بسمه تعالی

مقدمه ای بر یادگیری ماشین

نیمسال اول ۹۹-۰۰

تمرین عملی سری سوم — مدرس: صابر صالح

- **مقدمه**

در تمرین پیشین، با تعدادی از مشکلات موجود در داده از جمله داده های از دست رفته و داده های پرت آشنا شدید و روش هایی را برای برطرف کردن این مشکلات و پیش پردازش داده ها به کار گرفتید. در این تمرین، علاوه بر آشنایی با متد متفاوتی برای پیش پردازش داده ها، با تحلیل آماری داده ها و طبقه بندی با استفاده از SVM و AdaBoost آشنا خواهید شد.

- **پیش پردازش داده ها**

۱. در تمرین سری دوم با دو روش مقابله با داده های ناموجود آشنا شدید. در اینجا می خواهیم از دو روش متفاوت برای پر کردن خانه های خالی استفاده کنیم. می خواهیم با استفاده از ستون ویژگی sulphates خانه های خالی مربوط به ویژگی PH را پر کنیم. به این منظور، میانگین PH داده هایی را که دارای میزان سولفات یکسان هستند، محاسبه کنید و از آن برای پر کردن خانه های نامشخص ستون PH استفاده کنید.

۲. در تمرین قبل با پیاده سازی Linear regression آشنا شدید. در اینجا می خواهیم با استفاده از Linear regression ، خانه های خالی یک ویژگی را با استفاده از ویژگی های دیگر پر کنیم. چهار ویژگی 'A Indx'، 'fixed acidity' ، 'citric acid' و 'chlorides' را در نظر بگیرید. می خواهیم خانه های خالی مربوط به ستون ویژگی 'A Indx' را با استفاده از ویژگی های دیگر پر کنیم. به این منظور، نمونه هایی را که خانه مربوط به ویژگی 'A Indx' در آنها خالی است، به عنوان داده های تست جدا می کنیم. سپس با استفاده از Linear Regression مقادیر ناموجود را تخمین می زنیم. با استفاده از ۳ ویژگی ارائه شده، این مراحل را انجام دهید. کدام ویژگی نتایج بهتری را ارائه خواهد کرد؟ می توانید همزمان از چند ویژگی استفاده کنید. دو معیار برای بررسی مناسب بودن ویژگی ها برای تخمین 'A Indx' ارائه کنید.

۳. در تمرین گذشته با روش تبدیل داده های categorical به داده های کمی آشنا شدید. برای استفاده از این ویژگی ها در تخمین مقادیر ناموجود، ابتدا بایستی آنها را کمی کنید. ستون ویژگی 'Vit Indx' پس از کمی کردن ۴ مقدار به خود می گیرد. برای کمی کردن این ویژگی از دو ستون جدید که هر کدام دارای عناصر ۰ و ۱ هستند، استفاده کنید. اگر برای کمی کردن این ویژگی از اعداد ۰، ۱، ۲، ۳ استفاده می کردید، چه مشکلی پیش می آمد؟ سه ویژگی categorical را با اضافه کردن ستون مناسب کمی کنید. سپس با استفاده از این سه ویژگی، مقادیر ناموجود ویژگی 'A Indx' را با استفاده از Linear Regression پر کنید. کدام ویژگی نتایج بهتری خواهد داشت؟ آیا این روش، برای تخمین مقادیر ناموجود مناسب است؟

- مرحله دوم : تحلیل آماری داده‌ها

آشنایی با تحلیل آماری

در این بخش قرار است ابتدا با برخی آماره ها آشنا شویم و با رسم تعدادی نمودار با ماهیت داده بیش تر آشنا شویم. همان طور که در بخش قبل نیز دیدید یک راه خوب برای پیدا کردن درک درست از ویژگی ها و روابط بین آن ها استفاده از نمودارها است. عملیاتی که در این بخش یاد می گیرید در مورد هر داده ای می تواند مفید باشد.

آماره های مهم:

آماره ها در واقع توابعی هستند از فضای آماری داده ها به اعداد حقیقی که پارامتری از داده را محاسبه می کنند. در اولین برخورد با یک داده باید آماره های متنوع ولی ساده ای را در مورد آن به دست آوریم و تحلیل کنیم. مهم ترین آماره ها میانگین، واریانس، مد و میانه هستند. به عنوان تمرین، آماره ها فوق را برای ویژگی های citric acid,A Indx,PH,volatile acidity یک بار به طور کلی و یک بار برای دو دسته ی Survived 0,1 محاسبه کرده و نتایج حاصل را تحلیل کنید. منظور از تحلیل این است که برای مثال بیان کنید میانگین ویژگی محصول با کیفیت بالا چه تفاوتی با محصول با کیفیت پایین دارد یا چه گزاره هایی در مورد کیفیت درست است؟

نمودارهای مهم:

با وجود این که آماره ها، اعداد بسیار مهمی در کار با داده هستند اما آن ها روح زنده ی نمودارها را با خود ندارند. در این بخش قرار است با رسم نمودارهای مختلف داده را ارزیابی کنیم. نمودارهای نام برده در هر بند را برای داده ی خواسته شده رسم کنید و نتایج آن را تحلیل کنید. در این قسمت می توانید از کتابخانه های matplotlib و seaborn استفاده کنید.

۱. با استفاده از ویژگی sulphates نمودار میله‌ای محصول با کیفیت و بی کیفیت را رسم کنید.

۲. نمودار جعبه‌ای (Boxplot) :این نمودار بر حسب مفاهیم میانه و چارک کشیده می‌شود. در نمودار جعبه‌ای چارک‌های اول و سوم و میانه به همراه داده‌ی کمینه و بیشینه رسم می شود. این نمودار شهود بسیار خوبی در مورد پراکندگی مقادیر داده می دهد. نمودار جعبه ای مقدار chlorides را برای ۴ گروه sulphates و برای دو حالت Survived 0,1 (یعنی در مجموع ۸ نمودار) رسم کنید. چه نتیجه ای میتوان گرفت؟

۳. نمودار هیستوگرام (Histogram): با این نمودار میتوان توزیع داده های عددی را در تعداد دسته بندی دلخواه مشاهده کرد. هیستوگرام 'free sulphor dioxide' را رسم کنید.یک بار نیز این کار را برای دو دسته ی محصول با کیفیت و بی کیفیت انجام دهید. حال این نمودارها را با جداسازی نوع آبمیوه بکشید تا چهار نمودار داشته باشید. مشاهدات خود را ثبت کنید.

۴. نمودار Heatmap یک نمایش گرافیکی دو بعدی از دادگان است که ارتباط ویژگی های مختلف را نشان می دهد. علاوه بر ارتباط بین هر ویژگی با کیفیت محصول، بین خود ویژگی ها نیز ممکن است ارتباط معناداری باشد. برای دریافتن این ارتباطات میتوان همبستگی یا Correlation بین ستون ها را مقایسه کرد. میتوان برای مقایسه جامع از Heatmap استفاده کرد و همبستگی بین ستون ها را نشان داد. Correlation Heat map ویژگی‌ها را رسم نمایید. کدام ویژگی‌ها دارای همبستگی زیادی هستند؟ آیا می‌توان یکی از این ویژگی‌ها را حذف کرد یا خیر؟

- مرحله سوم : AdaBoost

در این قسمت می خواهیم عملکرد طبقه بند AdaBoost بر روی داده ها را بررسی کنیم. می دانیم که در استفاده از AdaBoost باید از یک طبقه بند پایه استفاده کنیم. برای طبقه بند پایه می توان از Random Forest ، Gaussian Naive Bayes و Decision Tree استفاده کنیم. در اینجا از Decision Tree به عنوان طبقه بند پایه استفاده می کنیم. در نظر داشته باشید عمق درخت تصمیم در نتیجه طبقه بندی تاثیر گذار است.

۱. تعداد درخت ها و همچنین عمق آنها را با استفاده از 5-fold cross validation تعیین کنید.
۲. دقت طبقه بندی، F1 score و Confusion Matrix را برای پارامتر های بهینه ارائه دهید.
۳. نمودار Decision Scores دو کلاسه را برای نمونه ها ارائه کنید.

- مرحله چهارم : SVM

۱. در این بخش میخواهیم طبقه بندی قسمت قبل را توسط SVM انجام دهیم. مدنظر این است که ما بین کرنل های rbf، linear، poly، و sigmoid بهترین آن ها را انتخاب کنیم. همچین باید پارامتر C را نیز به همراه پارامتر γ (وقتی که از rbf استفاده می کنیم) tune کنیم.
بنابراین به ازای:

$$C \in \{0.1, 1, 10, 100, 1000, 10000\}$$

$$\gamma \in \{0.001, 0.01, 0.1\}$$

پاسخ نهایی شما باید تمامی طبقه بند ها را به ترتیب F1 Score آن ها از بیشترین به کمترین مرتب شده به همراه accuracy و precision آن ها نشان دهد. همچنین حتما از 5-fold cross validation استفاده کنید.
برای بهترین طبقه بند، confusion matrix را نیز نمایش دهید.

۲. از طبقه بند هایی که تا به حال دیده ایم استفاده کنید و یک hard voting classifier را به اجرا بگذارید و نتیجه را توضیح دهید. برای توضیحات بیشتر به نوت بوک مراجعه کنید.
۳. در این بخش decision boundry های SVM را برای دیتاست دایره های هم مرکز ترسیم میکنیم. توضیحات بیشتر در نوت بوک آمده است.
۴. در این بخش میخواهیم پاسخی که از تمرین تئوری برای سوال quadratic programming بدست آوردید را توسط کتابخانه‌ی cvxopt دوباره چک کنیم.