تمرین دوم هوش مصنوعی و یادگیری ماشین (رگرسیون)

فروردین ۱۴۰۲

هدف این تمرین آشنایی با مهارت‌های پایهٔ داده‌پردازی و کاربرد روش‌های مختلف رگرسیون در پیش‌بینی و دسته‌بندی دوگانه است.

بخش اول: پیش‌بینی نوع تومور سرطان سینه

در سال‌های اخیر استفاده از الگوریتم‌های یادگیری ماشین نقش مهمی در افزایش سرعت و دقت تشخیص نوع و شدت بیماری‌ها داشته‌اند. دادگان (dataset)ی که در این بخش مورد استفاده قرار می‌گیرد (فایل breast_cancer.csv) شامل اطلاعات ۶۸۴ بیمار با تشخیص سرطان سینه است. هر نمونه شامل ۹ ویژگی (مانند مشخصات تومور و کیفیت بافت اطراف آن) (ستون‌های ۱ تا ۹) و یک خروجی (نوع تومور) (ستون دهم) است که در این ستون عدد ۲ به معنای تومور خوش‌خیم و عدد ۴ به معنای تومور بدخیم است.

الف) ابتدا نشان دهید دادگانی که در اختیار شما قرار گرفته دارای نقصان، پرتی داده و عیوبی از این دست نیست، سپس عددهای نشان دهندهٔ نوع تومور را از ۲ و ۴ به ترتیب به ۰ و ۱ تغییر دهید. در پایان در صورت نیاز ویژگی‌های ۹ گانه را نرمال‌سازی کنید. (راهنمایی: برای نرمال‌سازی داده‌ها می‌توانید از روش های مختلفی مانند اسکیل کردن تمامی داده‌ها بین ۰ و ۱ استفاده کنید.)

ب) توزیع دادگان را برای تمام ۹ ویژگی بدست آورید. داده‌های با نوع تومور مختلف را با رنگ های متفاوت از یکدیگر نمایش دهید. پراکندگی داده‌ها را بر حسب ویژگی‌های مختلف نشان دهید. بر پایهٔ این پراکندگی‌ها برداشت خود را از ماهیت داده‌ها بیان کنید (راهنمایی: برای نمایش توزیع داده‌ها میتوانید از نمودارهای نقشه گرمایی، گسسته، هیستوگرام یا جعبه ای استفاده کنید.)

ج) رابطه و تاثیرگذاری هر کدام از پارامترها بر نوع تومور را پیدا کنید. (راهنمایی: می‌توانید از معیارهای آماری مانند کوریلیشن استفاده کنید تا میزان تاثیر داده‌ها بر روی خروجی و حتی رابطهٔ آن‌ها با یکدیگر را مشاهده کنید. برای نمایش هم می‌توانید از pair-plot در کتابخانه seaborn استفاده کنید.)

د) داده ها را به سه بخش آموزش (training)، ارزیابی (validation) و آزمایش (test) تقسیم کنید. پیشنهاد می‌شود ۸۰٪ کل داده‌ها به آموزش، ۱۰٪ به ارزیابی و ۱۰٪ به آزمایش اختصاص داده شود. در گام بعد با استفاده از الگوریتم رگرسیون لجستیک، مدلی را برای پیش‌بینی خروجی تربیت کنید و سپس با استفاده از روش k-fold cross validation (با k=5) بهترین عملکرد مدل را بدست آورید (این روش در ادامهٔ درس معرفی خواهد شد.). (توضیح بیشتر: در یادگیری ماشین هر مدل برای تنظیم پارامترهای خود به داده‌های آموزش نیاز دارد. داده‌های ارزیابی داده‌هایی هستند که برای آموزش مدل از آن‌ها استفاده نشده و به نوعی داده‌های جدیدی برای مدل به حساب می‌آیند. با استفاده از داده‌های ارزیابی عملکرد مدل بر روی داده از قبل دیده نشده، سنجیده می‌شود و هایپرپارامترهای مدل در جهت افزایش دقت بر روی داده‌های ارزیابی تنظیم می‌شود و سپس آموزش دوباره مدل از سر گرفته

می‌شود اما داده‌های آزمایش داده‌هایی هستند که فقط در انتها (پس از آموزش کامل مدل) توسط مدل دیده می‌شوند. این دسته از داده‌ها تنها برای ارزیابی عملکرد مدل در دنیای واقعی صورت می‌گیرد.)

ه) بر روی داده‌های آزمایش، ماتریس سردرگمی را تشکیل دهید (این ماتریس نیز در ادامهٔ درس معرفی خواهد شد) و نتایج را تحلیل کنید. میزان دقت بدست آمده را مناسب می‌دانید یا خیر؟ پیشنهادهای خود را برای افزایش دقت ارائه دهید.

و) **امتیازی**: بعد از آموزش، مواردی که به غلط توسط مدل پیش‌بینی شده را جدا کنید. با استفاده از تحلیل آماری یا تفسیر بصری علت این پیش‌بینی غلط را توضیح دهید و با ذکر دلیل، در جهت بهبود دقت مدل تلاش کنید.

بخش دوم: پیش‌بینی عمر مفید مواد دی‌الکتریک

مواد دی‌الکتریک به دلیل توانایی ذخیره‌ی بار الکتریکی (مانند عملکرد خازن‌ها) بصورت گسترده در صنعت مورد استفاده قرار می‌گیرند. از جمله بررسی‌هایی که در مورد این مواد صورت می‌گیرد، تعیین حداکثر ولتاژ قابل اعمال به آن‌ها در دمای کاری مشخص و برای عمر مفید مشخص است.
دادگانی که در این بخش مورد استفاده قرار می‌گیرد (فایل Performance-Degradation Data Nelson.xlsx) شامل نتایج ۱۲۸ آزمایش تعیین ولتاژ بیشینه است. هر نمونه شامل دو ویژگیِ عمر مفید (بر حسب هفته) و دمای کاری (بر حسب درجه سلسیوس) در ستون‌های اول و دوم و یک خروجیِ ولتاژ بیشینهٔ مجاز دی‌الکتریک (بر حسب کیلوولت) در ستون سوم است.

الف) به کمک نرم افزار پایتون و کتابخانه‌های مناسب، داده‌های فایل را وارد و به کمک دستور مناسب از کتابخانه‌ی sklearn، رگرسیون را با کرنل‌های خطی، RBF، چندجمله‌ای درجه ۲ و سیگموئیدی انجام داده و خطای مطلق میانگین (mean absolute error) و امتیاز R2 (R2-score) را در هر کدام به کمک روش k-fold cross validation (با k=4 یعنی در هر حالت ۲۵٪ از داده‌ها با انتخاب تصادفی پیش فرض sklearn برای آزمایش در نظر گرفته شود) به دست آورده و عملکرد چهار تابع کرنل (میانگین امتیاز به دست آمده برای داده‌های آزمایش) را مقایسه نمایید.

ب) خواسته‌ی مورد الف را به کمک L2-regularization با پارامتر alpha=1, 2 (دو مقدار) برای توابع کرنل بخش الف (چهار تابع) به دست آورده و نتایج را مقایسه نمایید. با توجه به تغییرات به وجود آمده در دقت رگرسیون درداده‌های آزمایش نسبت به بخش قبل، چه نتایجی می‌توان گرفت؟

ج) با تغییر پارامتر Regularization در مقادیر {0.2, 0.8, 1,5, 10, 20, 50, 300} به ازای توابع کرنل خطی، چندجمله‌ای درجه ۲ و ۳ و ۴ و هم چنین RBF مقادیر بهینه را به ازای هر کرنل و هم چنین بهترین کرنل را با بهترین امتیاز R2 به دست بیاورید (از دستور gridsearchcv استفاده نمایید). مقادیر نزدیک صفر برای امتیاز R2 به چه معنا هستند؟

د) در نرم افزار متلب به کمک دستور مناسب، فایل داده‌ها (با فرمت Excel) را وارد کرده و رگرسیون غیر خطی را با تابع زیر بر روی داده‌ها اعمال و امتیاز R2 و ریشه‌ی میانگین مربعات خطا را برای تابع برازش شده (به دست آوردن ضرایب مجهول b) به دست آورید.

$$\log(y) = b_1 - b_2.x_1.\exp(-b_3.x_2)$$

ه) به کمک دستور cftool در متلب و با استفاده از داده‌های آزمایش، تابع چندجمله‌ای را به ازای مقادیر مختلف درجات x1 و x2 بر داده‌ها برازش کرده و در حالتی که امتیاز R2 بیشینه می‌شود، مقادیر امتیاز R2 و RMS خطا را به همراه ضرایب چندجمله‌ای و رویه‌ی ایجاد شده ارائه نمایید.

چند توضیح:

- برای یادگیری مفاهیمی که در تمرین مطرح شده و احتمالا تدریس نشده‌اند از منابع موجود در اینترنت استفاده کنید.
- برای انجام بخش‌های مختلف تمرین می‌توانید از کتابخانه‌های آماده‌ای مانند numpy، matplotlib، pandas ، sklearn و seaborn استفاده کنید.
- 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 می‌کنید 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. از ارائهٔ توضیحات کلیشه‌ای و همانند بَرداری از منابع موجود بپرهیزید.
- در فرایند ارزیابی گزارش، کدهای شما لزوما اجرا نخواهد شد. بنابراین همهٔ نتایج و تحلیل‌های خود را به‌طور کامل ارائه کنید.
- شباهت بیش از حد گزارش و کدها باعث از دست دادن نمرهٔ 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.
- در صورت استفاده از گیت هاب جهت ارائهٔ گزارش و کد، نمره امتیازی به شخص تعلق میگیرد.
- پرسش‌های خود را از طریق ایمیل یا تلگرام از دستیاران آموزشی مربوطه بپرسید:

	ایمیل	تلگرام
محمد مهدی عابدی	mohammadabedi@ut.ac.ir	@mohammadabedi1179
پارسا شفیعی	parsa.shafiei@ut.ac.ir	@blind_side