دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

درس یادگیری ماشین

تکلیف کامپیوتری سوم

تاریخ تحویل: ۹ دی

سوال ۱

در این تمرین قصد داریم بر روی مجموعه داده Iris مدل Logistic Regression را پیاده سازی کنیم. این مجموعه داده دارای چهار ویژگی (طول و عرض کاسبرگ و گلبرگ) برای ۱۵۰ نمونه است که شامل ۳ دسته مختلف هستند. در جدول زیر چند رکورد از این مجموعه داده قابل مشاهده است.

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2

داده‌ها را می‌توانید از کتابخانه scikit-learn بخوانید.

ابتدا داده ها را نرمال کرده و به سه دسته آموزش [۱] و آزمایش [۲] و اعتبارسنجی [۳] تقسیم کنید(۸۰ درصد کل داده‌ها برای آموزش، ۱۰ درصد برای آزمایش و ۱۰ درصد برای اعتبارسنجی).

الف) مدل Logistic Regression را از پایه پیاده سازی کنید. (۲۰ نمره)

ب) نمودار خطا بر روی کل داده‌ها در هر گام بروز رسانی وزن‌ها برای حالت GD و یا Mini-batch GD رسم کنید. (۵ نمره)

ج) با استفاده از داده اعتبارسنجی، هایپرپارامتر نرخ یادگیری [۴] را تنظیم [۵] کنید. (۵ نمره)

د) اختیاری: با استفاده از کتابخانه scikit-learn و استفاده از Logistic Regression مدل را آموزش داده و دقت [۶] را بر روی مجموعه داده آزمایش، گزارش کنید. (۳ نمره)

سوال ۲

در این سوال با استفاده از توابع موجود در کتابخانه sklearn ماشین بردار پشتیبان یک طبقه بند دو کلاسه طراحی خواهید کرد.

الف) با استفاده از داده های آموزشی موجود در فایل sin.csv یک طبقه بند ماشین بردار پشتیبان طراحی کرده و مرز تصمیم را همراه با داده های آموزشی رسم کنید. (۱۰ نمره)

ب) با تغییر دادن پارامتر C در این طبقه بند مرزهای تصمیم حاصل شده را رسم کنید و تغییرات را توجیه کنید. (۵ نمره)

[1] Train data
[2] Test data
[3] Validation data
[4] Learning Rate
[5] Tune
[6] Accuracy

ج) اختیاری: داده های موجود در قسمت الف به صورت خطی تفکیک پذیر بودند. داده های موجود در فایل sin_p.csv به صورت خطی تفکیک پذیر نیستند. یک ماشین بردار پشتیبان برای طبقه بندی داده های موجود در این فایل با استفاده از کرنل های گوسی طراحی کنید. برای این کار ۹۰ درصد داده ها را برای آموزش و بقیه را برای تست در نظر بگیرید. در انتها مرز تصمیم را همراه داده های تست رسم کنید. (۴ نمره)

د) اختیاری: با آزمون وخطا طبقه بند هایی با ترکیب هایی از پارامترهای C و γ را آموزش داده و مرزهای تصمیم ترکیب های مختلف را رسم کرده و مشاهدات خود را توجیه کنید. (۳ نمره)

سؤال ۳

در این سوال می‌خواهیم بر روی مجموعه داده Iris که در سوال ۱ با آن آشنا شدیم، الگوریتم KNN را پیاده‌سازی کنیم.

الف) تابعی بنویسید که با دریافت داده‌ها و نیز مقدار K پس از نرمال‌سازی داده‌ها (برای نرمال‌سازی میتوانید از توابع آماده استفاده کنید) طبقه‌بندی را به روش KNN انجام دهد. (برای این کار نمی‌توانید از تابع آماده KNeighborsClassifier در پایتون استفاده کنید.) هم‌چنین برای جداسازی داده آموزشی از داده آزمایشی از روش Leave One Out که هربار یکی از رکوردهای داده را برای تست و مابقی را برای آموزش استفاده می‌کند، بهره بگیرید. (می‌توانید از تابع LeaveOneOut از کتابخانه sklearn استفاده کنید.) (۲۰ نمره)

ب) تابعی که در قسمت قبلی پیاده‌سازی کردید را به ازای مقادیر K بین ۱ تا ۵۰ اجرا کرده، در هر مرحله نرخ خطای کلاس‌بندی را رسم کنید؛ نتایج نمودار را تحلیل کنید. کمترین نرخ خطا مربوط به چه مقدار K است و میزان آن چقدر است؟ (۵ نمره)

ج) اختیاری: این بار برای پیاده‌سازی KNN از روش weighted voting استفاده کنید و مراحل الف و ب را اجرا کرده، آن را با روش unweighted مقایسه کرده و نتایج را تحلیل کنید. (۵ نمره)

سؤال ۴

در این سوال می‌خواهیم بر روی مجموعه داده Iris که در سوال ۱ با آن آشنا شدیم، به کمک درخت تصمیم طبقه‌بندی انجام دهیم. در حل این سوال شما هیچ محدودیتی در استفاده از کتابخانه‌های آماده پایتون ندارید.

الف) ابتدا داده‌ها را بخوانید و آن را به دو قسمت آموزش و آزمایش تقسیم کنید. (۳۰ درصد داده‌ها را برای آزمایش در نظر بگیرید.) (۲ نمره)

ب) در این قسمت قصد داریم تا به کمک تابع DecisionTreeClassifier از کتابخانه sklearn پایتون، یک طبقه‌بند [۷] بسازیم. می‌خواهیم برای پیدا کردن پارامترهای بهینه برای این طبقه‌بند از روش grid search استفاده کنیم. بدین منظور میتوانید از GridSearchCV از کتابخانه sklearn استفاده کنید. پارامترهای مورد بررسی برای درخت تصمیم در این سوال به شرح زیر هستند:

- criterion که می‌تواند entropy یا gini باشد
- max_depth

[7] Classifier

- min_samples_split
- max_leaf_nodes

به کمک GridSearchCV مقدار بهینه برای هر یک از این پارامترها را برای درخت تصمیم بیابید و در خروجی چاپ کنید. (۱۸ نمره)

ج) بهترین مدلی که در قسمت قبل به دست آوردید را بر روی داده آموزشی آموزش داده و سپس آن را بر روی داده آزمایش تست کنید. دقت پیش‌بینی مدل روی داده آزمایشی را نمایش دهید. (۱۰ نمره)

د) اختیاری: ماتریس Confusion مربوط به این طبقه‌بند را رسم کرده و نتایج آن را تفسیر کنید. (می‌توانید از confusion_matrix از کتابخانه sklearn استفاده کنید.) (۵ نمره)

نکات تکمیلی

۱. برای انجام این تکلیف استفاده از زبان پایتون الزامی است.

۲. تکالیف را در محیط jupyter notebook پیاده‌سازی کنید و فایل ipynb. را ارسال کنید.

۳. توضیح کدی که نوشته‌اید، بررسی و تحلیل نتایج آن و بیان علت نتایج و نیز مقایسه نتیجه با آنچه مورد انتظارتان بوده است، از اهمیت بالایی برخوردار است. شما می‌توانید گزارش پروژه را در همان محیط jupyter notebook بنویسید و نیازی به فایل pdf جداگانه نیست. هم‌چنین اگر برای حل سوال فرضیات خاصی مدنظر دارید حتما آن را در متن گزارش قید کنید.

۴. فرمت نامگذاری تکلیف ارسالی باید به صورت زیر باشد: HWX_Programming_LastName_StudentID که X شماره تکلیف LastName نام خانوادگی شما و StudentID شماره دانشجویی شما است.

۵. انجام این تکلیف به صورت تک نفره است. در صورت مشاهده تقلب، نمرات هم مبدا کپی و هم مقصد آن صفر لحاظ می‌شود.

۶. شما می‌توانید تا یک هفته پس از پایان مهلت تکلیف آن را در یکتا بارگذاری کنید. در این صورت به ازای هر روز تاخیر ۷ درصد از نمره تکلیف کسر می‌شود. پس از اتمام این یک هفته امکان ارسال با تاخیر وجود ندارد.

۷. در صورت وجود هر گونه ابهام و یا سوال می‌توانید سوالات خود را در گروه سروش بپرسید. هم چنین می‌توانید برای رفع ابهامات با دستیاران آموزشی از طریق تلگرام و یا اسکایپ در تماس باشید.

آیدی‌ها:

@Fatemeh2114P

@amir7d0

@mastaraan

live:.cid.7f0be16d612107cc

و یا سؤال خود را با موضوع "تکلیف درس مبانی یادگیری ماشین" به ایمیل زیر ارسال کنید:

arsh.2001.1379@gmail.com

موفق باشید.