



دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

درس یادگیری ماشین

تکلیف کامپیوتری اول

تاریخ تحویل: ۸ آبان ۱۴۰۰

سوال ۱

در این تمرین قصد داریم بر روی مجموعه داده insurance یک برازش خطی به روش Gradient Descent انجام دهیم. این فایل شامل اطلاعات Age, BMI, Children است. هدف پیش بینی Charges است. در جدول زیر چند رکورد از این مجموعه داده قابل مشاهده است.

Age	BMI	Children	Charges
19	27.9	0	16884.924
18	33.77	1	1725.5523
28	33.0	3	4449.462
33	22.705	0	21984.47061
32	28.88	0	3866.8552

الف) با در نظر گرفتن تابع خطای میانگین مربعات^۱ سه روش SGD^۲، Batch GD و Mini-batch GD را پیاده سازی کرده و نمودار خطا بر روی کل داده‌ها در هر گام بروز رسانی وزنها را رسم کنید. (برای رسم نمودار، محور افقی گام‌های بروز رسانی وزنها (Iteration) و محور عمودی میزان خطا بر روی کل داده‌ها می‌باشد) (۲۰ نمره)

ب) سرعت همگرایی و کمینه خطا برای سه روش پیاده سازی شده در قسمت قبل را با یکدیگر مقایسه کنید. (۵ نمره)

ج) با آزمایش تاثیر کوچک بودن مقدار نرخ یادگیری^۳ را بررسی کنید. (۵ نمره)

د) با آزمایش تاثیر بزرگ بودن مقدار نرخ یادگیری را بررسی کنید. (۵ نمره)

ه) با نرمال سازی داده، قسمت‌های الف و ب را تکرار کرده و نتایج را با حالت بدون نرمال سازی مقایسه کنید. (۱۰ نمره)

و) با استفاده از کتابخانه scikit-learn مدل Linear Regression را آموزش داده و با مدلی که خودتان پیاده سازی کرده اید از نظر میزان خطا و سرعت همگرایی مقایسه کنید. (۵ نمره)

ز) راه حل بسته^۴ برای مسئله برازش خطی را محاسبه کرده و سپس مقدار پارامترها و مقدار خطا را با مدل خودتان و مدل پیاده سازی شده در scikit-learn مقایسه کنید. (۱۰ نمره)

سوال ۲

در این سوال اثر تابع خطا را در برازش خطی بررسی خواهیم کرد. مجموعه دادگان insurance شامل متغیرهای پیشین سن، شاخص توده بدنی و تعداد فرزندان است و هدف پیش‌بینی هزینه درمان است. ۸۰ درصد داده‌ها را به صورت تصادفی برای آموزش و ۲۰ درصد داده‌ها را برای تست استفاده کنید.

¹ Mean Square Error

² Stochastic Gradient Descent

³ Learning Rate

⁴ Closed-form solution

الف) برازش خطی را روی مجموعه دادگان آموزشی با استفاده از دو تابع خطای MAE و MSE با روش بهینه سازی SGD انجام داده و نمودار تغییرات خطا را پس از هر مرحله به روز رسانی رسم کنید. (۲۵ نمره)

ب) تغییرات خطا را بین دو روش بخش الف مقایسه کرده و مشاهدات خود را توجیه کنید. (۵ نمره)

ج) داده ها را نرمال کنید و بخش های الف و ب را تکرار کرده و تغییرات مشاهده شده را توجیه کنید. (۱۰ نمره)

د) (امتیازی) بخش های الف و ب را با استفاده از کتابخانه های آماده پیاده سازی 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 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 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. هم‌چنین می‌توانید برای رفع ابهامات با دستیاران آموزشی از طریق تلگرام در تماس باشید.

آیدی‌ها:

@Fatemeh2114P

@amir7d0

@mastaraan

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

arsh.2001.1379@gmail.com

موفق باشید.