دانشگاه صنعتی اصفهان

ترم دوم سال تحصیلی 1403_1404

مبانی یادگیری ماشین

تکلیف عملی اول

Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data.

نکات تکمیلی

1. پاسخ ها در یکتا بارگذاری شوند و ارسال از روش های دیگر مورد قبول نیست.
2. تحویل تکلیف با تاخیر تا 7 روز امکان پذیر است و به ازای هر روز،7 درصد از نمره آن تکلیف کسر میشود که به این نکته **حتما** توجه شود.
3. ساختار نامگذاری تکلیف ارسالی باید به این صورت باشد.HWX_ Programming_LastName_StudentID

   X شماره تکلیف ، studentID شماره دانشجویی وLastName نام خانوادگی شماست.
4. انجام تکالیف به صورت تک نفره هست و در صورت مشاهده تقلب نمرات هم مبدا و هم مقصد آن صفر خواهد شد.
5. براي انجام اين تكليف استفاده از زبان پايتون الزامي است و استفاده از توابعی جز pandas، numpy و matplotlib درصورتی که در سوال ذکر شود، مجاز نمیباشد.
6. تکالیف را در محیط jupyter notebook یا google colab پیاده سازی کنید و فایل ipynb. را ارسال کنید.
7. توضیح کدی که نوشته اید، بررسي و تحلیل نتایج آن و بیان علت نتایج و نیز مقایسه نتیجه با آنچه مورد انتظارتان بوده است، از اهمیت بالایی برخوردار است. شما مي توانيد گزارش پروژه را در همان محیط jupyter notebook بنویسید و نیازی به فایل pdf جداگانه نیست. هم چنين اگر براي حل سوال فرضيات خاصي مدنظر داريد حتما آن را در متن گزارش قید کنید.
8. در صورت پاسخگویی به هر 4 سوال بخش دوم، کمبود نمره ناشی از پاسخ اشتباه در دیگر سوالات تمرین جبران خواهد شد. (توجه کنید که نمره نهایی تکلیف، بیشتر از 100 نخواهد شد.)
9. در صورت هرگونه ابهامی می توانید سوالات خود را در گروه تلگرام بپرسید. و همچنین میتوانید با دستیاران آموزشی از طریق تلگرام در ارتباط باشید.

آیدی تلگرام:

@arash_azhand

تکلیف عملی اول

سوال اول - آشنایی با کتابخانه ها (40 نمره)

برای حل این سوال یک فایل ژوپیتر در اختیار شما قرار گرفته که باید آن را کاملش کنید. هدف این سوال آشنایی با کتابخانه های مهم و مورد نیاز برای پروژه های یادگیری ماشین است و با انجام کامل این نوت‌بوک شما به یک تسلط نسبی برای پروژه های واقعی تر خواهید رسید. مکان هایی که شما باید برای آن کد بنویسید با # your code here مشخص شده اند. با توجه به توضیحات و خواسته ی سوال کد خواسته شده را بنویسید و خروجی بگیرید. در این سوال مجموعه داده را خودمان میسازیم و با آن کار میکنیم.

سوال دوم – برازش خطی[1] (80 نمره که 20 نمره آن اختیاری است)

در این تمرین قصد داریم با توجه به داده های موجود در فایل Student_Performance.csv یک برازش خطی به روش نزول گرادیان[2] انجام دهیم که بتوانیم نمره‌ی دانشجوها را با توجه به اطلاعاتی که از آنها داریم تخمین بزنیم. این فایل شامل اطلاعات زیر میباشد که برای فهم بهتر مسئله مناسب است این اطلاعات مطالعه شود.

Hours Studied: تعداد ساعاتی که هر دانشجو برای امتحان درس خوانده است.

Previous Scores: دانشجو در امتحان قبلی چه نمره ای کسب کرده است.

Extracurricular Activities: آیا دانشجو فعالیت فوق برنامه داشته است یا خیر.

Sleep Hours: تعداد ساعات خواب دانشجو قبل از امتحان.

Sample Question Papers Practiced: تعداد نمونه سوالاتی که دانشجو برای این امتحان حل کرده است.

Performance Index: و در نهایت نمره ای که دانشجو در این امتحان کسب کرده است.

[1] Linear Regression
[2] Gradient Descent

تکلیف عملی اول

1. یکی از مهمترین گام های پروژه های یادگیری ماشین پیش پردازش داده ها و تحلیل آماری مجموعه داده یا EDA[3] است که هم به فهم بهتر داده ها کمک میکند و هم کار ما را در آینده برای انتخاب مدل های مختلف و تغییرات احتمالی راحت تر میکند. (20)
   1.1. مجموعه داده را بارگذاری کنید سپس داده هایی که بعضی از مقادیر آنها داخل جدول نیستند را حذف کنید. (3)
   1.2. سعی کنید بجای حذف، آنها را با دو روش مختلف جایگذاری کنید. (5)
   1.3. با نمودار های مختلف مثل histogram، scatter plot و heat map داده ها را به نمایش در بیاورید. (5)
   1.4. ستون هایی که شامل رشته هستند را باید به عدد تبدیل کنید. (3)
   1.5. داده های عددی را نرمال[4] کنید. (2)
   1.6. ممکن است رکورد هایی در مجموعه داده تکرار شده باشند. آنها را در صورت وجود پیدا کنید و رکوردهای تکراری را حذف کنید. (2)

2.
   2.1. هر دو روش GD و Mini-batch sgd را بر روی تابع خطای میانگین مربعات[5] پیاده سازی نموده و برای هر کدام از این روشها، نمودار تغییرات خطا بر روی کل داده ها را در هر گام بر وزرسانی وزنها رسم نمایید. علت اعوجاج های مشاهده شده در هر نمودار را توضیح دهید. (برای رسم نمودار توجه شود که محور عمودی نشان دهنده میزان خطا بر روی کل داده ها بوده و محور افقی نشان دهنده گام های به روزرسانی وزنها است. پارامتر های اولیه را صفر درنظر بگیرید. نرخ آموزش بهینه را هم با روش های مختلف مثل grid search بدست آورید.) (15)
   2.2. دو روش پیاده سازی شده در قسمت قبل را از نظر سرعت همگرایی و کمینه خطا با یکدیگر مقایسه کنید. (5)

3.
   3.1. برای روش GD ، برازش را بار دیگر با مقداردهی تصادفی انجام داده و نتایج را با مقداردهی اولیه صفر مقایسه کنید. (2)
   3.2. یکبار دیگر عمل برازش را با در نظر گرفتن تابع خطای MAE[6] و به روش SGD پیاده سازی نموده و با تابع خطای MSE مقایسه نمایید. در طول به روزرسانی وزنها چند بار به نقاط مشتق ناپذیر برخورد کردید؟ اگر برخورد کردید راه حل شما چه بود؟ (15)
   3.3. نشان دهید اگر نرخ یادگیری بزرگ باشد مدل همگرا نمیشود. (با توجه به خطای مدل در زمان آموزش. ترجیحا نمودار آن هم رسم شود) (3)

[3] Exploratory Data Analysis
[4] Normalization
[5] Mean Square Error
[6] Mean Absolute Error

تکلیف عملی اول

4. به مدل چند ویژگی کوادراتیک اضافه کنید و بعد با استفاده از رگیولاریزیشن ریج و نزول گرادیان مسأله را حل کنید و مقدار بهینه لامبدا را با استفاده از داده های اعتبارسنجی و جستجو بدست آورید. (20)