به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده‌ی مهندسی کامپیوتر

تمرین سری اول یادگیری ماشین

دکتر احسان ناظرفرد

طراح سوال:

سید اردلان قریشی

محمدرضا امامی ناصری

پاییز ۱۳۹۹

توضیحات مهم:

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان #StudentId_HW1.zip بارگذاری نمایید (به عنوان مثال 99131000_HW1.zip).
- مهلت انجام تمرین تا **ساعت ۲۳:۵۵ روز شنبه مورخ ۱۵ آذر** می‌باشد و به هیچ وجه تمدید نمی‌شود.
- تمرین بدون گزارش فاقد ارزش می‌باشد و **نمره‌ای به آن تعلق نمی‌یابد**.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید (به بهترین گزارش نمره تشویقی تعلق می‌گیرد).
- مطابق قوانین دانشگاه هرگونه کپی برداری **ممنوع** می‌باشد و در صورت مشاهده نمره‌ی **هر دو طرف** صفر در نظر گرفته می‌شود.
- شما مجاز هستید برای تمامی تمرین‌ها **۷ روز در کل** و **با سقف حداکثر ۳ روز برای هر تمرین**، تاخیر بدون کسر نمره داشته باشید. به ازای هر روز تاخیر بیشتر، ۱۰٪ از نمره‌ی تمرین مربوطه کسر می‌شود.
- در صورت داشتن هرگونه ابهام می‌توانید از طریق ایمیل زیر سوال خود را مطرح نمایید:

MLAUTFALL99@gmail.com

سوالات تشریحی

۱- مفاهیم زیر را تعریف کنید ٪۲۰

الف) یادگیری نظارتی (Supervised Learning)

ب) یادگیری نیمه نظارتی (Semi-Supervised Learning)

پ) یادگیری بدون نظارت (Unsupervised Learning)

ت) یادگیری تقویتی (Reinforcement Learning)

ث) یادگیری عمیق (Deep Learning)

ج) رگرسیون (Regression)

چ) یادگیری برخط (Online Learning)

ح) یادگیری فعال (Active Learning)

خ) دسته بندی (Classification)

د) خوشه بندی (Clustering)

ذ) بیش برازش و کم برازش (Overfitting & Underfitting)

۲- همبستگی[1] بین ویژگی‌ها به چه معنی است و چگونه می‌توان آن را تشخیص داد؟ (کامل توضیح دهید) ٪۵

۳- معیارهای ارزیابی MSE، MAE و RMSE را با هم مقایسه کرده و بگویید در صورت داشتن داده‌های پرت و نویزی کدام یک بهتر عمل می‌کند؟ چرا؟ ٪۵

۴- روش‌های گرادیان نزولی و معادله نرمال را با یکدیگر مقایسه کرده و برتری هر کدام را شرح دهید. ٪۵

۵- رگرسیون Lasso را توضیح داده و تفاوت آن را با رگرسیون خطی شرح دهید. ٪۵

[1] correlation

سوالات پیاده‌سازی

توضیحات مهم:

- در روند اجرا انتخاب مقادیر برای تقسیم داده‌ها به مجموعه آموزش، ارزیابی و... به عهده دانشجو می‌باشد.
- حتما پارامترهای انتخاب شده برای برنامه خود و هرگونه شرایطی که درنظر گرفته‌اید را در گزارش خود بیاورید.
- برای بهبود سرعت برنامه توصیه می‌شود از عملیات ماتریسی استفاده کنید.
- در هر مرحله، نتایج خود را تحلیل کنید.
- کدهای خود را برای خوانایی بیشتر **کامنت گذاری** کنید.
- در تمامی سوال‌ها تنها مجاز به استفاده از کتابخانه‌هایnumpy، matplotlib و pandas می‌باشید.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال ذکر شده است.**
- گذاشتن عنوان برای نمودارها و برچسب گذاری محورهای نمودار الزامی می‌باشد.

بخش اول

مجموعه داده‌ی Dataset1.csv را بارگذاری کنید.

۱- داده‌ها را رسم کنید.

۲- شافل[۲] کردن و نرمال‌سازی[۳] داده‌ها به چه منظور انجام می‌شوند؟ آیا مجموعه داده‌ی Dataset1.csv نیازی به این اقدامات دارد؟ توضیح داده و در صورت لزوم این موارد را بر روی مجموعه داده اعمال کنید. ۵٪

۳- تابعی بنویسید که با استفاده از روش گرادیان نزولی و با دریافت داده‌ها، درجه، تعداد تکرار، نرخ یادگیری و ضریب رگولاریزیشن یک خط/منحنی بر روی داده‌ها برازش کند. با استفاده از این تابع به ازای پنج درجه‌ی مختلف یک نمودار بر روی نقاط برازش کنید (ضریب رگولاریزیشن را صفر قرار دهید). برای هر درجه، سه مقدار تکرار با فاصله‌ی مناسب انتخاب کرده و با توجه به آن مقادیر MSE را بیابید و گزارش کنید (به ازای هر درجه و تعداد تکرار مقدار حدودی مناسب را برای نرخ یادگیری بیابید و بر اساس آن برازش را انجام دهید). ۲۰٪

خروجی مورد نظر: تصویر پنج نمودار تخمین زده شده با درجات مختلف به همراه داده اصلی در یک پلات. این تصویر را برای هر سه حالت تعداد تکرارها رسم کنید (در مجموع سه تصویر). نرخ یادگیری مناسب یافته شده و

[2] shuffle
[3] normalization

مقدار خطا برای هر دو فاز آموزش و آزمون برای کلیه مقادیر را گزارش کنید. تحلیل نمودارها و مقادیر در گزارش آورده شود.

۴- به ازای بهترین مقادیر یافته شده برای پارامترها در مرحله‌ی قبل و به ازای ۳ مقدار مختلف با فاصله‌ی مناسب برای ضریب رگولاریزیشن نمودار را بر روی نقاط برازش کرده و به همراه داده‌ها رسم کنید. مقدار خطای MSE برای هر دو فاز آموزش و آزمون و بردار ضرایب θ را گزارش کنید. تغییر ضریب رگولاریزیشن چه تاثیری بر روی اندازه بردار ضرایب θ دارد؟ ۵ ٪

خروجی مورد نظر: تصویر نمودار برازش شده بر روی نقاط به ازای بهترین مقدار پارامترهای یافته شده در مرحله‌ی قبل و مقادیر انتخاب شده برای ضریب رگولاریزیشن در یک پلات به همراه داده‌های اصلی. مقدار خطای MSE در هر دو فاز آموزش و آزمون، بردار ضرایب θ و تاثیرضریب رگولاریزیشن بر روی اندازه بردار ضرایب θ به ازای حالات ذکر شده.

۵- تابعی بنویسید که با دریافت داده‌ها و درجه، یک خط/منحنی به روش معادله نرمال بر روی داده‌ها برازش کند، سه درجه‌ی مختلف با فاصله مناسب را امتحان کنید و نمودار خط را همراه با داده‌ها رسم کرده و نتایج را بررسی کنید. ۲۰٪

خروجی مورد نظر: تصویر سه نمودار تخمین زده شده با درجات مختلف به همراه داده‌ی اصلی در یک پلات. مقدار خطا برای هر دو فاز آموزش و آزمون برای کلیه مقادیر باید گزارش شود. تحلیل نمودارها و مقادیر در گزارش آورده شود.

بخش دوم ۱۰٪

مجموعه داده‌ی Dataset2.csv را بارگذاری کنید.

۱- داده‌ها را رسم کنید.

۲- با استفاده از تابعی که در قسمت ۳ بخش اول نوشته‌اید و با درجه‌ی ۱، خطی را بر روی نقاط برازش کنید.

خروجی مورد نظر: نمودار تخمین زده شده با درجه ۱ به همراه داده‌ی اصلی در یک پلات. خطای MSE.

۳- با استفاده از یک کتابخانه‌ی آماده، با استفاده از LinearRegression خطی را بر روی نقاط برازش کنید.

خروجی مورد نظر: نمودار تخمین زده شده به همراه داده‌ی اصلی در یک پلات. خطای MSE.