

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 توانسته است معادله رگرسیون خطی زیر را با توجه به داده های موجود به دست آورد:

نمره آزمون = $60 + 5 * \text{ساعت مطالعاتی}$

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 مشاهده شده می شناسید؟

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 :

$$p(\text{Spam}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * \text{email Length})}}$$

و فرض کنید نتایج به صورت :

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

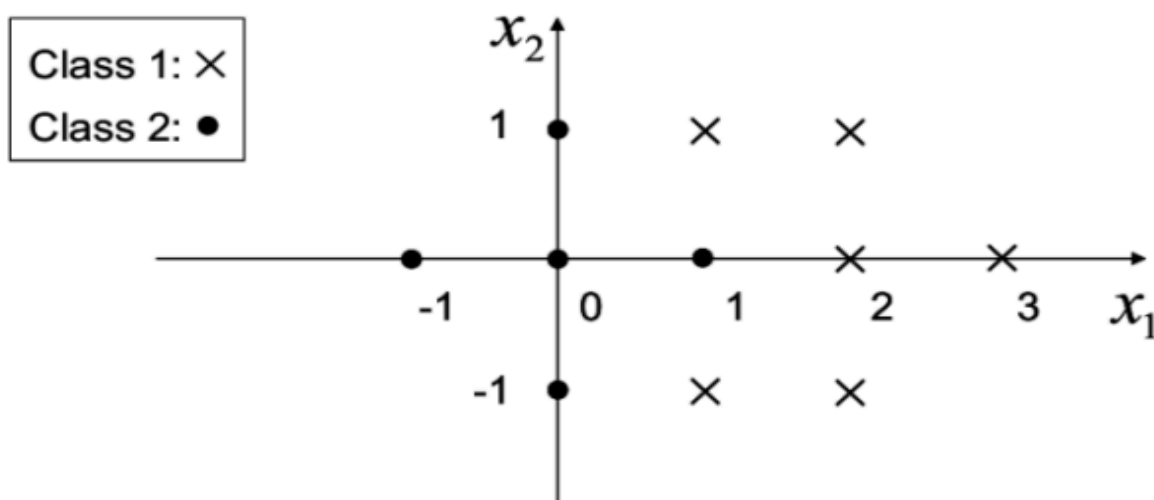
باشد .

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

KNN

سوال اول

در تصویر زیر تعدادی نمونه از دو کلاس مختلف مشخص شده اند. داده ی تست (0.5,0) را با روش KNN طبقه بندی کنید .



K نزدیک ترین همسایه با K=3 با دو فاصله ی 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 روش های یادگیری ماشین¹ جهت پیش بینی 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 داده ها پرداخته و با توزیع، انواع داده های موجود در مجموعه داده ها و اطلاعات آماری مربوط به مجموعه داده ها آشنا می شویم. به طور کلی این بخش را تحلیل دادگان² می گویند، برای آشنایی بیشتر با چیرستی و نحوه عملکرد این بخش می توانید از [این](#) لینک استفاده کنید.

بخش دوم که مهمترین بخش در یک پروژه یادگیری ماشینی است، بخش پیش پردازش دادگان است. در این بخش، با استفاده از نتایج و تحلیل های بخش قبلی، دادگان دنیای واقعی را به داده ای قابل پردازش و مناسب برای عملکرد یک مدل یادگیری ماشین تبدیل می کنیم. برای آشنایی بیشتر با این بخش می توانید از [این](#) لینک استفاده کنید.

در بخش سوم به ایجاد مدل های مختلف یادگیری ماشین و در نهایت ارزیابی آن ها می پردازیم. 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) می پردازیم، سپس متد گرادیان کاهشی³ و polynomial regression را پیاده سازی کرده و سپس با کمک کتابخانه Scikit-Learn اقدام به تخمین قیمت خانه ها می کنیم. در فازهای بعدی نیز با مدل های پیشرفته تر آشنا شده و با استفاده از کتابخانه ها به پیاده سازی آن ها می پردازیم. در نهایت نیز با ارزیابی تمام مدل ها استنتاج نهایی را انجام می دهیم.

در فاز اول و دوم این بخش لازم است که فایل نوت بوک قرار داده شده در سایت را دانلود کرده و بخش های مشخص شده را کامل نمایید. پیاده سازی فاز سوم نیز در ادامه آن ها و در همان نوت بوک انجام می شود. لازم به ذکر است تمیزی پیاده سازی کدها، استفاده از شیء گرایی در پیاده سازی مدل ها و توابع و دسته بندی منظم آن ها حائز نمره امتیازی خواهد بود. برای آشنایی بیشتر با شیء گرایی در پروژه های یادگیری ماشین می توانید 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 خانه‌ها را بر اساس ویژگی‌هایشان 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	$1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
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 آن‌ها را تجزیه و تحلیل کنیم و یک دید کلی نسبت به داده‌ها به دست آوریم و با ویژگی‌های آن‌ها آشنا شویم.

به فاز اولیه تجزیه و تحلیل داده‌ها اصطلاحاً EDA می‌گویند.

برای اجرا این فاز گام‌های زیر را انجام دهید:

۱. ساختار کلی داده‌ها را به دست آورید. (برای این کار می‌توانید 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 مقادیرهای منحصر به فرد برای هر ویژگی را رسم کنید و درباره آن‌ها توضیح دهید.
۴. نمودار وابستگی^۴ 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 معمولاً برای بررسی ارتباط ویژگی‌ها استفاده می‌شوند از این نمودارها برای بررسی وابستگی‌ها با ستون هدف استفاده کنید. کاربرد و چیرستی هر یک را مختصراً توضیح دهید.
۶. درباره بررسی‌های 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 می‌شود مدل کارا تری بتوانیم داشته باشیم و سرعت یادگیری بالاتر می‌رود.
۷. روش‌های پر کردن Missing Value را توضیح دهید و حداقل سه روش را پیاده سازی کنید. دلیل استفاده از هر روش را مختصراً 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 معمولاً بصورت یک string یا object در مجموعه داده ذخیره شده‌اند، در آموزش مدل چه پیش‌پردازش‌هایی مفید هستند؟
 ۱۲. درباره داده‌های test , train , validation تحقیق کنید و روش‌های 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 پیاده سازی مدل های یادگیری ماشینی با نظارت هستیم. تفاوت این مدل ها با مدل های بدون ناظر، نیمه نظارتی و یادگیری تقویتی در چیست؟ برای هر یک مثال بزنید.
15. رگرسیون چیست و چه تفاوت هایی با روش های دسته بندی می کند؟
16. روابط ارائه شده در خصوص روش رگرسیون خطی را مختصراً 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 می توان به عنوان ورودی این تابع انتخاب نمود. به نظر شما کدام ویژگی نسبت به سایر ویژگی ها خروجی دقیق تری به ما می دهد؟ علت انتخاب خود را توضیح دهید.
18. پس از انتخاب ویژگی مناسب از داده های train و پیش بینی داده های آزمون، می بایست معیاری برای ارزیابی کارایی خروجی بدست آمده تعیین کنیم. از آنجایی که مدل ما اصطلاحاً در حال انجام task رگرسیون است و دسته بندی روی آن انجام نداده ایم، نمی توان از متدهای ارزیابی کارایی مربوط به classification استفاده کرد. درباره متدهای RMSE, MSE, RSS و R2 score مطالعه کنید و هر کدام 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 می کنید؟
20. مقادیر پیش بینی شده را با مقادیر واقعی با استفاده از scatter plot مقایسه کنید، که در آن محور x مقادیر واقعی را نشان می دهد و محور y مقادیر پیش بینی شده را نشان می دهد. 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 وزن ها 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 نوت بوکی که در اختیارتان قرار گرفته است، یک رابطه برای محاسبه آن با استفاده از ماتریس ویژگی ها 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 پیش بینی شده را با مقادیر واقعی با استفاده از scatter plot مقایسه کنید، که در آن محور x مقادیر واقعی را نشان می دهد و محور y مقادیر پیش بینی شده را نشان می دهد. 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 درخت‌های تصمیم‌گیری چیست؟ مزایا و معایب استفاده از این روش را ذکر کنید.

22. استفاده از درخت‌های تصمیم‌گیری چه زمانی می‌تواند نسبت به سایر مدل‌ها دارای مزیت باشد؟

23. تفاوت ذاتی طبقه‌بند KNN با سایر روش‌های طبقه‌بندی مثل شبکه‌های 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 روش‌های سنجیدن فاصله در الگوریتم KNN تحقیق کنید و چند مورد از آن‌ها 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 فرآیندها⁵ 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 فرآیندهای 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 مدل‌های شما *underfitting* یا *overfitting* رخ داده است؟ به طور کلی چه زمانی این پدیده رخ می‌دهد؟ هر یک را توضیح دهید.

⁵ Hyperparameters

فاز چهارم: روش های Ensemble

روش های Ensemble در یادگیری ماشین به مجموعه ای از مدل ها اشاره دارند که به صورت همکاری برای بهبود دقت پیش بینی ها کار می کنند. این روش ها معمولاً با ترکیب چندین مدل ساده تر، مدل نهایی را می سازند که در مجموع از هر یک از مدل های تکی بهتر عمل می کند. دو روش اصلی در متدهای Ensemble وجود دارد: Bagging و Boosting به منظور کاهش واریانس مدل ها استفاده می شود و در آن چندین نمونه از داده ها به طور تصادفی انتخاب شده و برای هر نمونه یک مدل ساخته می شود. این مدل ها سپس ترکیب می شوند 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 تفاوت های 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 جنگل های تصادفی را به طور مختصر توضیح دهید.

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

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

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

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

37. در این بخش پس از توضیح مختصری درباره هر یک از فرآیندهای جنگل تصادفی مجدداً با استفاده از تابع [GridSearchCV](#)، این مدل را آموزش داده و بهترین فرآیندها را گزارش کنید. (نیازی به آموزش مدل و انتخاب فرآیندها 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، کارایی بالا و افزایش دقت در پیش بینی ها را دارا می باشد.

38. نحوه عملکرد XGboost را مختصراً توضیح دهید.

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

39. در این بخش پس از توضیح مختصری درباره هر یک از فرآپارامترهای XGBoost با استفاده از تابع [GridSearchCV](#) که پیش‌تر با آن آشنا شدیم، این مدل را آموزش داده و بهترین فرآپارامترها را گزارش کنید. (نیازی به آموزش مدل و انتخاب فرآپارامترها 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 درهم‌ریختگی⁶
- 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 می توان 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 می توان از منحنی ROC برای تعیین بهترین آستانه تصمیم گیری برای مدل دسته بندی 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 می کنید؟ برای رفع این مشکل چگونه از ROC استفاده می کنیم؟
- اکثر مدل های دسته بندی قابلیت دریافت یک توزیع احتمالاتی برای کلاس ها را دارند. از هر یک از مدل ها به جای دریافت کلاس پیش بینی شده احتمال قرار گرفتن در کلاس ها را بدست آورده و سپس threshold 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 پیاده سازی کدها، استفاده مناسب از شیء گرایی در پیاده سازی مدل ها و توابع و دسته بندی منظم آن ها 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.