تمرین سری اول

درس: یادگیری ماشین- پاییز ۱۴۰۳

استاد درس: دکتر فاطمه زمانی

دستیار آموزشی: حسین آقاگل زاده

دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی نوشیروانی بابل

- این تمرین دارای سه بخش **تئوری**، **پیاده سازی** و **مطالعاتی** است.
- پاسخنامه بخش تئوری می تواند به صورت دست نویس یا تایپی باشد. در هر صورت اما می بایست در قالب **یک فایل pdf** تحویل گردد.
- پاسخ بخش پیاده سازی می بایست حاوی **کد پیاده سازی** شده **به همراه گزارش آن در قالب فایل pdf** باشد. از آوردن کد در گزارش اجتناب کنید (مگر آوردن بخشی از آن از نظر شما ضرورت داشته باشد).
- پاسخ بخش مطالعاتی می بایست به صورت تایپ شده در قالب یک فایل pdf باشد.
- در تصحیح تمارین ابتدایی، نسبت به گزارش پیاده سازی سخت گیری کمتری در نظر گرفته می شود.
- کل پاسخ (کد و فایل های pdf) در **یک فایل zip** با فرمت زیر ارسال شود.

HW1_شماره دانشجویی.zip

تئوری

مسئله ۱

برای دو مسئله زیر مشخص کنید که اگر بخواهیم با رویکرد یادگیری ماشین آن ها را حل کنیم جزو کدام یک از مسائل supervised learning، unsupervised learning و یا reinforcement learning قرار میگرند. چرا؟

الف) مجموعه ای از داده جدولی با ۳ ستون در اختیار داریم: متراژ خانه، تعداد اتاق ها و قیمت خانه. می خواهیم بر اساس متراژ خانه و تعداد اتاق ها برای یک نمونه خارج از این جدول قیمت خانه را حدس بزنیم.

ب) بازی شطرنج (بیشتر شرح دهید)

مسئله ۲

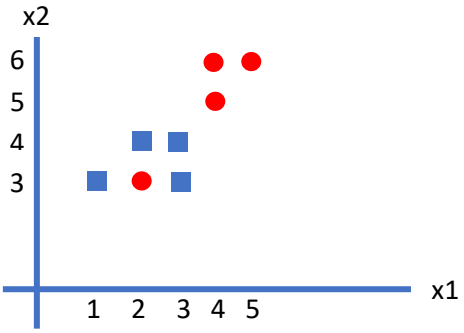
حسن و سعید دو دانشجوی درس ریاضیات هستند که می خواهند بهترین عملکرد را در آزمون پایان ترم داشته باشند. این دو فرد دو رویکرد متفاوت را برای آماده شدن برای امتحان انتخاب کردند. حسن تمام مثال ها و تمرین ها را حفظ می کند طوری که اگر دقیقا همان مثال در آزمون پایانی بیاید جواب کامل و دقیق آن را خواهد داد. سعید اما سعی می کند مفاهیم درسی رو به خوبی درک کند و تا قبل از امتحان نیز با بخشی از تمارین و مثال های درسی باقی مانده خود را آزمایش کند. این مسئله به چه بحثی در یادگیری ماشین شباهت مفهومی دارد؟ پیش بینی شما از عملکرد دو دانشجو چیست؟ چرا؟

مسئله ۳

یک مسئله Classification دو کلاسه را که نمونه های آموزشی برچسب دار آن دارای دو ویژگی x1 و x2 هستند. پراکندکی این نمونه ها در شکل زیر قابل مشاهده است (مقادیر x1 و x2 این نمونه ها را رند در نظر بگیرید). فرض کنید یک نمونه آزمون در موقعیت (x1=2, x2=3.1) قرار گیرد. می خواهیم با روش KNN با این مسئله رو به رو شویم.

الف) بدون انجام محاسبات به نظر شما چه مقداری برای k مناسب است؟ (ممکن است چند جواب صحیح وجود داشته باشد آوردن و توضیح یکی از آن ها کافی است)

ب) با نشان دادن محاسبات مرحله آزمون را با در نظر گرفتن k=1 و k=3 برای نمونه آزمون انجام دهید و کلاس این نمونه را پیش بینی کنید.



پیاده سازی

* کتابخانه های پایتون پیشنهادی حل این بخش:

numpy

matplotlib

scikit learn

Google Colab یک بستر برنامه نویسی رایگان (در حالت پایه) برای زبان پایتون است که علاوه بر محیط برنامه نویسی، سخت افزار آن را نیز در اختیار می گذارد. پیشنهاد می شود که برای انجام بخش پیاده سازی تمارین از این محیط بهره ببرید.

colab.research.google.com

الف) مطالعه مختصری در مورد دیتاست MNIST و اهمیت آن انجام دهید و در یک بند گزارش کنید.

ب) دیتاست MNIST را به برنامه اضافه کنید و یک نمونه آن را به تصویر بکشید. برای اضافه کردن این دیتاست می توان از راهکار آسان زیر استفاده کنید.

https://keras.io/api/datasets/mnist/

پ) هر نمونه از این دیتاست شامل یک ماتریس ۲۸ در ۲۸ (معادل یک تصویر ۲۸ در ۲۸) است. همه ی نمونه های درون این دیتاست را به بردار ۷۸۴ تایی تبدیل کنید. برای مثال برای بخش training:

$$(60000,28,28) \rightarrow (60000,784)$$

$$28 * 28 = 784$$

سمت چپ shape ورودی و سمت راست shape هدف است.

راهنمایی: استفاده از دستور reshape در کتابخانه numpy.

ت) دیتاست را به سه بخش train، test و validation تقسیم کنید. اگر بخش test در ابتدا خود جدا است فقط بخش train را به دو بخش validation (۱۰ درصد) و train جدید تقسیم کنید. کاربرد هر یک از این بخش ها را در روند حل یک مسئله یادگیری ماشین بیان کنید.

ث) مطابق با انچه از الگوریتم KNN فرار گرفته اید پیاده سازی ای دستی از آن انجام دهید (مجاز به استفاده از دستورات آماده ای که در یک یا چند خط کوتاه این الگوریتم را به برنامه اضافه می کنند نیستید).

ج) با در اختیار داشتن سه بخش train، test و validation چه راهکاری را برای یافتن k مناسب پیشنهاد می دهید؟ آن راهکار را پیاده کنید و مقدار k مناسب را بدست بیاورید (اگر از روش اعتبارسنجی چند بخشی استفاده می کنید انجام یک مرحله از آن کافی است). در این بخش معیار فاصله را اقلیدسی در نظر بگیرید.

چ) پس از یافتن k مناسب بخش validation و train را ادغام کنید و عملیات آموزش را بر روی این بخش ادغام شده انجام دهید. دقت شود بسته به کد پیاده سازی شده عملیات آموزش می تواند شامل هیچ کدی نباشد! مرحله آزمون (test) را بر روی بخش test انجام دهید و accuracy را گزارش کنید. یک بار با معیار فاصله اقلیدسی و یک بار Manhattan.

$$accuracy = \frac{\text{تعداد نمونه های آزمونی که برچسب آن ها درست پیش بینی شده است}}{\text{تعداد کل نمونه های آزمون}}$$

ح) بخش **چ** را این بار با دستورات آماده کتابخانه scikit learn انجام دهید (فقط با معیار فاصله اقلیدسی).

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

مطالعاتی

مقاله ی پیوست شده را مطالعه کنید و برداشت خود از آن را در حدود دو صفحه شرح دهید.

- از آوردن ترجمه به صورت مستقیم خودداری کنید.
- شرح مورد نظر شامل شبه کد به زبان فارسی و گویا از روش پیشنهادی مقاله باشد.

😊 موفق باشید