بسمه تعالی

تکلیف دوم درس هوش محاسباتی موعد تحویل: ۲۱ فروردین ۱۴۰۰

در این تکلیف قصد داریم با مبانی رگرسیون خطی و رگرسیون لاجیستیک آشنا شویم. برای اینکار، مفاهیم این دو نوع رگرسیون را با استفاده از پیاده سازی آنها روی داده هایی که در اختیار شما قرار داده میشود تحلیل، و نتایج خود را در قالب یک فایل جداگانه یا کامنت داخل محیط jupyter گزارش کنید. توجه کنید که تنها آپلود کردن کد بدون توضیحات و تحلیل ممکن است نمره ای به شما اختصاص ندهد. در این تکلیف از زبان برنامه نویسی پایتون و jupyter notebook به عنوان محیطی برای کدنویسی و تست روش های خود استفاده کنید. ویدئوی آموزشی به این منظور روی سامانه الکترونیکی دروس قرار داده شده است. در ضمن توجه شود که کدهای مربوط به الگوریتم هایی که هدف آموزشی کلاس درس بوده اند (مثل گرادیان دیسنت) باید توسط خود شما پیاده سازی شوند. گزارش نهایی باید شامل توضیح پیاده سازی ها و نتایج و تحلیل های خواسته شده در متن تمرین باشد. تاخیر در تحویل تکلیف باعث کسر ۱۰ درصد تا ۲۴ ساعت پس از آخرین مهلت تحویل، و بعد از آن باعث کسر ۲۰ درصد از کل نمره کسب شده میشود. به تکالیف تحویل داده شده بعد از ۴۸ ساعت نمره ای تعلق نخواهد گرفت.

رگرسیون خطی

در این قسمت از تکلیف قصد داریم از رگرسیون خطی برای پیش بینی قیمت اتومبیل با استفاده از ویژگی های عددی آن استفاده کنیم. پایگاه داده ای که در اختیار شما قرار دارد شامل اطلاعاتی مثل طول اتومبیل، عرض اتومبیل، تعداد سیلندر، جنس بدنه و ... است. توجه کنید که الگوریتم های یادگیری ماشین تنها میتوانند با ویژگی های عددی کار کنند. بنابراین برای شروع، باید ویژگی های دسته ای[1] را از پایگاه داده حذف کنید.

اختیاری (نمره تشویقی): دو روش را برای تبدیل ویژگی های دسته ای به ویژگی عددی پیشنهاد و پیاده سازی کنید.

۱- رگرسیون خطی تک متغیره

سعی کنید با استفاده از یکی از ویژگی ها، مقدار هدف را با استفاده از روشی که برای رگرسیون خطی تک متغیره در کلاس ارائه شد تخمین بزنید. نمودار رگرسیون خطی بر حسب متغیر ذکر شده و تابع هزینه $J(\theta)$ را بر حسب متغیر های θ_0 و θ_1 رسم کنید. در مورد روشهای مقدار دهی اولیه پارامتر ها تحقیق کنید و اثر هر روش را روی نتایج به دست آمده بررسی کنید (حداقل دو روش).

۲- رگرسیون خطی چند متغیره

سعی کنید با استفاده از تمامی اطلاعات پایگاه داده، مقادیر هدف را با استفاده از روشی که برای رگرسیون خطی چند متغیره در کلاس ارائه شد را تخمین بزنید. نمودار MSE را برای تمامی مراحل روش gradient descent گزارش کنید. در مرحله بعد، با استفاده از با حذف تعدادی از متغیر ها که به نظر شما اهمیت زیادی در یادگیری ندارند تعداد پارامترهای یادگیری را کاهش دهید. اینکار را با استفاده از رسم نمودارهای دوبعدی هر ویژگی نسبت به مقدار هدف انجام دهید. توضیح دهید انتخاب نامناسب ویژگی ها چه تاثیری روی overfitting مدل خواهد داشت و چه روش هایی برای انتخاب ویژگی و کاهش بعد داده ها وجود دارد؟ روش های پیشنهادی را به صورت مختصر معرفی کنید.

ارزیابی مدل رگرسون خطی

برای ارزیابی مدل رگرسیون خطی، باید چند فرض را در مورد مدل رگرسیون خطی بررسی کنیم. برای مدل رگرسیون خطی چند متغیره در قسمت قبل، این فرض ها را بررسی و در صورت ابهام در مورد هرکدام از فرض ها از اینترنت برای آموزش استفاده کنید. در نهایت بعد از بررسی و تحلیل هرکدام از این معیار ها بگویید آیا مدل شما به خوبی پایگاه داده داده شده را توصیف کرده است یا خیر.

۱- اگر مدل به درستی داده ها را توصیف کرده باشد، یابد رابطه خطی بین مقادیر تخمین زده شده و مقادیر واقعی هدف وجود داشته باشد. (برای بررسی این مورد نمودار دوبعدی این رابطه ها را رسم کنید.)

---

[1] Categorical

۲- نمودار خطای باقی مانده[۲] باید دارای توزیع نرمال باشد.

۳- میانگین خطای باقیمانده باید صفر یا خیلی نزدیک به صفر باشد.

۴- در مدل رگرسیون خطی تمامی ویژگی ها باید نرمان چند متغیره باشند. برای این تست میتوان از نمودار Q-Q استفاده کرد.

۵- در رگرسیون خطی فرض میشود ویژگی ها چند خطی[۳] نیستند. چند خطی بودن زمانی اتفاق میافتد که ویژگی ها همبستگی[۴] زیادی نسبت به هم داشته باشند. برای اندازه گیری این معیار میتوانید از ضریب VIF[5] استفاده کنید.

رگرسیون لاجیستیک

در این قسمت از تمرین قصد داریم با استفاده از روش هایی که برای رگرسیون لاجیستیک معرفی شد، ابتلا یا عدم ابتلا به دیابت را برای بیماران با رنج سنی و پیشینه پزشکی متفاوت پیش بینی کنیم. برای اینکار ابتدا یکی از ویژگی های ورودی را انتخاب و مقادیر هدف این ویژگی را با استفاده از نمودار نمایش دهید. سپس توضیح دهید چرا برای یادگیری این مساله نیاز به استفاده از روش feature mapping داریم. با روش feature mapping که در کلاس معرفی شد، داده ها را به فضایی با بعد بیشتر انتقال دهید. بعد از انجام بهینه سازی تابع هزینه و انتخاب بهترین پارامتر ها با استفاده از روش gradient descent، سعی کنید با اتخاذ مقادیر مختلف پارامتر regularization و رسم مرزهای تصمیم گیری (برای یک ویژگی)، دو پدیده overfitting و underfitting را در یادگیری پارامترها توضیح دهید.

موفق باشید

نسرین صالحی

[2] Residual error

[3] Multicollinearity

[4] Correlated

[5] Variance inflation factor