دانشگاه تربیت مدرس

دانشکده مهندسی برق و کامپیوتر

پروژه اول درس یادگیری ماشین

مبانی یادگیری نظارت شده و مدل‌های پارامتریک

استاد دکتر قادری

دستیاران آموزشی

نیلوفر مقدس

عادل گلدسته

پارسا اسدنژاد

نیمسال تحصیلی ۱۴۰۴ – ۱۴۰۵

فهرست

مقدمه

مارکوس لیسینیوس کراسوس فرماندار سوریه با یک سپاه هفت لژیونی و نیرومند بدون اعلام جنگ در سال ۵۴ پیش از میلاد از رود فرات گذشت. سپاهیان روم در آغاز پیروزی‌هایی بدست آوردند، گرچه موفقیت‌های کراسوس چیزی بیش از تاراج روستاها و آبادی‌های بی‌پناه میانرودان نبود. ارد دوم، فرستاده ای به نزد کراسوس فرستاد تا درباره این دشمنی از او توضیح بخواهد؛ او با بی اعتنایی به واگیز(فرستاده اشکانیان) گفت، پاسخم را در سلوکیه خواهم داد. بدین ترتیب جنگ بین این دو ابر قدرت قطعی شد. سورنا، فرمانده ایرانی همراه با سوارکاران دلیرش به مقابله با رومیان پرداختند و نهایتا بعد از نبرهای متعدد توانستند، پوبلیوس و کراسوس را از پای در آورده و ۷۰۰ سال صلح میان این دو امپراطوری به ارمغان آورند. این شکست برای رومیان به قدری سهمگین و مهلک بود که رومیان با شنیدن نام اشکانیان دچار اضطراب اجتماعی (social anxiety) می‌شدند.

دانشمندان و پژوهشگران یادگیری ماشین روم باستان که از این رخداد میان مردمانشان به ستوه آمده بودند، شروع به جمع آوری داده از میان رومیان کردند و توانستند داده‌هایی بیش از ۱۰۰۰۰ نفر را جمع آوری کنند. داده‌هایی همچون ضربان قلب، زمان استفاده از شبکه‌های اجتماعی و پاسخ به چند پرسش روانشناسی و ... . کاسیوس و وینیسیوس که از علاقه مندان به الگوریتم‌های یادگیری ماشین روم باستان بودند و مفهوم یادگیری نظارت شده را به تازگی آموخته بودند، متوجه شدند که در کنار هر نمونه از مجموعه

داده، یک برچسب وجود دارد که میزان اضطراب مردم را در یک محدوده ۱الی ۱۰ نشان می‌دهد. (عدد بالاتر نشان‌دهند میزان اضطراب بیشتر است). آنها تصمیم گرفتند که به کمک این مجموعه داده یک مدلی طراحی نمایند که بتواند با دقت مناسبی افراد مضطرب از خاطرات نبرد حران را از افراد دیگر جدا نمایند.

هدف پروژه

هدف از انجام این پروژه، یادگیری شما دانشجویان گرامی با مفاهیم زیر است:

- مدل‌های پارامتریک
- یادگیری با نظارت
- تفاوت میان مدل‌های Generative با Discriminative

مدل‌های پارامتریک

در این نوع از روش‌ها، مدل با فرض تعداد پارامتر ثابت، که این تعداد بدون توجه به اندازه مجموعه داده است، تلاش به یادگیری می‌کند. برای نمونه (رگرسیون خطی، رگرسیون لجستیک، پرسپترون). در مقابل مدل‌های پارامتریک، نوعی دیگری به نام مدل‌های غیر پارامتریک وجود دارد که روش یادگیری آنها تعداد خاصی پارامتر در نظر ندارد. به عبارت بهتر، تعداد پارامترها با توجه به اندازه مجموعه داده ممکن است متفاوت باشد، که در درس‌های جلوتر با این نوع از مدل‌ها آشنا خواهید شد.(KNN, SVM ، شبکه عصبی پیچیده)

یادگیری با نظارت

شاخه‌ای از یادگیری ماشین است که در آن ما مجموعه‌ای از نمونه‌ها داریم که برای هر نمونه، ورودی x و برچسب y مشخص است. هدف کلی ما این است که تابعی بسازیم که از ورودی جدید x، برچسب y را پیش‌بینی کند. که شامل دو دسته کلی است:

- طبقه‌بندی (Classification)
- رگرسیون (Regression)

❖ در این پروژه، ما یک مسئله از نوع طبقه‌بندی را بررسی می‌کنیم.

مدل‌های مولد در مقابل مدل‌های تمایز دهنده

مدل‌های مولد دسته‌ای از مدل‌ها هستند که یاد می‌گیرند، چگونه داده تولید شده است. مانند نظریه Naïve Bayes که در اسلایدها فصل ۳ فراگرفته‌اید. در مقابل این دسته، مدل‌هایی وجود دارند که مستقیما آموزش میابند که مرز بین کلاس‌ها کجا است و به توزیع داده‌ها کاری ندارند. مانند روش Logistic Regression.

- ❖ در این پروژه ما قصد داریم که روش‌های Naïve Bayes و Logistic Regression را پیاده سازی نموده و مرز تصمیم آنها را مقایسه و تحلیل نماییم.

مجموعه داده

این مجموعه داده شامل بیش از ۱۰۰۰۰۰ نمونه است که نشان‌دهنده رومیان بازمانده از نبرد حران با سطوح مختلف اضطراب اجتماعی، از خفیف تا شدید هستند که برای کاربردهای یادگیری ماشین و علوم داده، به ویژه در تجزیه سلامت روان طراحی شده است. این مجموعه در یک فایل csv جمع آوری شده به گونه‌ای که شامل ۱۸ ویژگی و نهایتا یک ستون به عنوان برچسب می‌باشد. به طور کلی ویژگی‌ها موجود از زمینه‌های متفاوتی اعم از ویژگی‌های شناختی همانند سن، جنیست، شغل و ویژگی‌های سبک زندگی همچون ضربان قلب، تعداد تنفس، میزان تعریق و سوابق سلامت روان مانند استفاده از دارو، رویداد مهم رخ داده به تازگی و ... را شامل می‌شود. که خود نشان دهنده تنوع داده‌های عددی و رشته‌ای می‌باشد.

- ❖ مجموعه داده در کنار صورت سوال به شما داده خواهد شد، به علاوه می‌توانید آن را از طریق این ریپازیتوری دریافت نمایید.

کتابخانه‌های مورد نیاز

برای حل این پروژه شما به تعدادی از کتابخانه‌های پایتونی نیاز خواهید یافت:

- Numpy
- Pandas
- Scikit-learn
- Matplotlib

این کتابخانه‌ها جهت بارگذاری داده‌ها، پردازش، ترسیم نمودارها و پیاده‌سازی الگوریتم های یادگیری ماشین استفاده می‌شوند. برای نصب این کتابخانه‌ها می‌توانید از دستور زیر استفاده نمایید. پیشنهاد می‌شود که از یک محیط مجازی چون venv یا موارد دیگر استفاده کنید. (نسخه‌های متفاوت کتابخانه‌ها همواره در حال توسعه هستند و ممکن است با یکدیگر سازگار نبوده و دچار conflict شوند، لذا بهتر است در یک محیط کپسوله شده کتابخانه‌ها را نصب کنید.)

pip install pandas numpy matplotlib scikit-learn

معیارهای ارزیابی

معیارهای ارزیابی یا Evaluation Metics قسمت مهمی از هر پروژه یادگیری ماشین هست و یادگیری اون می‌تونه شما را در سایر تمرین‌های این درس و یا حتی دروس دیگرتون در دوره ارشد یاری بده. شما باید پس از آنکه مدل‌های (Logistic Regression, Naïve Bayes) را آموزش دادید از این معیارها جهت بررسی کیفتشان استفاده نمایید.

- ماتریس در هم‌ریختگی یا confusion matrix
- دقت یا Accuracy
- دقت مثبت یا Precision
- بازیابی Recall (برخی به این معیار حساسیت یا Sensitivity نیز می‌گویند)
- F1-score
- Specificity
- Cross-Entropy Loss یا Log Loss
- AUC و ROC

معیار	مفهوم	زمان مناسب استفاده
Accuracy	درصد پیشبینی درست	داده‌ها متوازن باشند
Precision	دقت تشخیص کلاس مثبت	خطای مثبت مهم باشد
Recall	درصد کشف کلاس مثبت	از دست ندادن موارد مثبت مهم باشد
F1	تعادل بین precision و recall	داده نامتوازن
AUC	قدرت کلی مدل در نمایز بین کلاس‌ها	مقایشه بین مدل‌ها
Log Loss	کیفیت احتماهای خروجی	مدل‌های احتمالاتی(logistic Regression)

❖ معیارهایی که در بالا مشاهده می‌کنید ممکن است متناسب با شرایط مسائل گوناگون، استفاده شوند یا اصلا معیار مناسبی برای آن مسئله نبوده و استفاده نشوند. جهت سادگی، استفاده از سه معیار آخر اختیاری می‌باشد. در ادامه درس بیشتر با این معیارها آشنا خواهید شد.

مسیر انجام پروژه

در این قسمت، شما راهنمایی خواهید شد که مراحل را به چه صورتی انجام دهید.

گام اول – بارگذاری مجموعه داده و تحلیل اولیه

در این بخش شما باید داده را بارگذاری نموده و برخی آمار اولیه را نمایش دهید. خب برای این کار شما می‌توانید از کتابخانه‌های numpy و pandas استفاده کنید. کتابخانه numpy برای بارگذاری داده مورد نیاز است. اجباری به استفاده از pandas نیست اما اگر بتوانید داده‌ها رو بر روی یک دیتافریم قرار دهید، آنگاه درک داده برایتان ساده‌تر خواهد شد. در این قسمت سعی کنید داده ها را حس کنید.

- به نظر شما چه شغلی بیشتر در معرض اضطراب قرار دارد؟ (پاسخ این سوال جزو اهداف این پروژه نیست، اما به شما برای درک بهتر مجموعه داده کمک خواهد نمود.)

گام دوم – پیش پردازش

در این بخش بررسی کنید که آیا داده گمشده وجود دارد یا خیر، به نظرتان اگر داده گمشده بود چه کاری باید انجام دهیم؟ کار دیگری که بایستی در این مرحله انجام شود. مقیاس کردن داده‌هاست. در این قسمت کتابخانه scikit-learn می‌تواند به شما کمک نماید. داده‌ها را به دو قسمت train و test تقسیم نمایید. همینطور برای اینکه درک و حسی از داده بگیرید، ایده بدی نیست که به کمک کتابخانه matplotlb آن را رسم کنید.

- چرا feature standardization در قسمت پیش پردازش اهمیت دارد و اگر آن را انجام ندهیم چه اتفاقی رخ خواهد داد که دقت مدل را لکه دار خواهد نمود؟
- شاید چالشی که با آن مواجه شوید، رسم نمودار باشد، چرا که این مجموعه داده ۱۸ ویژگی دارد. به نظر شما چطوری میشه در یک نمودار که فضای دو بعدی دارد، ۱۸ ویژگی را نشان داد؟ (راهنمایی: دو ویژگی را به سلیقه خود انتخاب کنید، حال شما بایستی تصمیم بگیرید آن دو ویژگی کدام باشند. در ادامه درس یادگیری ماشین با روش‌های آشنا خواهید شد که بتوانید تمام این ۱۸ ویژگی را در یک فضای دو بعدی نمایش دهید.)

گام سوم – پیاده سازی الگوریتم logistic Regression

خب در این قسمت باید الگوریتم LR را پیاده سازی کنید اما خبر بد این است که نمی‌توانید از کتابخانه sklearn استفاده کنید(مگر در سال۵۴ پیش از میلاد چنین کتابخانه ای وجود داشت!!). در این قسمت باید تمامی توابع شامل:

- Sigmoid
- Cost function
- Gradient descent

را پیاده سازی نمایید.(اگه احساس می‌کنید تابع دیگری نیاز است می‌توانید آنها را بنویسید) هدف این قسمت پیاده سازی الگوریتم Logistic Regression از ابتدا یا scratch است. که برای راحتی می‌توانید از numpy استفاده کنید.

❖ دستیابی به دقت بالاتر منوط به انتخاب درست هایپرپارامترهاست. به نظر شما چه مقادیری بایستی به آنها بدهیم تا خطا کمتر بشود؟

گام چهارم – نمودار همگرایی و تحلیل نرخ یادگیری

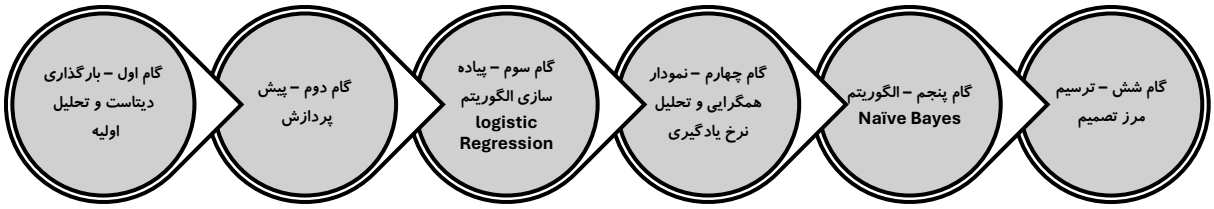
در این قسمت نمودار تابع هزینه در مقابل با تعداد تکرار یا iterationها متفاوت رو رسم کنید. یکی از هایپر پارامترهای مهم، learning rate یا ضریب یادگیری است. سه مقدار متفاوت بدهید و برای هر کدام یک نمودار رسم نمایید. (یا هر سه را در یک نمودار بکشید.) از این نمودار چه نتیجه‌ای می‌گیرید؟

گام پنجم – الگوریتم Naïve Bayes

خب در این بخش باید از این الگوریتم استفاده کنیم. خبر خوب اینه که می‌تونید از کتابخانه sklearn استفاده کنید و نیازی به پیاده سازی توسط خودتون نیست. اما کاری که باید انجام بدید تحلیل و مقایسه این روش با الگوریتم logistic regression است. به نظرتون کدامشان مناسب تر است؟ پاسخ به این سوال رو می‌تونید به کمک مقایسه معیارهای ارزیابی بدید. همینطور ایده بدی نیست که نتایج رو با الگوریتم logistic regression که توسط کتابخانه sklearn پیاده سازی شده است، مقایسه کنید.

گام شش – ترسیم مرز تصمیم

نمودار مرز تصمیم برای روش Naïve Bayes و Logistic Regression را رسم کنید. این نمودارهارا مقایسه کنید. تحلیل شما از نتایجتان چگونه است؟



- یک نمونه دمو از حل این پروژه، بر روی دیتاست Breast Cancer Wisconsin در این ریپازیتوری موجود است.