به نام خدا

تکلیف چهارم یادگیری ماشین

نیمسال تحصیلی ۰۱-۰۰

موعد تحویل: ۱۵ دی ساعت ۲۳:۵۹

- در این تمرین میتوانید از توابع آماده موجود در کتابخانه هایی نظیر scikit-learn استفاده نمایید.

۱. مجموعه داده‌های آموزشی که در فایل Q1.csv قرار دارد شامل نمرات دانشجویان یک کلاس می‌باشد. ستون اول نمرات امتحان شفاهی و ستون دوم نمرات امتحان کتبی درس از ۱۰۰ نمره می‌باشد. ستون سوم نیز نشان‌دهنده قبولی یا عدم قبولی شخص در آن درس است. هدف از این سوال طراحی یک طبقه‌بند رگرسیون لجستیک به منظور پیش بینی احتمال قبولی بر اساس نمرات امتحان هر شخص است.

الف) ابتدا ۷۰ درصد داده‌ها را به منظور آموزش طبقه‌بند و ۳۰ درصد باقی مانده را برای تست به صورت تصادفی، تقسیم کرده و داده‌های آموزشی را توسط دو رنگ متفاوت نمایش دهید. (۰/۵ نمره)
ب) یک طبقه‌بند رگرسیون لجستیک را بر روی داده‌های آموزشی آموزش داده و سپس دقت آن را هم بر روی داده‌های آموزشی و هم بر روی داده‌های تست به دست آورید. (۲ نمره)
ج) مرز تصمیم این طبقه‌بند و داده‌های تست را رسم کنید. (۱/۵ نمره)

۲. مجموعه داده‌های آموزشی که در فایل Q2.csv قرار دارد، نتایج دو آزمون بر روی میکروچیپ‌های تولیدی یک شرکت است. ستون سوم نشان‌دهنده پذیرش یا رد شدن هر چیپ می‌باشد. در این سوال یک طبقه‌بند رگرسیون لجستیک منتظم (Regularized Logistic Regression) برای طبقه‌بندی این نتایج این آزمون‌ها طراحی خواهد شد.

الف) ابتدا ۷۰ درصد داده‌ها را به منظور آموزش طبقه‌بند و ۳۰ درصد باقی مانده را برای تست به صورت تصادفی، تقسیم کرده و داده‌های آموزشی را توسط دو رنگ متفاوت نمایش دهید. (۰/۵ نمره)

ب) در این سوال با توجه به عدم امکان تفکیک خطی، نیاز است که داده‌های ورودی به یک فضای ویژگی پیچیده تر تصویر شده و سپس آموزش طبقه‌بند صورت گیرد. با بردن داده‌ها به یک فضای ویژگی پیچیده تر شامل تمام چند جمله ای های تا مرتبه ۴، طبقه‌بند را آموزش دهید. (۲/۵ نمره)
ج) برای جلوگیری از بیش‌برازش بر روی داده‌های آموزشی از پارامتر منتظم‌سازی استفاده می‌شود. با تغییر دادن این این پارامتر و آموزش مجدد دقت را هم بر روی داده‌های آموزشی و دادهای تست با هم مقایسه کنید. (۳ نمره)

۳. در این سوال با استفاده از ماشین بردار پشتیبان، یک طبقه‌بند برای دسته‌بندی دو کلاس طراحی خواهید کرد.

الف) با استفاده از داده‌های آموزشی موجود در فایل Q3_1.csv یک طبقه‌بند ماشین بردار پشتیبان آموزش دهید همچنین داده‌ها و مرز تصمیم را ترسیم کنید. با تغییر دادن پارامتر C در این طبقه‌بند تغییرات در مرز تصمیم را ترسیم و علت را توضیح دهید. (۲ نمره)
ب) داده‌های قسمت الف به صورت خطی تفکیک پذیر بود. به منظور تفکیک داده‌هایی که به صورت خطی تفکیک پذیر نیستند یک روش استفاده از کرنل‌های گوسی است. در این قسمت داده‌های موجود در فایل Q3_2.csv را توسط ماشین بردار پشتیبان و کرنل‌های گوسی طبقه‌بندی کرده و داده‌ها را به همراه مرز تصمیم ترسیم نمایید. (۳ نمره)
ج) داده‌های آموزشی موجود در فایل Q3_3.csv را ابتدا به دو مجموعه آموزش و تست با نسبت ۷۰ به ۳۰ تقسیم کرده سپس با استفاده از آموزش یک طبقه‌بند ماشین بردار پشتیبان و تغییر پارامترهای C و σ بر روی تمامی ترکیب‌های (0.01,0.03,0.1,0.3,1,3,10,30) بهترین مرز تصمیم برای داده‌های تست را یافته و آن را به همراه داده‌های تست رسم نمایید. (۲/۵ نمره)

۴. داده‌های آموزشی موجود در فایل Q4.csv دیتاست مشهور Iris است که ویژگی های اندازه‌گیری شده از سه نوع مختلف گل زنبق را نشان می‌دهد.
الف) ابتدا ۷۰ درصد داده‌ها را به منظور آموزش طبقه‌بند و ۳۰ درصد باقی مانده را برای تست به صورت تصادفی، تقسیم نمایید. (۰/۵ نمره)
ب) یک طبقه‌بند K – نزدیکترین همسایگی به منظور منظور تشخیص نوع گل زنبق آموزش دهید. (۱ نمره)
ج) میزان دقت بر روی داده‌های تست به ازای مقادیر مختلف K محاسبه کرده و بر روی یک نمودار نمایش دهید. با این کار مقدار بهینه K بر روی داده‌های تست محاسبه می‌شود. (۱ نمره)