش تحویلی شما باید بین دو الی سه صفحه باشد.

۱. **Decision Tree Pruning (link)**
۲. **Random Forest Regressor (link (turn on your vpn))**
۳. **Explained AI using Random Forest (link / link)**

- نحوه تخصیص موضوعات به شکل تصادفی بوده و از طریق فرمول (3 % n - 3) به دست می‌آید که n دو رقم آخر شماره دانشجویی شما است.
- می‌توانید از هر منبع دیگری در کنار منابع کمکی استفاده کنید.

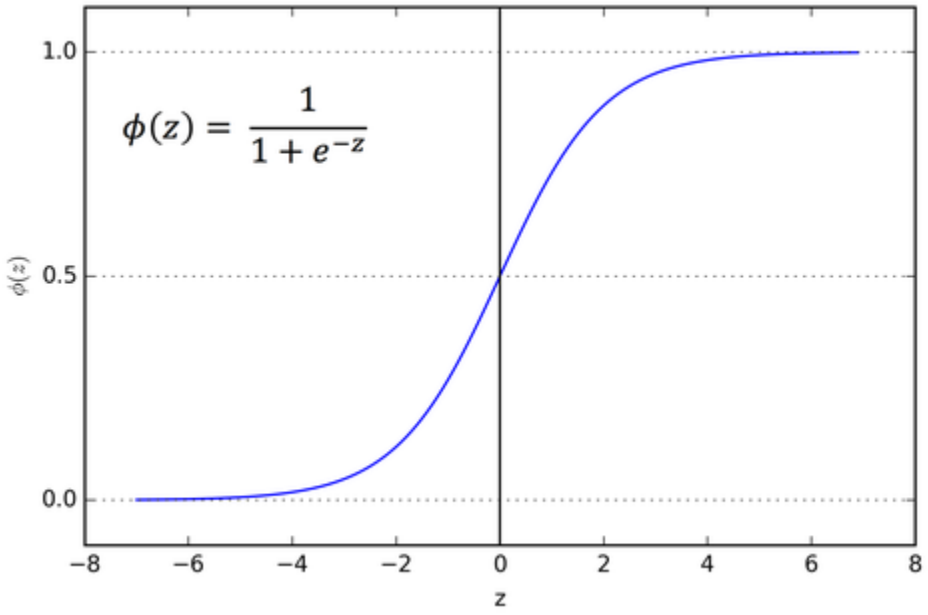
بخش سوم : پیاده‌سازی

بخش پیاده‌سازی این تمرین از سه قسمت تشکیل شده که در ادامه به توضیح هر قسمت می‌پردازیم.

قسمت اول : Logitic Regression from scratch vs Sklearn

در این قسمت از شما خواسته شده که همانند تمرین سری اول مدل Logistic Regression را خودتان پیاده‌سازی کنید و نتایج مدل پیاده‌سازی شده را با مدل از پیش تعریف شده Sklearn مقایسه کنید. برای پیاده‌سازی این قسمت به موارد زیر توجه کنید.

- ✓ از الگوریتم Gradient Descent برای بهینه‌سازی مقدار وزن‌ها استفاده کنید.
- ✓ مدل را به شکل یک Class پیاده‌سازی کنید. کلاس پیاده‌سازی شده باید شامل method های مختلفی مثل compute_loss() ، gradient_descent() ، fit() ، predict() و ... باشد.
- ✓ همان‌طور که می‌دانید، مدل‌های Logistic Regression و Linear Regression شباهت بسیار زیادی به یکدیگر دارند و به‌نوعی می‌توان گفت مدل Logistic Regression صرفاً به کمک تابع Sigmoid مقدار پیش‌بینی شده را به بازه [0,1] می‌برد؛ بنابراین می‌توانید از توابع پیاده‌سازی در تمرین قبلی ایده بگیرید و در صورت نیاز آن‌ها را تغییر ایده بگیرید. طبیعتاً برخی از توابع مثل compute_loss() ، gradient_descent() و ... نیازمند تغییر خواهند بود.



- ✓ برای این قسمت، یک مجموعه‌داده برای پیش‌بینی بیماری قلبی در نظر گرفته شده است که این مجموعه‌داده را می‌توانید از پوشه Data و در فایل heart.csv مشاهده کنید. همین‌طور جزئیات بیشتری از ستون‌های این مجموعه‌داده در فایل heart_Description آمده است.
- ✓ بعد از پیاده‌سازی مدل، از 80 درصد مجموعه‌داده برای آموزش و از 20 درصد مجموعه‌داده برای ارزیابی عملکرد مدل استفاده کنید. برای تقسیم کردن مجموعه‌داده می‌تواند از تابع آماده کتابخانه Sklearn استفاده کنید. هنگام استفاده از این تابع دقت کنید که توزیع برچسب‌ها در مجموعه‌داده آموزش و تست یکسان باشد. در صورت نیاز می‌توانید مجموعه‌داده را پیش‌پردازش کنید.
- ✓ عملکرد مدل پیاده‌سازی شده را با مدل از پیش تعریف شده کتابخانه Sklearn به کمک معیارهای ارزیابی موجود برای مسائل طبقه‌بندی و همین‌طور ماتریس آشفتگی مقایسه کنید و نتایج به‌دست‌آمده را گزارش دهید.
- ✓ همان‌طور که می‌دانید در این مسئله خروجی تابع sigmoid مقدار احتمال متعلق بودن به برچسب 1 را نشان می‌دهد و روند کار به این صورت است که با تعیین یک threshold (در مسائل binary-classification معمولاً 0.5 در نظر گرفته می‌شود) پیش‌بینی انجام می‌شود. نتایج به‌دست‌آمده را به‌ازای threshold های مختلف بررسی کنید و مقدار threshold بهینه را گزارش دهید.
- ✓ معمولاً زمانی که تعداد نمونه‌های مجموعه‌داده کم هستند، بهتر است از رویکرد k-fold-cross-valdation برای ارزیابی عملکرد مدل استفاده شود. این رویکرد را به‌عنوان یکی از method های کلاس تعریف کنید و ارزیابی را با این رویکرد انجام دهید. انتظار می‌رود تابعی که تعریف می‌کنید مدل، مجموعه‌داده و مقدار k را به‌عنوان ورودی دریافت کند و ارزیابی را انجام دهد (**امتیازی**).

قسمت دوم : Multi-Class Classification

در این قسمت از شما خواسته شده که مسئله طبقه‌بندی چند کلاسه را حل کنید. در این قسمت می‌توانید از مدل از پیش تعریف کتابخانه Sklearn استفاده کنید. هدف از این قسمت مقایسه عملکرد سه مدل جنگل تصادفی، درخت تصمیم و Logistic Regression روی مجموعه‌داده خواهد بود.

برای پیاده‌سازی این قسمت مسئله طبقه‌بندی شیشه‌ها بر اساس خواص شیمیایی آن‌ها در نظر گرفته شده که مجموعه‌داده مربوط به قسمت را می‌توانید در پوشه Data و فایل Glass.csv مشاهده کنید. این مجموعه‌داده از 214 و 7 کلاس مختلف تشکیل شده که هریک از کلاس‌ها مربوط به یکی از انواع شیشه‌ها است. در صورت نیاز به اطلاعات بیشتر می‌توانید فایل Glass_Description را مطالعه کنید.

برای پیاده‌سازی این قسمت به موارد زیر توجه کنید.

- ✓ در صورت نیاز می‌توانید قبل از آموزش مدل‌ها پیش‌پردازش‌های مختلفی مثل نرمال کردن مقدار ویژگی‌ها و ... روی مجموعه‌داده انجام دهید.
- ✓ همان‌طور که در قسمت قبل گفته شد هنگامی که تعداد نمونه‌های کافی برای ارزیابی مدل وجود ندارد رویکرد k-fold-cross-valdation یکی از رویکردهایی که برای ارزیابی به کار می‌رود. در این قسمت برای ارزیابی عملکرد مدل‌ها از دو رویکرد استفاده کنید. ابتدا مجموعه‌داده را به کمک تابع آماده کتابخانه Sklearn با نسبت 20-80 به مجموعه‌داده آموزش و تست تقسیم کنید و نتایج به‌دست‌آمده را گزارش کنید. دقت کنید که تقسیم‌بندی به صورتی انجام شود که توزیع برچسب‌ها در مجموعه‌داده آموزش و تست یکسان باشد. سپس از رویکرد k-fold-cross-validation برای ارزیابی استفاده کنید و تحلیل خودتان را از نتایج به‌دست‌آمده گزارش دهید.
- ✓ همان‌طور که می‌دانید، مدل‌های درخت تصمیم و جنگل تصادفی سریعاً با مشکل overfit شدن مواجه خواهند شد. برای جلوگیری از مشکل هایپرپارامترهای مدل را تنظیم کنید.
- ✓ برای هر یک از مدل‌ها ماتریس آشفتگی و معیار ارزیابی را به دست آورید و عملکرد مدل‌ها را با دو رویکرد ذکر شده مقایسه کنید.

قسمت سوم : Fraud Detection

در این قسمت از شما خواسته شده یک مدل جنگل تصادفی برای شناسایی تراکنش‌های جعلی پیاده‌سازی کنید. برای این قسمت مجموعه‌داده Credit Card Fraud Detection در نظر گرفته شده که می‌توانید آن را از این لینک دریافت کنید. این مجموعه‌داده شامل جزئیات مربوط تراکنش‌های انجام شده توسط کارت های اعتباری است که ستون Trarget حاوی برچسب و سایر ستون ها حاوی ویژگی‌هایی هستند که می‌توانید برای پیاده‌سازی مدل از آن‌ها استفاده کنید. در صورت نیاز به اطلاعات بیشتر می‌توانید فایل Fraud_detection_description را مطالعه کنید. هنگام پیاده‌سازی این قسمت به موارد زیر توجه کنید.

- ✓ برای ارزیابی عملکرد مدل پیاده‌سازی شده از 25 درصد مجموعه‌داده استفاده کنید. برای تقسیم‌بندی مجموعه‌داده می‌توانید از تابع آماده sklearn استفاده کنید. فقط دقت کنید که تا حد امکان لیبل های مجموعه‌داده آموزش و تست توزیع یکسانی داشته باشند.
- ✓ توجه کنید که در برخی از ستون این مجموعه‌داده ممکن است مقادیر null وجود داشته باشد و برای آموزش و ارزیابی مدل، باید آن‌ها را با مقادیر مناسبی جایگزین کنید یا آن ستون را حذف کنید.

- ✓ باتوجه‌به ابعاد بالای مجموعه‌داده و همین‌طور باتوجه‌به imbalance توزیع برچسب‌ها، باید پیش‌پردازش خوبی روی مجموعه‌داده انجام شود. روندی (روشی) که برای انتخاب ستون‌هایی که برای آموزش مدل از آن‌ها استفاده کردید را در گزارش ذکر کنید.
- ✓ برای ما در این مسئله عملکرد مدل روی برچسب‌های 1 اهمیت بیشتری دارد. هنگام ارزیابی عملکرد مدل، معیارهای ارزیابی precision ، recall ، f1-score را به‌صورت جداگانه و برای برچسب‌های صفر و یک محاسبه کنید. برای سادگی کار می‌توانید از کتابخانه sklearn کمک بگیرید.
- ✓ تحلیل خودتان را از بررسی ماتریس آشفتگی مدلی که پیاده‌سازی کردید، گزارش دهید و برای رسیدن به بهترین نتیجه هایپرپارامترهای مدل را تنظیم کنید.
- ✓ شما می‌توانید به کمک predict_proba() مقدار احتمال متعلق بودن یک نمونه به کلاس یک را پیش‌بینی کنید. همان‌طور که در قسمت اول گفته شد threshold=0.5 لزوماً پیش‌بینی خوبی ارائه نمی‌دهد. برای این مسئله مقدار threshold بهینه را پیدا کنید و مقدار آن را به همراه معیاری که برای ارزیابی عملکرد مدل در نظر گرفتید را گزارش دهید.
- ✓ feature_importances_ یکی از attribute های مفید کلاس RandomForestClassifier() که کاربرد زیادی در انتخاب ویژگی‌ها دارد. به کمک این attribute ده ویژگی‌ای که بیشترین تأثیر را در پیش‌بینی مدل داشته‌اند، گزارش دهید و تحلیل خودتان را درباره دلیل بالاتر بودن اهمیت این ویژگی‌ها بنویسید. آیا بالا بودن اهمیت این ویژگی ها منطقی است.
- ✓ یک‌بار دیگر مدل را روی 10 ویژگی ای که بیشترین تأثیر را داشته‌اند آموزش دهید و نتایج جدید را بررسی کنید. به نظر شما برای این مسئله بهترین روش برای انتخاب ویژگی ها به چه شکل است ؟ آیا نیاز است از تمامی ویژگی ها استفاده کرد یا می توان با تعداد کمتری از ویژگی ها به نتیجه خوبی رسید ؟!

معیار ارزیابی شما

بخش اول سؤالات تشریحی (۲۰ نمره) : سؤال اول (۱۰ نمره) / سؤال دوم و سوم (هرکدام ۵ نمره)

بخش دوم Reading Assignment (۲۰ نمره)

بخش سوم پیاده‌سازی (۱۶۰ نمره) :

- قسمت اول (۴۵ نمره)
- قسمت دوم (۳۰ نمره)
- قسمت سوم (۸۵ نمره)
- قسمت امتیازی (۱۰ نمره)

این تمرین حدود ۲ نمره از ۲۰ نمره نهایی شما را شامل می‌شود.

نکات تکمیلی

- ✓ با توجه به حجم بالای مجموعه داده قسمت سوم، برای پیاده سازی این تمرین ترجیحا از **Google Colab** استفاده کنید و مجموعه داده را مستقیما به کمک دستور .... gdown --id! دانلود کنید. درنهایت بعد از اتمام پیاده سازی می توانید نوت بوک تمرین را مستقیما از colab یا از google drive خودتان دانلود کنید.
- ✓ انجام این تمرین بسته به تسلط شما به مطالب درس و زبان پایتون حداقل بین ۴ الی ۷ روز از وقت مفید شما را خواهد گرفت. به همین علت انجام این تمرین را به‌روزهای پایانی موکول نکنید. همین‌طور باتوجه‌به برنامه فشرده کلاس و حجم زیاد مطالب، مهلت تحویل این تمرین تمدید نخواهد شد.
- ✓ ارسال گزارش اجباری است. نکته مهم در گزارش‌نویسی و سؤال تشریحی روشن بودن پاسخ است نه حجم زیاد، اگر فرضی برای حل سؤال استفاده می‌کنید حتماً آن را ذکر کنید، و پاسخ نهایی را به‌صورت واضح بیان کنید. گزارش کد و پاسخ سؤال تشریحی به‌صورت فایل pdf باشد.
- ✓ هرگونه شباهت در گزارش و پاسخ تشریحی به منزله تقلب است و کل نمره تمرین را نخواهید گرفت.
- ✓ فایل pdf مربوط به بخش اول و دوم تمرین و همین‌طور گزارش مربوط به بخش پیاده‌سازی را به همراه کدها را به‌صورت یکجا در قالب یک فایل zip در سامانه کورسز آپلود کنید (نام فایل = شماره دانشجویی)
- ✓ در صورت هرگونه ابهام درباره این تمرین می‌توانید در کلاس رفع اشکال سؤالات خودتان رو بپرسید و یا از طریق ایمیل‌های زیر با ما در ارتباط باشید.

مریم نظرلو : maryamnazarloo@aut.ac.ir

فاطمه رجبی : imfatemerajabi@gmail.com

محمدعلی سفیدی اصفهانی : mohammadali.esfahani@aut.ac.ir

با آرزوی سلامتی و موفقیت برای شما عزیزان