



دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

درس مبانی یادگیری ماشین

تکلیف کامپیوتری اول

مهلت تحویل: ۲۰ آبان ۱۴۰۲

سؤال ۱

در این سوال قصد داریم بر روی مجموعه داده house_prices یک برازش خطی^۱ به روش نزول گرادیانی^۲ انجام دهیم. این فایل شامل اطلاعاتی از جمله area, bedrooms, bathrooms و ... است. و هدف پیش‌بینی مقدار Price است.

الف) ابتدا داده‌ها را از فایل house_prices.csv خوانده، ماتریس X, Y تشکیل دهید و داده‌ها را نرمال سازی کنید(۵ نمره)

ب) با در نظر گرفتن تابع خطای میانگین مربعات^۳ روش Batch GD را پیاده‌سازی کرده و نمودار مقدار خطا به تعداد گام‌های روزرسانی شده را رسم کنید(۱۵ نمره)

ج) فردی قصد دارد منزل خود که دارای ویژگی‌های زیر است را با قیمت ۹۵۰۰۰۰ به فروش برساند با توجه به مدل پیاده سازی شده در قسمت قبل نظر شما در مورد قیمت پیشنهادی او چیست؟(۵ نمره)

area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	parking
7700	3	2	1	1	0	0	2

د) روش‌های mini-batch GD و SGD^۴ را پیاده‌سازی کنید و نمودار مقدار خطا به تعداد گام‌های روزرسانی شده را رسم کنید پس از آن سه روش SGD, mini-batch GD و batch GD را از نظر سرعت همگرایی و مقدار خطا مقایسه کرده و مشاهدات خود را توجیه کنید(۱۰ نمره)

ه) تاثیر بزرگ بودن یا کوچک بودن مقدار نرخ یادگیری^۵ را بررسی کنید(۱۰ نمره)

و) مدل برازش خطی را با استفاده از کتابخانه‌های آماده پایتون آموزش دهید و میزان خطا و سرعت این روش را با مدل پیاده‌سازی شده در قسمت ب مقایسه کنید(۵ نمره)

سؤال ۲

در این سوال قصد داریم که راه‌حل بسته^۶ را پیاده‌سازی کرده و همچنین به بررسی مقدار پیچیدگی مدل بر روی مقدار خطا در داده‌های آموزش و ارزیابی بپردازیم.

الف) مجموعه داده‌های Data.csv در اختیار شما قرار گرفته است این مجموعه داده شامل مقادیر x و y است که مربوط به تعدادی نقطه در جدول مختصات است. داده‌ها را از این فایل خوانده سپس ۸۰ درصد آن‌ها را به صورت تصادفی برای آموزش و ۲۰ درصد باقی‌مانده را برای تست در نظر بگیرید(۵ نمره) (به منظور جلوگیری از نتایج نامطلوب برای تقسیم داده‌ها از random_state=10 استفاده کنید)

¹Linear Regression

²Gradient Descent

³Mean Squared Error

⁴Stochastic Gradient Descent

⁵Learning Rate

⁶Closed-form Solution

ب) قسمتی از داده‌های آموزش را به داده‌های اعتبارسنجی^۱ اختصاص دهید. در ادامه یک مدل برازش خطی با استفاده از روش نزول گرادینتی به صورت برداری (بدون حلقه) بر روی داده‌ها آموزش دهید. برای این منظور از داده اعتبارسنجی برای تنظیم هایپرپارامتر نرخ یادگیری استفاده کنید. سپس نمودار مربوط به نقاط موجود در مجموعه داده‌ها و خط برازش شده بر روی آن‌ها را رسم کنید. در نهایت مقدار خطای نهایی را بر روی مجموعه داده آموزش و تست بررسی کنید و مشاهدات خود را توجیه کنید (۱۰ نمره)

ج) برازش خطی را بر روی مجموعه داده‌ها با استفاده از راه‌حل بسته انجام داده نمودار مربوط به نقاط موجود در مجموعه داده‌ها و خط برازش شده بر روی آن‌ها را رسم کنید. همچنین مقدار خطای نهایی را بر روی مجموعه داده آموزش و تست بررسی کنید. مدل پیاده‌سازی شده در این قسمت را با مدل قسمت قبل از نظر سرعت همگرایی و میزان خطا مقایسه کنید و مشاهدات خود را توجیه کنید (۱۰ نمره)

د) پیچیدگی مدل پیاده‌سازی شده در قسمت قبل (ج) را با اضافه کردن توان‌های اعداد موجود در ماتریس X افزایش دهید. این کار را با چند جمله‌ای‌های درجه ۵ و ۷ و ۹ و ۱۱ و ۱۳ و ۱۵ تکرار کنید و در هر مرتبه میزان خطا بر روی داده‌های آموزش و تست را بررسی کرده و نمودار مربوط به نقاط و منحنی برازش شده بر روی آن‌ها را رسم کنید سپس مشاهدات خود را توجیه کنید (۱۵ نمره)

ه) مدلی را بر اساس راه‌حل بسته Ridge Regression پیاده‌سازی کنید. سپس این مدل را بر روی چند جمله‌ای‌های درجه ۱۳ و ۱۵ اجرا کنید. سپس مقدار خطای نهایی بر روی داده‌های آموزش و تست را بررسی کرده و نمودار مربوط به نقاط و منحنی برازش شده بر روی آن‌ها را رسم کنید سپس مشاهدات خود را توجیه کنید (۱۰ نمره)

و) (امتیازی) مدلی را بر اساس Lasso Regression با استفاده از کتابخانه‌های آماده پایتون پیاده‌سازی کنید. سپس وزن‌های مدل آموزش دیده را با مدل‌های قسمت‌های د و ه بر روی چند جمله‌ای‌های درجه ۱۳ و ۱۵ مقایسه کنید و مشاهدات خود را توجیه کنید (۱۰ نمره)

¹ Validation

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 حتماً 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 ۵ 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 تلگرام بپرسید. هم‌چنین می‌توانید برای رفع ابهامات با دستیاران آموزشی از طریق تلگرام در تماس باشید.

آیدی‌ها:

@R3Z4_JVN

موفق باشید.