به نام خدا

تمرین کامپیوتری دوم درس مقدمه‌ای بر یادگیری ماشین- گروه ۱

نیمسال دوم ۱۴۰۳-۱۴۰۲

موعد تحویل: ۳۱ اردیبهشت ۱۴۰۳

آداب انجام تمرینات کامپیوتری

- **همفکری** دانشجویان در انجام تمرینات مانعی ندارد ولی تمرینات باید به صورت مشخص توسط خود تحویل دهنده انجام گرفته باشد. در زمان تصحیح تمرینات شباهت تمرینات بررسی خواهد شد و تمرین‌های با شباهت غیرعادی مورد قبول نیست.
- استفاده از منابع اینترنتی برای گرفتن ایده ی حل سوالات مانعی ندارد ولی کپی کردن پاسخ تمرینات از هر منبعی مورد قبول نیست.
- استفاده از کتابخانه های آماده در صورتی که در صورت تمرین قید نشده باشد، مانعی ندارد.
- کدهای تحویل شده باید قابلیت اجرای دوباره را داشته باشند. در صورت وجود مشکل در زمان تصحیح (حتی به دلیل خطای تایپی)، قسمت‌های غیرقابل اجرا مورد قبول نیست. لطفا قبل از تحویل تمرین حتما کد خود را بررسی فرمایید.
- فایل ZIP شامل کدها و خروجی به همراه فایل گزارش (توضیحات در مورد آنچه انجام دادید و سوالات پرسیده شده) در CW آپلود گردد (در صورت استفاده از فرمت ipynb. می‌توانید توضیحات را در همان نوتبوک بنویسید و نیازی به گزارش نیست).
- سوالات و ابهامات در رابطه با این تمرین را می‌توانید در تلگرام با آیدی @khpew مطرح کنید.

تشخیص بیماری‌های قلبی عروقی

بیماری‌های قلبی و عروقی مهمترین عامل مرگ و میر ایرانیان است، امراضی که بر اساس آمارهای وزارت بهداشت و ثبت احوال، سالانه دست کم جان حدود ۱۴۰ هزار نفر را می‌گیرد و هزاران نفر را بستری، ناتوان و از کار افتاده می‌کند. بیماری هایی که بیش از هر عامل دیگری به سبک زندگی مانند تغذیه، تحرک و ورزش، استعمال دخانیات و آلودگی هوا بستگی دارند و با کنترل این عوامل تا حد زیادی قابل پیشگیری هستند. هدف ما در این تمرین این است که با استفاده از اطلاعات استاندارد سلامتی افراد و روش‌های مختلف طبقه‌بندی در یادگیری ماشین، وجود این دسته از بیماری‌ها در افراد را پیش بینی کنیم. مجموعه داده‌ای که در اختیار شما قرار گرفته است، شامل اطلاعات ۲۰۰۰ بیمار است. در این مجموعه داده ۱۸ ستون وجود دارد که از چپ به راست به صورت زیر می‌باشند:

```
Column Names: ['age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo', 'smoke', 'alco', 'active', 'cardio',
'cholesterol_1', 'cholesterol_2', 'cholesterol_3', 'gluc_1', 'gluc_2', 'gluc_3', 'bmi', 'bp']
```

متغیر `'cardio'` در واقع همان برچسب (label) ما می‌باشد که تعیین می‌کند فرد بیمار است (cardio = 1) یا خیر (cardio = 0). با توجه به این موارد به سوالات زیر پاسخ دهید:

۱. آماده‌سازی مجموعه داده

۱,۱) در ابتدا مجموعه داده را بخوانید و پنج سطر اول آن را نمایش دهید. همچنین به کمک دستور ()info. اطلاعات مختلف مربوط به مجموعه داده را نمایش دهید.

۱,۲) در این مجموعه داده، برخی از خانه‌ها خالی هستند (missing values)، سطرهای متناسب با آن‌ها را از این مجموعه داده حذف نمایید. سپس، بخش قبل را تکرار کنید.

۱,۳) داده‌ها را به گونه‌ای تقسیم‌بندی کنید که به صورت رندوم ۸۰ درصد آن برای train و ۲۰ درصد آن برای تست test باشد. (مقدار random_state را برابر 2024 قرار دهید.)

۲. طبقه‌بند Logistic Regression

۲,۱) با آموزش یک مدل Logistic Regression، داده‌ها را طبقه‌بندی کرده و صحت مدل را بدست آورید.

۲,۲) حال می‌خواهیم ویژگی‌های مهم‌تر را شناسایی کنیم. ضرایب مدل تعریف شده در قسمت قبل را محاسبه کرده و رابطه‌ی بین هر ویژگی با خروجی را توضیح دهید. کدام ویژگی‌ها مهم‌ترند و چرا؟

۳. طبقه‌بند Support Vector Machine

۳,۱) این بار می‌خواهیم مجموعه داده‌ها را به کمک SVM طبقه‌بندی کنیم. در ابتدا با استفاده از کرنل خطی(linear) مدل خود را تعریف کنید و به ازای مقادیر مختلف C، صحت هرکدام را محاسبه کنید.

۳,۲) حال این‌بار از کرنل poly استفاده کرده و به ازای مقادیر مختلف پارامترهای C و degree، صحت طبقه‌بندها را محاسبه کنید.

۳,۳) در آخر نیز از کرنل rbf استفاده کرده و به ازای مقادیر مختلف پارامترهای C و gamma ، صحت طبقه‌بندها را محاسبه کنید.

۳,۴) صحت‌های محاسبه شده در بخش‌های قبل را باهم مقایسه کنید. کدام کرنل با چه پارامترهایی بهترین عملکرد را داشته است؟ دلیل آن چیست؟

٤. طبقه‌بند Perceptron

۴,۱) حال این بار به کمک الگوریتم perceptron، داده‌ها را طبقه بندی کنید و صحت آن را بدست آورید. از الگوریتم sample mode استفاده می‌کنید یا batch mode و چرا؟ (در این بخش مجاز به استفاده از تابع آماده Perceptron نیستید.)

۴,۲) این بار به کمک تابع آماده Perceptron طبقه‌بندی را انجام دهید و صحت آن را با بخش قبل مقایسه کنید.

٥. شبکه‌های عصبی MLP

۵,۱) حال در این بخش می‌خواهیم به کمک شبکه‌های عصبی(MLP) داده‌ها را طبقه‌بندی کنیم. در رابطه با پارامترهای مختلف MLPClassifier (تعداد لایه‌ها و نورون‌های پنهان، توابع فعالسازی، solver و ...) تحقیق کنید و با استفاده از GridSearchCV، شبکه عصبی خود را تعریف کنید و به کمک آن داده‌ها را طبقه‌بندی کرده و صحت آن را گزارش کنید.

٦. درخت تصمیم

۶,۱) حال در این بخش می‌خواهیم به کمک درخت تصمیم، داده‌ها را طبقه‌بندی کنیم. ابتدا بهترین پارامترهای درخت را بیابید (مانند عمق درخت، معیار ناخالصی و ...)، سپس یک مدل DecisionTreeClassifier با random_state = 2024 تعریف کنید و داده‌ها را به کمک آن و پارامترهایی که بدست آوردید، طبقه‌بندی کرده و صحت آن را گزارش کنید. (راهنمایی: می‌توانید از GridSearchCV استفاده کنید.)

۷. طبقه‌بندهای بالا را باهم مقایسه کنید. کدام طبقه‌بندها عملکرد بهتری دارند؟ علاوه بر صحت، از چه معیارهای دیگری می‌توان برای مقایسه عملکرد طبقه‌بندها استفاده کرد؟