دانشکدهٔ مهندسی مکانیک
دانشگاه تهران

هوش مصنوعی و یادگیری ماشین

تکلیف دوم (روش های یادگیری ماشین)

استاد درس:

دکتر شریعت پناهی

دستیاران آموزشی:

فاطمه مجاب

سعید دلیر

تاریخ تحویل: ۱۴۰۱/۰۲/۱۰

نیمسال دوم سال تحصیلی ۰۱-۱۴۰۰

فهرست مطالب

بخش اول: پیش‌بینی بارهای گرمایشی و سرمایشی یک ساختمان

در این بخش می‌خواهیم بارهای گرمایشی و سرمایشی یک ساختمان را بر پایهٔ ویژگی‌های هشت‌گانهٔ آن ساختمان پیش‌بینی کنیم. برای این کار ابتدا مجموعه دادهٔ مورد نظر را از این لینک دانلود کنید و با ویژگی‌های داده‌ها آشنا شوید. سپس با استفاده از ویژگی‌های ۸ ستون اول (X_1 تا X_8) مقدار بارهای گرمایشی و سرمایشی ساختمان (y_1 و y_2) را با استفاده از رگرسیون چند متغیره پیش‌بینی کنید.

در بسیاری از مسائل رگرسیون به علت پیچیده بودن رابطه بین ورودی و خروجی‌ها به جای رگرسیون خطی از رگرسیون غیرخطی (چند جمله‌ای) استفاده می‌شود. این موضوع در مسائل دسته‌بندی نیز با غیرخطی شدن مرز تصمیم‌گیری خود را نشان می‌دهد.

در این گونه مسائل به جای استفاده از رابطه خطی، از یک چند جمله‌ای استفاده می‌کنیم و در واقع فضای ویژگی‌ها را به فضای مرتبه بالاتری تغییر می‌دهیم. اگر فضای ویژگی‌ها را به درجه d ببریم، تمام جملات چند جمله‌ای "حداکثر" از درجه d در فضای ویژگی‌ها وجود خواهند داشت. به عنوان مثال اگر برای رگرسیون دو متغیره فضای ویژگی‌ها را به درجه ۳ ببریم فضای ویژگی‌ها به صورت زیر خواهد بود.

$$1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3$$

مراحل اجرایی و خواسته‌ها:

۱) ضرورت Normalize کردن داده‌ها را به اختصار توضیح دهید و ستون‌های X_1 تا X_8 (feature ها) را Normalize کنید.

۲) یک مدل خطی روی مجموعه دادگان نرمالایز شده برازش کرده و مقادیر mean squared error و r2 score را گزارش کنید.

۳) یک مدل غیرخطی از **درجه ۳** روی مجموعه دادگان نرمالایز شده برازش کرده و مقادیر mean squared error و r2 score را گزارش کنید. اگر از مجموعه دادگان اصلی (بدون نرمالایز کردن) استفاده کرده بودیم این مقادیر به چه صورت تغییر می‌کردند؟

لینک‌های مفید:

راهنمایی پیاده‌سازی مدل رگرسیون:

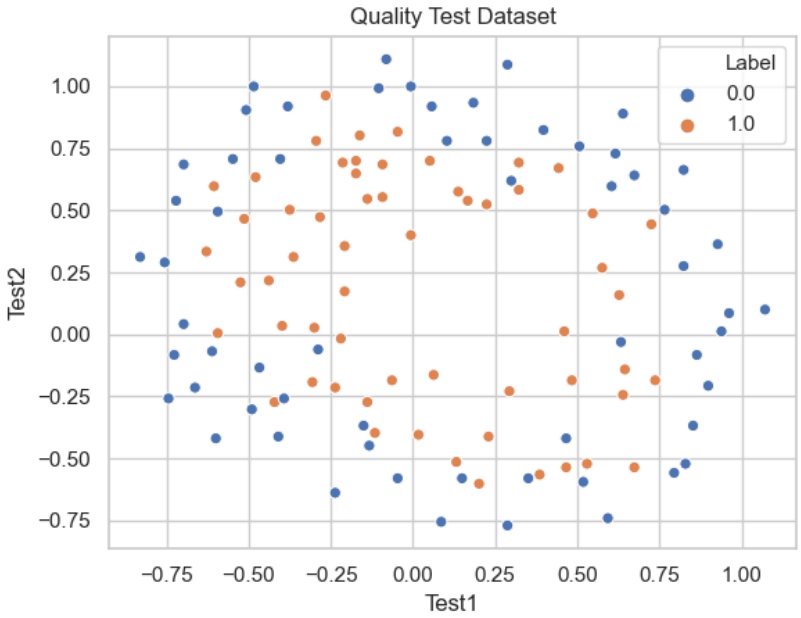
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

راهنمایی ایجاد فضای ویژگی چندجمله‌ای:

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

بخش دوم: دسته‌بندی قطعات به سالم و معیوب

در این بخش می‌خواهیم تراشه‌های تولید شده در یک خط تولید را بر پایهٔ دو ویژگی آن‌ها به دو دستهٔ سالم و معیوب دسته‌بندی کنیم. در مجموعه دادهٔ quality_test.csv دو ستون اول نشان دهندهٔ نتایج تست تراشه و ستون سوم نشان‌دهنده کیفیت آن (قابل قبول یا مردود) است.



شکل ۱: نمایش مجموعه دادهٔ quality_test.csv

همانگونه که در شکل ۱ دیده می‌شود این مجموعه داده به‌صورت خطی جدایی‌پذیر نیست، به همین دلیل همانند توضیحات بخش قبل برای تفکیک دو کلاس باید فضای ویژگی‌ها را به مرتبهٔ بالاتری برد.

مراحل اجرایی و خواسته‌ها:

۱) در مورد مفهوم Regularization و انواع آن تحقیق کنید و به اختصار توضیح دهید.

۲) فضای ویژگی‌های مجموعه داده را به درجهٔ ۷ برده و با استفاده از الگوریتم Logistic Regression و L2 Regularization دو کلاس مختلف را از هم جدا کنید. برای پارامتر C (پارامتر Regularization) سه مقدار مختلف 1، 0.01 و 10000 در نظر بگیرید و در هر حالت دقت دسته‌بند را گزارش کرده و مرز تصمیم‌گیری را رسم کنید.

۳) در مورد Cross Validation به اختصار توضیح دهید. با استفاده از روش K-fold CV درجهٔ چندجمله‌ای را از بین اعداد ۳، ۷ و ۱۰ تعیین کنید.

۴) مجموعه دادهٔ اولیه (درجه ۱) را به دو بخش آموزش و تست تقسیم کنید. سپس با استفاده از روش kNN و برای مقادیر k برابر ۱ و ۷ و ۱۳ و ۱۹ دسته‌بندی را انجام دهید. برای هر مقدار k، پارامترهای Accuracy، Precision و Recall را روی داده‌های تست گزارش کرده و نتایج را تفسیر کنید.

۵) برای بهترین مقدار k از میان مقادیر بالا، نتیجه را با استفاده از فاصلهٔ منهتن نیز گزارش کنید.

لینک‌های مفید:

راهنمایی پیاده‌سازی رگرسیون لوجستیک:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

راهنمایی پیاده‌سازی الگوریتم kNN:

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

مطالعه در مورد Cross Validation:

https://scikit-learn.org/stable/modules/cross_validation.html

مطالعه در مورد معیارهای سنجش مدل:

https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/

بخش سوم: دسته‌بندی داده‌ها به کمک ماشین بردار پشتیبان

۱) با استفاده از کتابخانه‌های مناسب ۱۰۰ داده رندم در بازه $0 < x_1 < 1$ و $-1 < x_2 < 1$ و همچنین ۱۰۰ داده رندم دیگر در بازه $2 < x_1 < 3$ و $3 < x_2 < 4$ رسم کنید. دو دسته داده به دست آمده را به کمک ماشین بردار پشتیبان (SVM) دسته‌بندی کنید. سپس دادگان و خط جدا کننده داده‌ها را نیز رسم کنید.

۲) با استفاده از کتابخانه‌های مناسب ۱۰۰ داده رندم در بازه $0 < x_1 < 2$ و $-1 < x_2 < 1$ و همچنین ۱۰۰ داده رندم دیگر در بازه $1.75 < x_1 < 3$ و $-1 < x_2 < 1$ رسم کنید. در مورد پارامتر C در SVM تحقیق کنید و به اختصار توضیح دهید. سپس با انتخاب C مناسب دسته‌بندی را انجام داده و خط جدا کننده داده‌ها را نیز رسم کنید.

۳) با استفاده از کتابخانه‌های مناسب ۱۰۰ داده رندم در بازه $1 < x_1^2 + x_2^2 < 2$ و همچنین ۱۰۰ داده رندم دیگر در بازه $4 < x_1^2 + x_2^2 < 5$ رسم کنید. به دلخواه خود دو کرنل غیرخطی انتخاب کرده و دسته‌بندی را انجام دهید و نتایج حاصل را مقایسه کنید. سپس منحنی جدا کننده داده‌ها را نیز رسم کنید.

بخش امتیازی: در صورتی که سؤالات این بخش را بدون استفاده از کتابخانه‌های آماده SVM حل کنید نمره امتیازی به شما تعلق خواهد گرفت. برای پیاده‌سازی کد بخش امتیازی امکان استفاده از نرم‌افزار متلب نیز وجود دارد.

لینک‌های مفید:

راهنمایی پیاده‌سازی SVM:

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

توضیحات

- یادگیری مفاهیمی که در تمرین مطرح شده و تدریس نشده‌اند با مطالعه شخصی ضروری است. برای این کار می‌توانید از لینک‌هایی که در تمرین معرفی شده است کمک بگیرید.
- برای حل و پیاده‌سازی سوالات (به جز قسمت امتیازی) تنها استفاده از زبان برنامه‌نویسی **پایتون** مجاز است. همچنین شما مجاز به استفاده از کتابخانه‌های آماده مانند numpy، matplotlib، pandas و sklearn می‌باشید.
- تحویل گزارش این تمرین ضروری است و به تمرین بدون گزارش نمره‌ای تعلق نمی‌گیرد. حجم گزارش معیاری برای ارزیابی نخواهد بود و لزومی به توضیح جزئیات کد نیست اما از آنجا که برای این تمرین از کتابخانه‌های موجود استفاده می‌شود، لطفا تمامی پارامترهای تنظیم‌شده در هر قسمت از کد را گزارش کرده و فرض‌هایی را که برای پیاده‌سازی‌ها و محاسبات خود به کار برده‌اید ذکر کنید.
- در فرایند ارزیابی گزارش، کدهای شما لزوما اجرا نخواهند شد. بنابراین همه‌ی نتایج و تحلیل‌های خود را به‌طور کامل ارائه کنید.
- شباهت بیش از حد گزارش و کدها باعث صفر شدن نمره تمرین خواهد شد. همچنین گزارش‌هایی که در آنها از کدهای آماده استفاده شده باشد پذیرفته نخواهند شد.
- گزارش شما باید به صورت تایپ شده و با فرمت pdf ارائه شود و کدهایی که به همراه گزارش تحویل می‌دهید باید قابل اجرا باشند. در انتها تمامی فایل‌های لازم را در یک فایل zip یا rar بارگذاری و ارسال کنید.
- پرسش‌های خود را از طریق ایمیل از دستیاران آموزشی مربوطه بپرسید:

fmojab@ut.ac.ir
saeed.dalir@yahoo.com