هوش مصنوعی **تمرین شماره‌ی 4**

مدرسین: **دکتر فدایی و دکتر یعقوب‌زاده**

طراحان: پرهام سازدار، اولدوز نیساری، محمد امانلو

مهلت تحویل: 19 اردیبهشت ماه ۱۴۰۳، ساعت ۲۳:۵۹

بخش کتبی

مبحث اول

سوال اول

نادقیقی، تمیز نبودن ،نقص داشتن و وجود داده های خارج از محدوده و نا معقول من جمله مشکلاتی هستند که در داده هایی که از دنیا واقعیت به دست می آوریم وجود دارد . به نظر شما برای حل مشکل زیر چه راهکارهایی ارائه می شود ؟

1) وجود نداشتن یک یا چند ویژگی در داده های آموزش
2) نا متعادل بودن توزیع داده ها در کلاس ها
3) وجود نویز در داده ها
4) وجود ویژگی های correlated (ویژگی هایی که ارتباط زیاد با یکدیگر دارند)

سوال دوم

یک مشاور تحصیلی در حال بررسی روی یک مجموعه داده درباره ساعت مطالعاتی دانشجویان و نمرات آزمون هایشان است . او توانسته‌است معادله رگرسیون خطی زیر را با توجه به داده های موجود به دست آورد :

نمره آزمون = 60 + 5 * ساعت مطالعاتی

اما باتوجه به تاثیر انکار ناپذیر آزمون دادن در آمادگی دانشجویان او قصد دارد که نقش این مسئله را هم در نمره آزمون در نظر بگیرد . به نظر شما او چه مدل ریاضیاتی برای درک ارتباط بین این دو ویژگی پیشنهاد خواهد داد ؟ با استفاده از least square method سعی کنید توضیح دهید چگونه ضرایب مناسب را پیدا می کند ؟ اگر از gradient descent استفاده کند چطور ؟ آیا تکنیک دیگری برای کم کردن اختلاف مجموع مربعات و مقادیر مشاهده‌شده می‌شناسید؟

سوال سوم

ارزیابی مدلی که برای پیش بینی استفاده کرده ایم بسیار ضروری است . فرض کنید برای پیش بینی spam بودن از مدل بر مبنای logistic regression زیر استفاده کرده ایم :

$$p(Spam) = \frac{1}{1+e^{-(ß_0+ß_1 * email\ Length)}}$$

و فرض کنید نتایج به صورت :

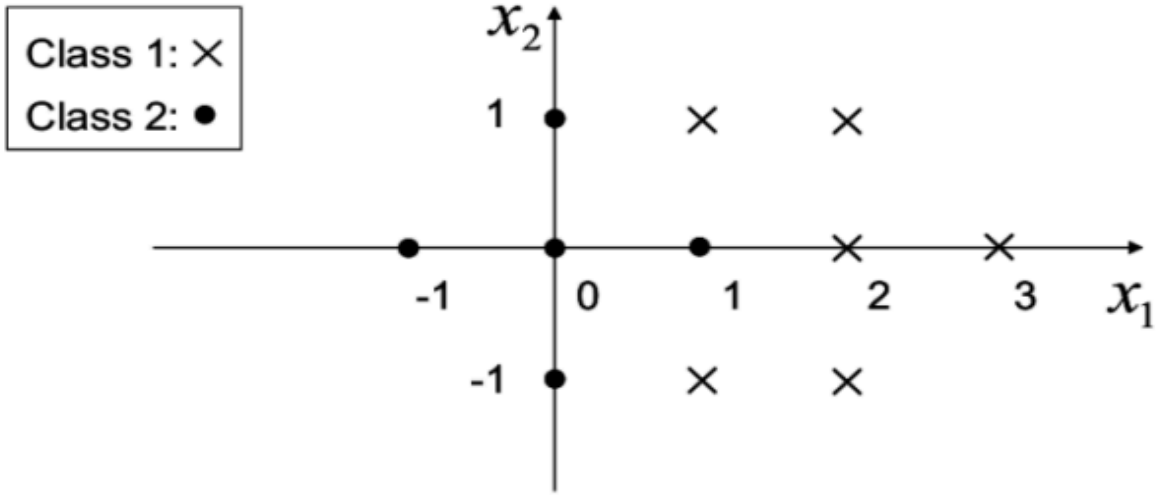
TP : 300 , TN :200 , FP :30 , FN : 20

باشد .
برای ارزیابی این مدل ابتدا confusion matrix را رسم کنید ، سپس , recall , precision ، accuracy F1-score را به دست آورید .

KNN

سوال اول

در تصویر زیر تعدادي نمونه از دو کلاس مختلف مشخص شده اند. داده ي تست (0.5,0) را با روش KNN طبقه بندي کنيد .



K نزدیک ترین همسایه با K=3 با دو فاصله ي زیر:

1. فاصله اقلیدسی
2. فاصله منهتن

Support Vector Machine

سوال اول

- به چه نقاطی support vector گفته می شود و آن را روی مثالی دلخواه نمایش دهید.
- به نظر شما طبقه بند SVM برای طبقه بندی چه نوع داده هایی مناسب نیستند؟

- درباره kernel ها و نقش آن ها در طبقه بندی توضیح دهید.(توضیح دهید وظیفه kernel ها چیست و چجوری به طبقه بندی کمک می کنند)
- تفاوت soft svm classifier با hard svm classifier بیان کنید.
- نحوه استفاده از SVM در مسائل رگرسیون رو را با کشیدن شکل توضیح دهید.

بخش عملی

مقدمه

هدف این تمرین، آشنایی با روش‌های یادگیری ماشین[1] جهت پیش‌بینی قیمت مسکن در شهر بوستون آمریکا است. این تمرین از سه بخش اصلی و یک بخش اختیاری تشکیل شده است؛
در بخش اول به آشنایی با داده‌ها پرداخته و با توزیع، انواع داده‌های موجود در مجموعه داده‌ها و اطلاعات آماری مربوط به مجموعه داده‌ها آشنا می‌شویم. به طور کلی این بخش را تحلیل دادگان[2] می‌گویند، برای آشنایی بیشتر با چیستی و نحوه عملکرد این بخش می توانید از این لینک استفاده کنید.
بخش دوم که مهمترین بخش در یک پروژه یادگیری ماشینی است، بخش پیش‌پردازش دادگان است. در این بخش، با استفاده از نتایج و تحلیل های بخش قبلی، دادگان دنیای واقعی را به داده‌ای قابل پردازش و مناسب برای عملکرد یک مدل یادگیری ماشین تبدیل می‌کنیم. برای آشنایی بیشتر با این بخش می‌توانید از این لینک استفاده کنید.
در بخش سوم به ایجاد مدل‌های مختلف یادگیری ماشین و در نهایت ارزیابی آن‌ها می‌پردازیم. در این بخش که شامل 6 فاز مستقل است. ابتدا به ساخت یک مدل Linear Regression مرتبه اول به صورت دستی(بدون استفاده از مدل آماده) می‌پردازیم، سپس متد گرادیان کاهشی[3] و polynomial regression را پیاده سازی کرده و سپس با کمک کتابخانه Scikit-Learn اقدام به تخمین قیمت خانه‌ها می‌کنیم. در فازهای بعدی نیز با مدل‌های پیشرفته‌تر آشنا شده و با استفاده از کتابخانه‌ها به پیاده سازی آن‌ها می‌پردازیم. در نهایت نیز با ارزیابی تمام مدل‌ها استنتاج نهایی را انجام می‌دهیم.
در فاز اول و دوم این بخش لازم است که فایل نوت‌بوک قرار داده شده در سایت را دانلود کرده و بخش‌های مشخص شده را کامل نمایید. پیاده‌سازی فاز سوم نیز در ادامه آن‌ها و در همان نوت‌بوک انجام می‌شود.
لازم به ذکر است تمیزی پیاده‌سازی کدها، استفاده از شیئ‌گرایی در پیاده‌سازی مدل‌ها و توابع و دسته بندی منظم آن‌ها حائز نمره امتیازی خواهد بود. برای آشنایی بیشتر با شیئ‌گرایی در پروژه‌های یادگیری ماشین می‌توانید از این لینک استفاده کنید.

---

[1] Machine Learning
[2] Data analytics
[3] Gradient Descent

بخش اول: آشنایی با مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار دارد، شامل اطلاعات مربوط به قیمت خانه های شهر بوستون به همراه ویژگی‌های خانه‌ها می‌باشد. در این تمرین می‌خواهیم با تحلیل ویژگی‌های خانه‌ها، معیارها و وابستگی‌ها را درک کنیم و بتوانیم قیمت خانه ها را بر اساس ویژگی هایشان درک کنیم.

توضیح ستون های این مجموعه داده در جدول زیر قرار داده شده است:

نام ستون	توضیحات
CRIM	per capita crime rate by town.
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of nonretail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 (otherwise
NOX	nitric oxides concentration (parts per 10 million).
RM	average number of rooms per dwelling.
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers.
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	.Bk – 0.63)^2 where Bk is the proportion of blacks by town)1000
LSTAT	% lower status of the population
MEDV (Target)	Median value of owner-occupied homes in $1000s

بررسی مجموعه داده

وقتی یک پروژه یادگیری را شروع می‌کنیم،. داده‌هایی که در ابتدا با آن‌ها شروع می‌کنیم، داده های خام هستند لذا نیاز داریم که آن ها را تجزیه و تحلیل کنیم و یک دید کلی نسبت به داده ها به دست آوریم و با ویژگی های آن ها آشنا شویم.
به فاز اولیه تجزیه و تحلیل داده ها اصطلاحا EDA می‌گویند.
برای اجرا این فاز گام های زیر را انجام دهید:

۱. ساختار کلی داده‌ها را به دست آورید. (برای این کار می توانید از متدهای info و describe استفاده کنید.)
۲. ممکن است برخی ستون‌های جدول دارای داده‌های از دست رفته باشند، تعداد و نسبت این داده‌ها را به دست بیاورید.
۳. نمودار تعداد مقدارهای منحصر به فرد برای هر ویژگی را رسم کنید و درباره آن‌ها توضیح دهید.
۴. نمودار وابستگی[4] ویژگی‌ها به یکدیگر را رسم کنید. درباره این نمودار توضیح دهید و بگویید کدام ویژگی‌ها وابستگی بیشتری به ستون هدف دارند؟
۵. نمودارهای scatter و hexbin معمولا برای بررسی ارتباط ویژگی‌ها استفاده می شوند از این نمودارها برای بررسی وابستگی‌ها با ستون هدف استفاده کنید. کاربرد و چیستی هر یک را مختصرا توضیح دهید.
۶. درباره بررسی های دیگری که می‌توانید برای مجموعه داده از آن‌ها استفاده کنید تحقیق کنید و یکی از آن‌ها را پیاده سازی کنید.

بخش دوم: پیش پردازش مجموعه داده

مهم ترین فاز هر پروژه یادگیری ماشین ، فاز پیش پردازش است. در این فاز فرمت داده ها را تغییر می دهیم، آن را ها را اصلاح و خلاصه می کنیم، تا بتوانیم برای آموزش یک مدل یادگیری ماشین از آن استفاده کنیم.
چرا که در دنیای واقعی، اطلاعات جمع‌آوری شده به راحتی کنترل نمی‌شوند و در نتیجه مقادیر خارج از محدوده، ناممکن، از دست رفته و به طور کلی گمراه‌کننده برای آموزش مدل در مجموعه داده‌ها وجود دارند. این فاز باعث می شود مدل کارا تری بتوانیم داشته باشیم و سرعت یادگیری بالاتر می‌رود.

۷. روش های پر کردن Missing Value را توضیح دهید و حداقل سه روش را پیاده سازی کنید. دلیل استفاده از هر روش را مختصرا ذکر کنید.
۸. آیا امکان حذف برخی ستون‌ها وجود دارد؟ چرا؟ در صورتی که این امکان وجود دارد با ذکر دلیل ستون‌های لازم را حذف کنید.
۹. کدام ویژگی‌ها را عددی و کدام‌ها را دسته‌ای می‌گویند؟ تفاوت این دو نوع از ویژگی‌ها در چیست؟ ویژگی‌های عددی و دسته‌ای را در این مجموعه دادگان مشخص کنید.
۱۰. در ویژگی‌های عددی، normalizing یا standardizing به چه منظور انجام می‌شود؟ تفاوت این دو روش در چیست؟ در این پروژه نیاز به انجام این کار هست؟
۱۱. برای ویژگی‌های دسته‌ای، که معمولا بصورت یک string یا object در مجموعه داده ذخیره شده‌اند، در آموزش مدل چه پیش‌پردازش‌هایی مفید هستند؟
۱۲. درباره داده های validation , train , test تحقیق کنید و روش های معمول تقسیم بندی را توضیح دهید. سپس دادگان خود را به این دسته‌ها تقسیم‌بندی کنید.
۱۳. درباره سایر روش‌های پیش‌پردازش تحقیق کنید. برخی از این روش‌ها را ذکر کرده و در صورت نیاز از آن استفاده کنید.

[4] Correlation

بخش سوم: آموزش، ارزیابی و تنظیم

فاز اول: Linear Regression

در این بخش بدون استفاده از کتابخانه سعی کنید که روابط ارائه شده برای رگرسیون خطی را درک کنید و بدون بهره گیری از مدل آماده آن را پیاده سازی کنید.

14. در این پروژه ما در حال پیاده‌سازی مدل‌های یادگیری ماشینی با نظارت هستیم. تفاوت این مدل‌ها با مدل‌های بدون ناظر، نیمه نظارتی و یادگیری تقویتی در چیست؟ برای هر یک مثال بزنید.
15. رگرسیون چیست و چه تفاوت‌هایی با روش‌های دسته‌بندی می‌کند؟
16. روابط ارائه شده در خصوص روش رگرسیون خطی را مختصرا توضیح دهید.
17. بخش های مشخص شده در notebook را تکمیل کنید. از آنجایی که تابع رگرسیون ساخته شده از مرتبه ۱ است، تنها یک ویژگی را می‌توان به عنوان ورودی این تابع انتخاب نمود. به نظر شما کدام ویژگی نسبت به سایر ویژگی‌ها خروجی دقیق‌تری به ما می‌دهد؟ علت انتخاب خود را توضیح دهید.
18. پس از انتخاب ویژگی مناسب از داده‌های train و پیش‌بینی داده‌های آزمون، می‌بایست معیاری برای ارزیابی کارایی خروجی بدست آمده تعیین کنیم. از آنجایی که مدل ما اصطلاحا در حال انجام task رگرسیون است و دسته بندی روی آن انجام نداده‌ایم، نمی‌توان از متدهای ارزیابی کارایی مربوط به classification استفاده کرد. درباره متدهای RMSE, MSE, RSS و R2 score مطالعه کنید و هرکدام را در گزارش خود توضیح دهید.
19. با استفاده از متد RMSE و R2 score، مقادیر پیش بینی شده را ارزیابی کنید. عملیات فوق را بر روی چند ویژگی دیگر نیز انجام دهید. از مقادیر بدست آمده چه استنباطی می‌کنید؟
20. مقادیر پیش‌بینی‌شده را با مقادیر واقعی با استفاده از scatter plot مقایسه کنید، که در آن محور x مقادیر واقعی را نشان می‌دهد و محور y مقادیر پیش‌بینی‌شده را نشان می‌دهد. همچنین خط x = y را نیز رسم کنید.

فاز دوم(اختیاری): polynomial Regression

در این قسمت، رگرسیون را با درجات بالاتر انجام می دهیم. در مرحله قبل، توانستیم با استفاده از دو معادله و دو مجهول به مقادیر بهینه وزن‌ها برسیم. با افزایش درجه به وضوح به لحاظ حل ریاضیاتی دشواری بیشتری متحمل می شویم و همین مسئله اهمیت حل گام به گام و نزدیک شدن تدریجی به وزن های بهینه را نشان می دهد. در نوت‌بوکی که در اختیارتان قرار گرفته است، یک رابطه برای محاسبه آن با استفاده از ماتریس ویژگی‌ها بیان شده است.

- ابتدا با استفاده از رابطه داده شده یک مدل رگرسیون چند جمله ای ایجاد کنید. سپس با استفاده از روش گرادیان کاهشی و توابع پیاده سازی شده در نوت بوک یک مدل چند جمله ای بسازید. دقت این دو مدل را با استفاده از متدهای پیشین سنجیده و با مدل رگرسیون خطی مقایسه نمایید.
- مقادیر پیش‌بینی‌شده را با مقادیر واقعی با استفاده از scatter plot مقایسه کنید، که در آن محور x مقادیر واقعی را نشان می‌دهد و محور y مقادیر پیش‌بینی‌شده را نشان می‌دهد. همچنین خط x = y را نیز رسم کنید.

فاز سوم: طبقه‌بندی

درخت تصمیم یک مدل پیش‌بینی است که از ساختار درختی برای تصمیم‌گیری در مورد مقدار یک متغیر هدف استفاده می‌کند. این درخت از گره‌ها و لیستی از تقسیم‌ها تشکیل شده است که به ازای هر گره، یک متغیر و یک مقدار تقسیم‌بندی انتخاب می‌شود تا داده‌ها به گره‌های فرزند تقسیم شوند. این فرآیند ادامه پیدا می‌کند تا ویژگی‌های مهم مجموعه داده درخت تصمیم را تشکیل دهند. هدف نهایی این است که با استفاده از این درخت، می‌توان پیش‌بینی‌هایی در مورد داده‌های جدید انجام داد. درخت تصمیم به دلیل قابل فهم بودن ساختار و نتایج آن، یکی از محبوب‌ترین روش‌های یادگیری ماشین است.

الگوریتم KNN یا همسایگان نزدیک‌ترین، یکی از ساده‌ترین الگوریتم‌های یادگیری ماشین است که برای دسته‌بندی و رگرسیون استفاده می‌شود. در این الگوریتم، تصمیم‌گیری بر اساس اکثریت آرای همسایه‌های نزدیک‌ترین به نمونه‌ای که می‌خواهیم طبقه‌بندی یا پیش‌بینی کنیم، انجام می‌گیرد. به بیان ساده‌تر، KNN با محاسبه فاصله بین نمونه جدید و تمام نمونه‌های موجود در داده‌های آموزش، نمونه‌هایی را که نزدیک‌ترین هستند شناسایی می‌کند و بر اساس بیشترین برچسب حاضر در همسایه‌های نزدیک، برچسب نمونه جدید را تعیین می‌کند. این الگوریتم نیاز به تنظیم پارامتر K دارد که تعداد همسایگانی را مشخص می‌کند که در تصمیم‌گیری شرکت می‌کنند.

21. مفهوم prune کردن در درخت های تصمیم گیری چیست؟ مزایا و معایب استفاده از این روش را ذکر کنید.
22. استفاده از درخت‌های تصمیم گیری چه زمانی می تواند نسبت به سایر مدل ها دارای مزیت باشد؟
23. تفاوت ذاتی طبقه‌بند KNN با سایر روش های طبقه بندی مثل شبکه های عصبی یا Logistic regression در چیست؟ (به نحوه train شدن هر کدام از طبقه‌بندها دقت کنید.)
24. در رابطه با الگوریتم one nearest neighbor تحقیق کنید و مزایا و معایب آن را ذکر کنید.
25. در رابطه با دیگر روش های سنجیدن فاصله در الگوریتم KNN تحقیق کنید و چند مورد از آن ها را بیان کنید.
26. در این فاز از پروژه، ابتدا همان ستون هدف که شامل میانگین قیمت خانه‌های تحت اشغال (MEDV) است را به 3 دسته تقسیم می‌کنیم، دو دهک بالای قیمت‌ها را خانه‌های لوکس، دو دهک پایین را خانه‌های اقتصادی و باقی خانه‌ها را بعنوان خانه‌های معمولی برچسب‌گذاری می‌کنیم و بعنوان یک ستون جدید ذخیره می‌کنیم. سپس دو مدل بر پایه Decision Trees، K-Nearest-Neighbours با استفاده از کتابخانه scikit learn پیاده‌سازی می‌کنید. سپس فراپارامترها[5] را تغییر دهید و مدل را تا حد امکان بهینه کنید. بهینه‌سازی مدل‌ها به این منظور است که تابع هزینه کمینه شود اما overfitting رخ ندهد. یکی از مدل‌های بهینه شده را که با آزمون و خطا به آن رسیده‌اید در گزارش خود نشان دهید.
27. برای هر دو این مدل‌ها با کمک تابع GridSearchCV، مقادیر بهینه برای پارامترها را بدست آورید. نحوه عملکرد این تابع را به طور مختصر توضیح دهید. و نتایج بدست آمده را با نتایج بدست آمده از مدل‌هایی که فراپارامترهای آن با آزمون و خطا بدست آمده بود، مقایسه کنید.
28. درخت تصمیم نهایی خود را رسم کنید . (برای این کار می‌توانید از plot_tree استفاده کنید.)
29. آیا در مدل های شما underfitting یا overfitting رخ داده است؟ به طور کلی چه زمانی این پدیده رخ می دهد؟ هر یک را توضیح دهید.

[5] Hyperparameters

فاز چهارم: روش های Ensemble

روش‌های Ensemble در یادگیری ماشین به مجموعه‌ای از مدل‌ها اشاره دارند که به صورت همکاری برای بهبود دقت پیش‌بینی‌ها کار می‌کنند. این روش‌ها معمولاً با ترکیب چندین مدل ساده‌تر، مدل نهایی را می‌سازند که در مجموع از هر یک از مدل‌های تکی بهتر عمل می‌کند. دو روش اصلی در متدهای Ensemble وجود دارد: Bagging و Boosting به منظور کاهش واریانس مدل‌ها استفاده می‌شود و در آن چندین نمونه از داده‌ها به طور تصادفی انتخاب شده و برای هر نمونه یک مدل ساخته می‌شود. این مدل‌ها سپس ترکیب می‌شوند تا نتیجه نهایی حاصل شود.حال ابتدا برای بررسی بهتر این مفهوم به سوالات زیر پاسخ کوتاه دهید.

30. درباره چرایی استفاده از روش های Ensemble و اینکه چرا امروزه این روش ها از اهمیت بالایی برخوردار هستند توضیح دهید.

31. مکانیزم کلی روش های Boosting و Bagging برای طبقه بندی را تشریح کنید و تفاوت‌های آن را بیان کنید.

جنگل تصادفی یکی دیگر از روش های یادگیری جمعی است که بر اساس ایده ای از تجمع از قوانین یا الگوریتم های ساده تر، به صورت تصادفی، تعدادی از مدل های یادگیری خود را اجرا می کند و سپس از ترکیب نتایج حاصل از این مدل ها برای پیش بینی مقادیر جدید استفاده می کند.

در واقع، جنگل تصادفی یک مجموعه از درخت های تصمیم است که هر کدام به صورت مستقل از دیگری آموزش داده می شوند و سپس نتایج آن ها ترکیب می شوند تا یک پیش بینی نهایی برای داده های ورودی انجام شود. این روش برای حل مسائل پیچیده و تعداد زیادی داده بسیار موثر و کارآمد است.

32. نحوه عملکرد روش جنگل‌های تصادفی را به طور مختصر توضیح دهید.

33. مفهوم Bootstrapping در جنگل‌های تصادفی چیست؟ کارکرد آن چگونه است و چگونه بر نتایج مدل تاثیرگذار است؟

34. آیا تعداد درخت‌های تصمیم گیری در جنگل تصادفی بر کارایی مدل تاثیر گذار است؟ بهترین مقدار آن به طور تجربی در حدود چه مقادیری است؟

35.استفاده از جنگل تصادفی چه زمانی مناسب نیست؟ این روش در چه زمانی توصیه می‌شود؟

36. استفاده از جنگل تصادفی چه تاثیری روی واریانس دارد؟

37. در این بخش پس از توضیح مختصری درباره هر یک از فراپارامترهای جنگل تصادفی مجددا با استفاده از تابع GridSearchCV، این مدل را آموزش داده و بهترین فراپارامترها را گزارش کنید. (نیازی به آموزش مدل و انتخاب فراپارامترها با آزمون و خطا نیست.)

امتیازی: XGBoost

در این تمرین قصد داریم تا با الگوریتم XGBoost آشنا شویم. XGBoost یک الگوریتم یادگیری ماشین است که بر پایه روش های گرادیان کاهشی است. این الگوریتم برای حل مسائل مختلف یادگیری ماشین از جمله طبقه بندی، پیش بینی و رتبه بندی مورد استفاده قرار می گیرد. XGBoost قابلیت اجرای سریع، کارآیی بالا و افزایش دقت در پیش بینی ها را دارا می باشد.

38. نحوه عملکرد XGboost را مختصرا توضیح دهید.

39. ابتدا مفهوم Gradient Boosting را توضیح دهید و سپس تفاوت بین Boosting Tree و Decision Tree را بیان کنید.

39. در این بخش پس از توضیح مختصری درباره هر یک از فراپارامترهای XGBoost با استفاده از تابع GridSearchCV که پیش‌تر با آن آشنا شدیم، این مدل را آموزش داده و بهترین فراپارامترها را گزارش کنید. (نیازی به آموزش مدل و انتخاب فراپارامترها با آزمون و خطا نیست.)

فاز پنجم: Support Vector Machine

SVM یا ماشین بردار پشتیبان یکی از روش‌های یادگیری ماشین برای طبقه‌بندی و رگرسیون است. این مدل با استفاده از یک خط یا صفحه‌ای که بین داده‌های دو دسته مختلف قرار می‌گیرد و فاصله بین این خط و نزدیک‌ترین نمونه‌ها از هر دو دسته را به حداکثر می‌رساند، کار می‌کند. این رویکرد به ایجاد یک مرز واضح بین دسته‌ها کمک می‌کند که امکان پیش‌بینی دقیق‌تر را فراهم می‌آورد. SVM برای داده‌هایی که دارای ویژگی‌های بسیاری هستند و نیازمند تفکیک دقیق‌تری می‌باشند، بسیار مفید است. در این تمرین قصد آن را داریم با مفهوم این روش بیشتر آشنا شویم.

40 . حال مراحل زیر را برای پیاده سازی این روش انجام دهید :

1. با استفاده از کتابخانه های موجود با 2 کرنل RBF و Linear داده های خود را دسته بندی کنید.
2. ماتریس Confusion و هم چنین معیار های ارزیابی مدل مثل F1 , accuracy , Recall,... را گزارش کنید و تحلیل خود را در گزارش ذکر کنید.
3. از کدام یک از روش های Grid search و Random search در اینجا بهتر است استفاده کنیم ؟
4. حال با استفاده از دو روش Random Search و Grid Search به ترتیب برای بازه دلخواه و مقادیر دلخواه و برای 2 کرنل RBF و Linear بهترین طبقه بند خود را پیدا کنید.(مجاز به استفاده از کتابخانه می باشید.)

فاز ششم: ارزیابی مدل‌ها

معیارهای زیادی برای سنجش و ارزیابی عملکرد مدلها وجود دارد. ارزیابی مدل های دسته بندی در یادگیری ماشینی به معنای ارزیابی عملکرد و کارایی مدل های مختلف است که برای دسته بندی داده ها استفاده می شوند. ارزیابی مدل های دسته بندی از اهمیت بسیاری برخوردار است زیرا به ما کمک می کند تا بتوانیم مدلی که می سازیم را با دقت بیشتری پیشرفت دهیم و اطمینان حاصل کنیم که عملکرد آن بهینه است.
با استفاده از این معیارها و ارزیابی کننده های دیگر می توان مدل های دسته بندی را مقایسه کرده و انتخاب بهترین مدل را برای مسئله خاص خود انجام داد.
برای ارزیابی مناسب از معیارهای زیر استفاده نمائید:

- ماتریس درهم‌ریختگی[6]
- Recall
- F1-Score
- Precision
- Accuracy
- میانگین‌گیری Macro و Micro و Weighted

[6] Confusion matrix

بخش اختیاری: ROC Curve

ROC Curve یک نمودار ارزیابی برای مدل های دسته بندی است که به تعیین کیفیت و عملکرد مدل در تفکیک داده ها کمک می کند. ROC Curve مخفف Receiver Operating Characteristic Curve است.
این نمودار به طور خاص مقدار عملکرد مدل را بر اساس تعداد تشخیص داده های صحیح (True Positive Rate) نسبت به تعداد تشخیص داده های نادرست (False Positive Rate) نشان می دهد. در واقع، ROC Curve نمایش دهنده تغییرات در تشخیص صحیح و نادرست مدل در سطوح مختلف آستانه (Threshold) است که برای تصمیم گیری در دسته بندی استفاده می شود.

برای درک بهتر این نمودار ابتدا به سوالات زیر پاسخ کوتاه دهید.

- منحنی ROC چیست و چگونه می‌توان آن را تفسیر کرد؟ در چه صورتی این نمودار بیانگر عملکرد بهتر مدل است؟
- مساحت زیر منحنی ROC چه اهمیتی دارد؟ درباره شاخص AUC تحقیق کنید و نحوه تفسیر و محاسبه آن را شرح دهید.
- چگونه می‌توان از منحنی ROC برای تعیین بهترین آستانه تصمیم‌گیری برای مدل دسته‌بندی استفاده کرد؟
- روشی برای رسم منحنی ROC در حالت چند کلاسه ارائه دهید.
- در این بخش با توجه به راه حلی که برای کشیدن نمودار ROC برای حالت چند کلاسه ارائه دادید منحنی ROC را برای هر قسمت رسم کنید و برای هر کلاس AUC را محاسبه کرده و نتیجه را گزارش کنید. همچنین تحلیل خود را بر روی عملکرد هر یک با توجه به معیار ROC در گزارش ذکر کنید.
- در صورتی که دادگان به طور چشمگیری نامتوازن باشند، چه راهکارهایی برای مدیریت این موضوع پیشنهاد می‌کنید؟ برای رفع این مشکل چگونه از ROC استفاده می‌کنیم؟
- اکثر مدل‌های دسته بندی قابلیت دریافت یک توزیع احتمالاتی برای کلاس‌ها را دارند. از هر یک از مدل‌ها به جای دریافت کلاس پیش‌بینی شده احتمال قرار گرفتن در کلاس‌ها را بدست آورده و سپس thereshold یا حد آستانه قرارگیری در هر کلاس را تغییر دهید. نتایج را با نتایج قبل مقایسه کنید. آیا تغییر حدود آستانه به طور یکنواخت بر تعداد مشاهدات پیش بینی شده در هر دسته تاثیر گذار است؟

نکات پایانی

- توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. از ابزارهای تحلیل داده مانند نمودارها استفاده کنید.
- تمیزی پیاده‌سازی کدها، استفاده مناسب از شیئ‌گرایی در پیاده‌سازی مدل‌ها و توابع و دسته بندی منظم آن‌ها حائز نمره امتیازی خواهد بود.
- پس از مطالعه کامل و دقیق صورت پروژه، در صورت وجود هرگونه ابهام یا سوال با طراحان پروژه در ارتباط باشید.
- نتایج، گزارش و کدهای خود را در قالب یک فایل فشرده با فرمت AI_CA3_[stdNumber].zip در سامانه ایلرن بارگذاری کنید.

- محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- دقت کنید که نیازی به آپلود مجموعه داده‌ها در سامانه ایلرن نیست.