دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

درس یادگیری ماشین

تکلیف کامپیوتری دوم

تاریخ تحویل: ۲۴ آبان ۱۴۰۰

سؤال ۱

تابع هزینه رگرسیون Ridge و Lasso به ترتیب مطابق شکل ۱ و ۲ می باشد.

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

شکل ۱: تابع هزینه رگرسیون Ridge

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

شکل ۲: تابع هزینه رگرسیون Lasso

در این دو رابطه:

- w_j : وزن‌های ویژگی‌ها
- x_{ij}: ویژگی‌ها
- y_i: خروجی های موجود
- $\widehat{y_i}$: خروجی های پیش‌بینی شده توسط مدل
- M: تعداد نمونه‌ها
- P: تعداد ویژگی‌ها
- λ: ضریب کاهش ضرایب (منظم سازی ضرایب)

حال در ادامه می‌خواهیم به پیاده سازی این دو نوع رگرسیون بدون استفاده از کتابخانه‌های آماده پایتون بر روی دیتا BigMart بپردازیم. (۶۵ نمره)

ابتدا به توصیف دیتاست مورد نظر می‌پردازیم: دانشمندان داده در BigMart داده های فروشسال ۲۰۱۳ را برای ۱۵۵۹ محصول در ۱۰ فروشگاه در شهرهای مختلف جمع آوری کرده‌اند. همچنین ویژگی های خاصی برای هر محصول و فروشگاه تعریف شده است. دیتاست مورد نظر شامل ۱۴۲۰۴ مورد است که ۸۵۲۳ مورد برای مجموعه آموزشی در نظر گرفته شده است و ۵۶۸۱ مورد برای مجموعه تست در نظر گرفته شده است. مجموعه آموزشی شامل ۱۲ ستون و مجموعه تست شامل ۱۱ ستون می‌باشد. متغیر هدف در این مجموعه داده متغیر Item_Outlet_Sales می باشد. در ادامه نمایی از دو مجموعه داده می‌بینید.

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052

شکل ۳: نمایی از مجموعه آموزشی

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type
0	FDW58	20.750	Low Fat	0.007565	Snack Foods	107.8622	OUT049	1999	Medium	Tier 1	Supermarket Type1
1	FDW14	8.300	reg	0.038428	Dairy	87.3198	OUT017	2007	NaN	Tier 2	Supermarket Type1
2	NCN55	14.600	Low Fat	0.099575	Others	241.7538	OUT010	1998	NaN	Tier 3	Grocery Store
3	FDQ58	7.315	Low Fat	0.015388	Snack Foods	155.0340	OUT017	2007	NaN	Tier 2	Supermarket Type1
4	FDY38	NaN	Regular	0.118599	Dairy	234.2300	OUT027	1985	Medium	Tier 3	Supermarket Type3

شکل ۴: نمایی از مجموعه تست

هدف این است که مدل‌هایی برای پیش بینی براساس دو نوع رگرسیون Lasso و Rdige و رگرسیون خطی گفته شده بسازیم و فروش هر محصول را در یک فروشگاه خاص پیدا کنیم. برای این هدف به صورت قدم قدم موارد خواسته شده را انجام دهید.

آماده‌سازی داده (که به عنوان «پیش‌پردازش داده‌ها» نیز گفته می‌شود) فرآیند تبدیل داده‌های خام است به طوری که دانشمندان و تحلیلگران داده می‌توانند آن را از طریق الگوریتم‌های یادگیری ماشین اجرا کنند تا بینش‌ها را کشف کنند یا پیش‌بینی کنند. فرآیند آماده‌سازی داده‌ها می‌تواند با مسائلی مانند:

- پاکسازی داده ها: شناسایی و تصحیح اشتباهات یا خطاهای موجود در داده ها
- انتخاب ویژگی: شناسایی متغیرهای ورودی که بیشترین ارتباط را با کار دارند
- تبدیل داده ها: تغییر مقیاس یا توزیع متغیرها
- کاهش ابعاد: ایجاد پیش بینی های فشرده از داده ها

در ارتباط باشد. روش‌های مختلفی برای آماده سازی داده‌ها وجود دارد که عبارتند از:

- کاهش داده‌ها
  - نمونه گیری ویژگی (attribute sampling)
  - ثبت نمونه (record sampling)
  - تجمیع (aggregating)
- مقیاس مجدد داده‌ها (rescaling data)
  - نرمال سازی داده‌ها (data normalization)
  - نرمال سازی حداقل-حداکثر (min-max normalization)
  - مقیاس دهی دهدهی (decimal scaling)
- گسسته سازی داده‌ها (discretize data)

الف) در این قسمت به عنوان پیش پردازش داده‌ها یکی از روش‌های ذکر شده در بالا را برای داده‌ها عددی پیاده سازی کنید. در صورت دانستن روشی دیگر پیاده سازی و استفاده از آن موردی ندارد. در این بخش دو پیش پردازش اهمیت بیشتری دارد:

- مدیریت داده های از دست رفته [۱]
- تبدیل داده های دسته بندی به داده های عددی مناسب

ب) بررسی میزان وابستگی متغیر(ویژگی)ها به یکدیگر (محاسبه ماتریس همبستگی)

ج) بررسی توزیع هر یک از متغیرها؛ یعنی نمودار هر یک از متغیرها را براساس تعداد و یا بازه رسم کنید و اگر از توزیع خاصی پیروی می‌کنند، نام توزیع را ذکر کنید.

د) بررسی تأثیر هر یک از متغیرها بر متغیر هدف [۲] (Item_Outlet_Sales) براساس رسم نمودارها وابستگی هدف به هر یک از متغیرها را بررسی کنید. سپس نتیجه را با نتیجه بدست آمده از ماتریس همبستگی براساس رابطه بین هدف و هر یک از متغیرها بررسی کنید.

و) برای پیاده سازی هر یک از مدل‌ها ابتدا چهار سری مجموعه آموزشی، ارزیابی و تست با استفاده از تابع train_test_split از روی مجموع آموزشی train مطابق با درصدهای داده شده درست کنید.

- ۸۰ درصد مجموعه داده train، ۱۰ درصد مجموعه داده ارزیابی و ۱۰ درصد مجموعه داده تست
- ۶۰ درصد مجموعه داده train، ۲۰ درصد مجموعه داده ارزیابی و ۲۰ درصد مجموعه داده تست
- استفاده ازKfold cross validationبرای تقسیم بندی مجموعه آموزشی و ارزیابی مدل(حداقل ۳ مقدار متفاوت k بررسی شود که مفاهیمoverfit و underfit را پوشش دهد و یکی از مقادیر یک مدل مطلوب باشد.)

حال در ادامه در هر قسمت رگرسیون ذکر شده را به صورت کلی پیاده‌سازی کنید و بر روی هر یک از تقسیم‌بندی‌ها موارد خواسته شده را انجام دهید.

ه) یک مدل براساس رگرسیون خطی بدون استفاده از کتابخانه‌های آماده پیاده سازی کنید و مدل را براساس مجموعه آموزشی و تقسیم بندی‌های در اختیار قرار داده شده آموزش دهید و نتایج را به صورت بصری(نموداری) مشاهده و تحلیل کنید.

ی) یک مدل براساس رگرسیون Ridge بدون استفاده از کتابخانه‌های آماده پایتون پیاده سازی کنید و مدل را براساس مجموعه آموزشی و تقسیم بندی‌های در اختیار قرار داده شده آموزش دهید و نتایج را به صورت بصری مشاهده و تحلیل کنید.

---

[۱] missing values

[۲] target variable

ک) یکی از دلایل استفاده از رگرسیون Lasso قابلیت feature selection می‌باشد. ابتدا توضیح دهید که چگونه این رگرسیون در feature selection کمک می‌کند؟ در ادامه مطلوب است:

ک - الف

پیاده سازی یک مدل براساس رگرسیون Lasso بدون استفاده از کتابخانه‌های آماده پایتون و قابلیت feature Selection سپس مدل را براساس مجموعه آموزشی و تقسیم بندی‌های در اختیار قرار داده شده آموزش دهید و نتایج را به صورت بصری(نموداری) مشاهده و تحلیل کنید.

ک - ب

پیاده سازی یک مدل براساس رگرسیون Lasso بدون استفاده از کتابخانه‌های آماده پایتون و قابلیت feature Selection سپس مدل را براساس مجموعه آموزشی و تقسیم بندی‌های در اختیار قرار داده شده آموزش دهید و نتایج را به صورت بصری مشاهده و تحلیل کنید.

گ) نخست هر مدل را با داده های آموزشی آموزش دهید و سپس بر روی داده های تست آزمایش کنید. نتایج آزمایش را یک بار با تابع هزینه MSE و بار دیگر با تابع هزینه MAE ارائه نمایید و این نتایج به دست آمده را با یکدیگر مقایسه کنید.

ل) مفاهیم underfit و overfit بر روی هر یک از مدل‌های پیاده سازی شده به ازای نرخ‌های یادگیری متفاوت و مقادیر متفاوت برای ضریب پنالتی در نظر گرفته شده برای رگرسیون‌های Ridge و Lasso مختلف بررسی کنید و مشاهدات خود را توصیف کنید.

م) با اضافه کردن توان دوم ویژگی‌های ITEM_MRP و Outlet_Year بار دیگر مدل‌ها را آموزش دهید و با توجه به تقسیم بندی مورد نظر روی داده‌های تست(۱۰ درصد داده‌ها) نتایج یک مرتبه با تابع هزینه MAE و بار دیگر با تابع هزینه MSE بررسی کنید. مفاهیم underfit و overfit نیز بررسی شوند.

نمره اضافه) می‌توانید مدل‌های خود را بر روی مجموعه داده تست قرار داده شده با نام BigMart_Dataset_Testset اجرا کنید و نتایج را با فرمت خواسته شده توسط چالش سایت Kaggle به نام Big Mart Sales Prediction بارگذاری کنید. سایت معیار مقایسه‌اش RMSE است و خود سایت با توجه به فایلی که با فرمت خواسته شده بارگذاری کرده‌اید، RMSE را محاسبه می‌کند و رتبه شما را نیز تعیین می‌کند. نمونه فایلی که باید بر روی سایت قرار گیرد به نام sample_submission در فایل زیپ قرار دارد. در صورت انجام این قسمت نتیجه هر مدل(شامل مقدار RSME و رتبه شما) در فایل نوت بوک باید مشخص باشد.

سؤال ۲

مجموعه داده data.csv در اختیار شما قرار گرفته است. این مجموعه داده شامل مقادیر X و y مربوط به چند نقطه است. در این سؤال قصد داریم توزیع این نقاط را بیابیم. (۳۵ نمره)

الف) ابتدا این مجموعه داده را بخوانید سپس پراکندگی داده‌ها را روی نمودار نشان دهید. (به کمک تابع scatter در کتابخانه matplotlib)

ب) در این قسمت فرض می‌کنیم توزیع داده‌ها از تابع گوسی با رابطه‌ای روبرو باشد: $p(x) = a.e^{-(x-\mu)^2}/2\sigma^2$. تابعی بنویسید که ورودی‌های آن σ ، μ ، a ، x و خروجی آن تابع توزیع گوسی باشد.

ج) حال برای هر یک از پارامترهای σ ، μ ، a یک حدس اولیه در نظر بگیرید و آنها را به عنوان ورودی به تابعی که در قسمت قبل پیاده سازی کردید بدهید. خروجی را y_pred بنامید و آن را روی نمودار داده‌ها نمایش دهید. (به کمک تابع plot در کتابخانه matplotlib) (یعنی نمودار نهایی هم شامل پراکندگی داده‌ها و هم شامل خروجی تابع باشد.) آیا حدس اولیه شما درست بوده است؟ میانگین مربعات خطا بین مقادیر واقعی و مقادیر به دست آمده را محاسبه و گزارش کنید. (تابع خطا را خودتان پیاده سازی کنید.) (نیازی نیست حدستان کاملا درست باشد، صرفاً برسی یک حدس کافی است.)

د) در این قسمت از تابع curve_fit از کتابخانه scipy در پایتون استفاده کنید و پارامترهای تابع گوسی را به کمک آن به دست آورده و این بار خروجی را براساس این پارامترها به دست آورده و آن را روی نمودار نمایش دهید. به کمک نمودار میانگین مربعات خطای واقعی و مقادیر بدست آمده در این قسمت را محاسبه و با قسمت قبلی مقایسه کنید.

ه) در این قسمت فرض می‌کنیم توزیع داده‌ها کسینوسی به شکل رابطه زیر است: $f(x) = a.cos(b, x)$ مجدداً به کمک تابع curve_fit پارامترهای این تابع را به دست آورید و خروجی را براساس این پارامترها محاسبه کرده، نمودار آن را روی داده‌ها برازش کرده و نشان دهید. و نیز میانگین مربعات خطا بین مقادیر واقعی و به دست آمده محاسبه و تفسیر کنید.

و) در صورتی که در قسمت قبلی خروجی مناسب نیست، با توجه به شکل اولیه داده ها (یعنی قسمت الف) خودتان یک حدس اولیه برای پارامترهای a و b در نظر بگیرید و این حدس را به عنوان ورودی به تابع curve_fit بدهید. در این قسمت باید بتوانید با حدس پارامترهای مختلف بهترین پارامترها را به دست آورید به گونه ای که خروجی تابع روی داده‌ها به خوبی برازش شود. مجدداً مراحل محاسبه خروجی، رسم نمودار و محاسبه خطا را با پارامترهای جدید محاسبه شده توسط تابع curve_fit به دست آورده و نتایج را تفسیر کنید.

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 حتما 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. هم چنین می‌توانید برای رفع ابهامات با دستیاران آموزشی از طریق تلگرام و یا اسکایپ در تماس باشید.

آیدی‌ها:

@Fatemeh2114P

@amir7d0

@mastaraan

live:.cid.7f0be16d612107cc

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

arsh.2001.1379@gmail.com

موفق باشید.