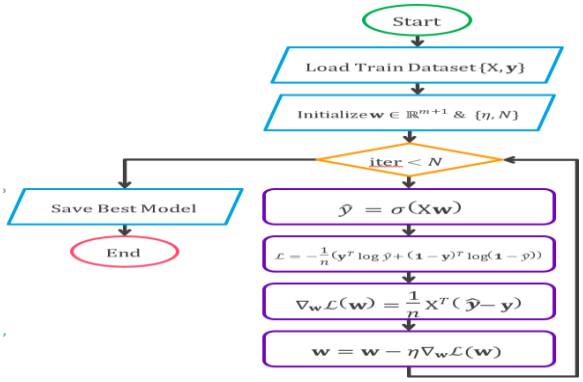
پاسخ سوالات سری اول درس یادگیری ماشین

محمدرضا حاجی نیا

سوال اول



فلوچارت کلی اموزش یک مدل طبقه بند خطی را در بالا مشاهده میکنیم.

مراحل فلوچارت به ترتیب شامل:

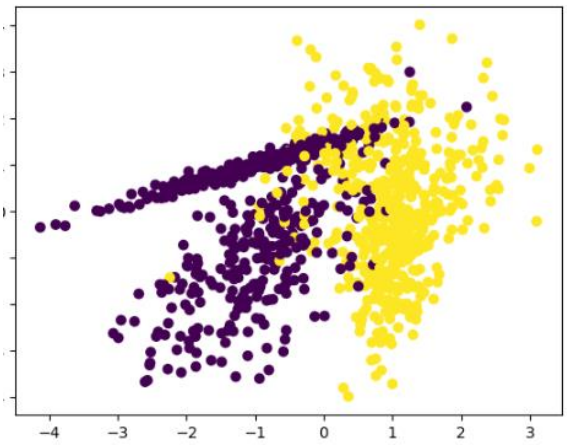
۱. **بارگذاری داده‌های آموزش** : در ابتدا، داده‌های آموزش (ورودی‌ها و برچسب‌ها) بارگذاری می‌شوند.
۲. **مقداردهی اولیه وزن‌ها و بایاس** : وزن‌ها و بایاس به صورت تصادفی مقداردهی اولیه می‌شوند.
۳. **پیش‌بینی خروجی:** در هر تکرار، خروجی پیش‌بینی شده با استفاده از تابع فعال‌ساز سیگموید محاسبه می‌شود.
٤. **محاسبه خطا :**خطای بین خروجی پیش‌بینی شده و برچسب واقعی با استفاده از تابع هزینه لگاریتمی محاسبه می‌شود.
٥. **به‌روزرسانی وزن‌ها** :گرادیان هزینه نسبت به وزن‌ها محاسبه می‌شود و سپس با استفاده از نرخ یادگیری، وزن‌ها به‌روزرسانی می‌شوند.
۶. **تکرار فرآیند:** باتوجه به تعداد دوره های اموزشی این فرایند تکرار می شود .و اگر به پایان برسد مدل ذخیره می شود.

تغییر حالت از دو کلاس به چند کلاس:

- ✓ تغییر تابع فعال ساز.تابع سیگموید دوحالت ۰ یا ۱ رو داره که باینری است میتوان از توابع دیگر برای این موضوع استفاده کرد.
- ✓ همچنین باتغییر کلاس قطعا در مقدار دهی وزن ها نیز تغییراتی ایجاد می شود.

سوال دوم

یک دیتاست با ویژگی ها و کلاس ها خواسته شده ایجاد شد.در ابتدا این دیتاست چالش برانگیز و سخت بود زیرا داده ها بسیار به یکدیگر نزدیک کلاس بندی ها دارای دو خوشه بودند.



از بعضی پارامتر ها استفاده شد تا نوع کلاس بندی بهتر مشخص بشود .توضیح برخی از پارامتر های مهم را در این قسمت داریم.

پارامتر n_clusters_per_class برای تعیین تعداد خوشه در هر کلاس است .در اینجا مقدار ۱است پس ، هر کلاس فقط یک خوشه دارد، یعنی داده‌های هر کلاس به یکدیگر نزدیک تر هستند.

پارامتر class_sep میزان تفکیک بین داده‌های دو کلاس را تعیین می‌کند.یعنی دوری و نزدیکی داده ها از هم.

پارامتر n_redundant تعداد ویژگی‌های اضافی و تکراری است که به داده‌ها اضافه می‌شود. با تنظیم این پارامتر به صفر، ویژگی‌های تکراری اضافه نمی‌شود.

برای چالش برانگیز تر کردن دیتاست میتوانیم مرز های تصمیم گیری را نزدیک تنظیم کنیم.یا خوشه بندی داده ها را بیشتر کنیم. مثلا میتوانیم class_sep را کاهش بدهیم. یا n_clusters_per_class افزایش بدهیم.همچنین می توانیم بین داده ها نویز ایجاد کنیم که بسیار تاثیر گذار است.

سوال سوم

برای ایجاد یک مدل طبقه بندی خطی از دو مدل

LogisticRegression , SGDClassifier

استفاده شده است.داده ها به دو قسمت اموزش و تست تقسیم بندی می شوند.در ابتدا پارامتر مهمی که باید لحاظ بشه پارامتر stratify است .این پارامتر مشخص می کند که از هر کلاس به اندازه تقریبا برابر در مجموعه داده اموزش و تست قرار گیرد.

از مدل رگرسیون لجستیک LogisticRegression برای اموزش و ارزیابی عملکرد مدل استفاده می‌شود. سپس از داده های اموزشی برای اموزش مدل سپس ارزیابی انجام می شود. از ماتریس درهم‌ریختگی (confusion matrix) برای ارزیابی دقیق‌تر عملکرد مدل استفاده می شود.

بهینه ساز این مدل sag است یک روش تقریبی برای بهینه سازی وزن ها.که دو روش تصادفی و تکراری را انجام می دهد.

Max_iter تعداد دوره های اموزشی می باشد.

random_state برای یکسان بودن یا متفاوت بودن نتایج کاربرد بسیار زیادی دارد.(تکرار پذیری نتایج)

برای بهبود در این مدل تعداد دوره های اموزشی را تا حد مشخصی می شود بالا برد.

و مهمترین پارامتر solver است که باید با روش های مختلف و با توجه به داده ها انتخاب بشود.

مدل SGDclassifier از روش گرادیان کاهشی تصادفی برای آموزش استفاده می‌کند. پارامترهای مختلف مدل را استفاده کرده ایم و در این قسمت توضیح می دهیم.

تابع هزینه hinge: تابع هزینه که برای اندازه‌گیری خطا استفاده می‌شود. این پارامتر باید تست شود زیر توابع دیگری وجود دارند در این مدل.

max_iter تعداد دوره های اموزشی

learning_rate یا نرخ یادگیری یک پارامتر مهم است که مشخص می‌کند که میزان تغییرات وزن‌ها در هر مرحله از آموزش چقدر باشد. از این پارامتر جهت تنظیم سرعت یادگیری و همگرایی زودتر به مقدار بهینه استفاده می شود روش های مختلفی برای تعیین کردن دارد.

از پارامتر alpha استفاده می کنیم تا بتوانیم بیش برازش را کنترل کنیم.این پارامتر بازه تغییرات وزن را کنترل میکند تا مقادیر وزن در طول اموزش زیاد نشوند.افزایش یا کاهش این مقدار تاثیر زیادی در اموزش دارد.

مقدار زیاد گاها باعث می‌شود مدل به سختی با داده‌ها هماهنگ شود و وزن‌های بسیار کوچکی داشته باشد.

مقدار کم باعث می شود مدل به راحتی با داده ها هماهنگ بشه و وزن های بزرگتری بگیره.

پس باید با امتحان کردن این مقدار بدست بیاد.

سوال پنج

در این تمرین از کتابخانه deawdata استفاده می کنیم.ابتدا سه کلاس a,b,c ایجاد میکنیم با توجه به شکل مدنظر.

و ستون رنگ را حذف میکنیم.از انجایی که مقادیر x,y زیاد می باشند از نرمال سازی استفاده میکنیم.همچنین کلاس هارا به صورت عددی تغییر می دهیم یعنی a=0 ,b=1,c=2 . سپس داده ها به مجموعه اموزشی و تست تقسیم می شوند.

از مدل های قبلی استفاده می شود تا ارزیابی بشود مدل.همچنین از ماتریس درهم ریختگی استفاده می شود تا تعداد داده های درست وغلط در هر کلاس مشخص بشود.در اخر هم مرز تصمیم گیری بین سه کلاس مشخص می شود.