مینی‌پروژهٔ شمارهٔ یک

۱ پیش‌بینی آب‌وهوا مبنتی بر یادگیری ماشین

در این سوال به شبیه‌سازی مقاله A real-time collaborative machine learning based weather forecasting system with multiple predictor locations خواهیم پرداخت. این مقاله برای بهبود کیفیت مدل خود از اطلاعات نواحی مختلف برای پیش‌بینی آب‌وهوای یک ناحیه استفاده می‌کند. در این تحقیق الگوریتم‌های مختلفی مورد بررسی قرار گرفته است. هر یک از الگوریتم‌ها نیز برای چند حالت مختلف آموزش داده شده‌اند؛ یعنی مدل‌های مختلف برای داده هدف متفاوت. در این سؤال قصد داریم علاوه‌بر تمرین پیاده‌سازی الگوریتم‌های خطی برای مسئله رگرسیون [۱]، با collaborative machine learning نیز آشنا شویم.

۱.۱ دادگان

۱.۱.۱

درباره دیتاست جمع‌آوری شده توسط مقاله به صورت مختصر توضیح دهید.

۲.۱.۱

داده این تمرین در فایل `weather_prediction_dataset.csv` قرار دارد. در این پیاده‌سازی از دادگانی غیر از دادگان مقاله مورد مطالعه استفاده می‌شود. در مقاله از نواحی نزدیک به هم برای پیش‌بینی آب‌وهوا استفاده شده است. برای شبیه بودن این تمرین به مقاله از داداگان شهرهای فرانسه که در فایل داده وجود دارد استفاده می‌کنیم. چه شهرهایی از فرانسه در این داده موجود هستند؟ آن داده‌ها را ذخیره کرده و داده مربوط به شهرهای دیگر را حذف کنید. (لینک دادگان)

۳.۱.۱

چند نمونه [۲] در این داده وجود دارد؟ این دادگان چه بازه زمانی را شامل می‌شوند؟ مقاله چه پیش‌پردازش‌هایی را روی داده خود اعمال کرده است؟ آن‌ها را روی داده خود اعمال کنید.

۴.۱.۱

این مقاله برای پیش‌بینی x_t از پنجره‌های مختلفی به صورت $[x_{t-1}, ..., x_{t-n}]$ استفاده کرده است. برای هر یک از این حالات باید مدل‌های مجزایی آموزش داده شود. برای کاهش حجم کاری پیش‌بینی لحظه t، یکبار صرفا از لحظه $t-1$ استفاده کنید و بار دیگر از یک بازه زمانی دلخواه. بعد از آموزش تمامی مدل‌ها، باید عملکرد کلی را برای هر یک از این دو حالت مقایسه کنید. (پنجره‌های انتخاب شده می‌توانند دارای همپوشانی [۳] نیز باشند)

داده‌های مربوط به سال ۲۰۰۹ را جدا کرده و به عنوان دادگان آزمون استفاده نماید. دادگان آموزش و آزمون را با پنجره‌های انتخاب شده بسازید؛ مثلا اگر اندازه پنجره انتخاب شده ۵ با همپوشانی ۴ باشد، داده آزمون باید از شکل $n \times 365$ به شکل $n \times 5 \times 361$ درآید.

[۱] regression
[۲] sample
[۳] overlap

۲.۱ آموزش مدل

در این بخش باید یک شهر را انتخاب کرده و تمامی مدل‌های خواسته شده را برای آن شهر پیاده‌سازی کنید.

۱.۲.۱

مفهوم collaborative machine learning را توضیح دهید. collaborative machine learning در چهارچوب این مقاله به چه صورت استفاده شده است؟ شرح دهید.

۳.۱

مقاله برای پیش‌بینی هر متغیر از دو متغیر دیگر استفاده کرده‌است. مثلا برای پیدا کردن دما در لحظه $t+1$ از دما و رطوبت در لحظه t در تمامی شهرهای دیگر و شهر هدف استفاده شده است. شما باید برای مدل‌های خود از تمامی ویژگی‌هایی که در اختیار دارید استفاده کنید. مثلا برای یک مدل رگرسیون خطی خواهید داشت:

$$x_{Ii_{t+1}} = \beta_{0_t} + \sum_{k=1}^{3}\sum_{j=1}^{8}\beta_{kj_t}x_{kj_t} \qquad (۱)$$

در معادله ۱، x_{kj_t} به معنای ویژگی j در شهر k است.

در این بخش شما باید بدون استفاده از مدل آماده، یک رگرسیون چندجمله‌ای [۱] را پیاده‌سازی کنید. از معیاری که مقاله برای ارزیابی مدل‌های خود استفاده کرده بهره بگیرید و آن معیار را در طول آموزش مدل خود برای دادگان آموزش و آزمون نمایش دهید.

نکته ۱: برای پیاده سازی مدل خود باید یک حلقه تشکیل دهید که آموزش مدل در آن صورت گیرد. این حلقه باید شامل پیش‌بینی ورودی، محاسبه خطا، محاسبه گرادیان و بروزرسانی وزن‌ها باشد. در انتهای هر حلقه باید مقدار خطا مدل برای دادگان آموزش و آزمون محاسبه شده و ذخیره شود تا در انتها آموزش مدل نمایش داده شود.

نکته ۲: باید حلقه آموزش را طوری طراحی کنید که هنگامی که خطا از حد مشخصی کمتر شد، آموزش متوقف شود.

نکته ۳: از کتابخانه `tqdm` استفاده کنید و نوار پیشرفت [۲] را برای هر حلقه آموزش نشان دهید. نوار پیشرفت باید خطا را برای دادگان آموزش و آزمون در انتهای هر حلقه نشان دهد.

۱.۳.۱

از بخش مدل‌های خطی کتابخانه `scikit-learn` ۳، مدل رگرسیون انتخاب کرده و به صورت خلاصه تئوری آن‌ها را با فرمول‌های ریاضی مرتبط توضیح دهید. سپس از هر سه آن مدل‌ها استفاده کرده و مقاله را پیاده‌سازی کنید. مدل‌های مختلف را با هم مقایسه کرده و بهترین مدل را انتخاب کنید.

امتیازی

مدل‌ها را برای دو شهر دیگر نیز آموزش دهید.

امتیازی

ویژگی‌های یکسانی از سه شهر دیگر انتخاب کنید. سپس مدل‌هایی که آموزش داده‌اید را روی داده‌های یکی از آن شهرها ارزیابی کنید. چه عملکردی را مشاهده می‌کنید؟

نکته: برای اطلاعات عمومی خود درباره یادگیری انتقالی [۳] مطالعه کنید.

---

[۱] polynomial regression
[۲] progress bar
[۳] transfer learning

۲ تشخیص عیب یاتاقان غلتشی بر مبنای دسته‌بندی‌های سلسله مراتبی

در این سوال، قصد داریم مراحل مختلف انجام یک پایان‌نامه را بررسی کرده و برخی از مراحل را گام‌به‌گام پیش ببریم. در نهایت، نتایج به‌دست‌آمده را تحلیل و مقایسه خواهیم کرد. (با تشکر ویژه از مهندس علی صادقی)

۱.۲ دادگان

دادگان MaFaulDa یک مجموعه‌ای داده ارتعاشاتی [۱] است که برای پایش وضعیت [۲] و تشخیص عیب [۳] ماشین‌های دوار جمع‌آوری شده است. این دادگان شامل شرایط عملیاتی مختلفی از حالت نرمال تا عیوب مختلف برای ماشین دوار می‌باشد. از این دادگان در تحقیقات برای توسعه مدل‌های یادگیری ماشین با هدف تشخیص عیب استفاده می‌شود. به دلیل وجود داده ارتعاشاتی برای عیب‌های گوناگون، این دادگان برای آموزش مدل تشخیص عیب برای ماشین‌ها و ابزارآلات صنعتی بسیار مناسب هستند. درباره این دادگان تحقیق کرده و به پرسش‌های زیر پاسخ دهید: (لینک دادگان)

۱.۱.۲

سازوکار داده‌برداری این مجموعه داده را شرح دهید. برای داده برداری از چه سنسورهایی استفاده شده است؟

۲.۱.۲

کلاس‌های مختلف عیب را که در این دادگان جمع‌آوری شده‌است را معرفی کرده و توضیح دهید. چه مقدار داده برای هر بخش ثبت شده‌است؟

۳.۱.۲

تمامی کلاس‌های عیبی که در پایان‌نامه بررسی شده‌اند را درنظر گرفته و به تعداد دلخواه، فایل داده از این پیوند دنلود کنید. (دقت کنید که لازم نیست تمامی فایل‌های داده برای هر کلاس را استفاده کنید، صرفا یک یا دو فایل داده برای این مینی‌پروژه کفایت می‌کند.) سیگنال‌های موجود در یک فایل را همانند شکل ۳-۲ پایان‌نامه نمایش داده و هر بخش را با رنگ مجزا و لیبل مشخص کنید.

۲.۲ پیش‌پردازش و استخراج ویژگی

در این بخش سعی میکنیم قدم به قدم با پایان‌نامه پیش رفته و همراه با درک فعالیت‌های انجام شده روی داده، پیش‌پردازش‌ها و استخراج ویژگی اعمال شده را دوباره انجام دهیم.

۱.۲.۲

پیش‌پردازش داده یکی از اساسی‌ترین بخش‌های ایجاد یک مدل تخشیص عیب است. در این مرحله فعالیت‌هایی همانند حذف نویز یا نرمال سازی روی داده انجام می‌شود. با مطالعه بخش ۳-۲ پایان‌نامه، مراحل مختلف پیش‌پردازش داده را توضیح دهید. به نظر شما مرحله استخراج ویژگی از کدام بخش شروع شده است؟ مراحل استخراج ویژگی را کنار گذاشته و تمامی پیش‌پردازش‌های انجام شده را روی داده خود اعمال کنید.

نکته: پس از انتخاب پنجره زمانی، داده‌های خود را به بخش‌های آموزش [۴] و آزمون [۵] تقسیم کنید.

---

[۱] vibration
[۲] condition monitoring
[۳] fault diagnosis/detection
[۴] train
[۵] test

۲.۲.۲

ارتعاشات یک نوع داده سری زمانی است. روش‌های مختلفی برای تحلیل این مدل داده وجود دارد. با رجوع به بخش ۲-۳ توضیح دهید این کار تحقیقاتی از چه ویژگی‌هایی برای آموزش مدل خود استفاده کرده‌است. سپس با بهره‌گیری از مراحل استخراج ویژگی، که در بخش قبل مشخص کردید، ویژگی‌هایی همانند ویژگی‌های پایان‌نامه برای داده‌های خود استخراج کنید.

نکته: انجام مراحل "رتبه‌بندی ویژگی‌ها توسط LightGBM" و "انتخاب ویژگی به وسیله SI" به عنوان بخش امتیازی محسوب می‌شوند. ابتدا هر یک از این روش‌های را توضیح دهید و سپس به پیاده‌سازی آن‌ها بپردازید.

۳.۲ آموزش مدل

در این کار از MLP [۱] به عنوان مدل تشخیص عیب استفاده شده است. به دلیل این که هنوز مباحث تدریس شده این مطلب را پوشش نمی‌دهند شما می‌توانید از هر مدل طبقه‌بندی خطی‌ای استفاده نمایید. بنابراین یک مدل طبقه‌بندی خطی انتخاب کرده و در ادامه این بخش فقط از همان مدل استفاده کنید.

۱.۳.۲

یکی از ایده‌های استفاده شده در این پایان‌نامه طراحی یک مدل سلسله مراتبی است. درباره این روش تحقیق کنید. در این تحقیق چگونه از این روش استفاده شده است؟ نویسنده چه مزیت‌هایی برای این روش ذکر کرده است؟ آیا نتایج نهایی مؤثر بودن این روش را تأیید می‌کنند یا خیر؟ توضیح دهید.

۲.۳.۲

با استفاده از دو رویکرد مختلف برای داده خود مدل تشخیص عیب خود را ایجاد کنید.

- **رویکرد اول:** بدون استفاده از ساختار سلسله‌مراتبی
- **زویکرد دوم:** با استفاده از ساختار سلسله‌مراتبی استفاده شده در پایان‌نامه

نکته ۱: شما می‌توانید برای راحتی کار به جای پیاده‌سازی کل ساختار سلسله مراتبی، ساختار را به صورت زیر پیاده‌سازی کرده و نیازی به طبقه‌بندی تمامی کلاس‌های موجود در دادگان ندارید :

- normal
- imbalance
- misalignment
  - misalignment horizontal
  - misalignment vertical
- under-hang
  - ball
  - cage
  - race outer
- over-hang
  - ball
  - cage
  - race outer

[۱] multi-layer perceptron

نکته ۲

داده آموزش هر مدل باید متعادل [۱] باشند. یعنی باید تعداد یکسانی از هر کلاس در آموزش هر مدل استفاده شود. به این منظور می‌توانید از روش‌های upsampling ، downsampling و یا هر روش دلخواهی استفاده کنید.

۳.۳.۲

نتایج بدست‌آمده از هر دو رویکرد را با هم مقایسه کنید. به این نکته دقت داشته باشید که باید برای تمامی مدل‌های آموزش داده شده گزارش طبقه‌بندی [۲] ماتریس درهم‌ریختگی [۳] را نمایش دهید. (هم برای دادگان آموزش و هم برای دادگان آزمون)

۴.۳.۲

با در نظر گرفتن تمامی جنبه‌های رویکردهای پیش‌برده شده در بخش قبل، شما از کدام روش استفاده می‌کنید؟ دلیل خود را توضیح دهید.

۴.۲ محصول

برنامه‌ای بنویسید که یک نمونه را به عنوان ورودی دریافت کند و کلاس عیب را اعلام نماید. برنامه باید از هر مدلی در صورت نیاز استفاده کند؛ مثلا اگر داده ورودی از کلاس vertical misalignment باشد، مدل باید misaligned بودن آن را با استفاده از مدل اول تشخیص دهد و سپس مدل ثانویه مناسبی را انتخاب کرده تا تشخیص عیب را کامل نماید.

امتیازی

ساختار سلسله‌مراتبی را با دو مدل طبقه‌بندی خطی دیگر دوباره آموزش دهید و نتایج مدل‌های مختلف را به صورت کامل با هم مقایسه کنید.

امتیازی

یکی از روش‌های UMAP و یا TSNE را انتخاب کرده و درباره آن تحقیق کنید و به صورت خلاصه و با استفاده از روابط ریاضی لازم روش را توضیح دهید. سپس داده‌هایی که در انتها بخش ۲.۲ ایجاد کردید را به صورت دوبعدی و سه‌بعدی نمایش دهید.

---

[۱] balanced

[۲] classification report

[۳] confusion matrix

در انجام این مینی‌پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این تمرین، ساعت ۱۸:۰۰ روز پنجشنبه ۱۴ فروردین‌ماه ۱۴۰۴ است.
- برای گزارش لازم است که پاسخ هر سوال و زیربخش‌هایش به‌ترتیب و به‌صورت مشخص نوشته شده باشند. بخش زیادی از نمره به توضیحات دقیق و تحلیل‌های کافی شما روی نتایج بستگی خواهد داشت.
- لازم است که در صفحهٔ اول گزارش خود لینک مخزن گیت‌هاب را و گوگل‌کولب مربوط به مینی‌پروژهٔ خود را درج کنید. درخصوص گیت‌هاب، یک مخزن خصوصی درست کنید و آی‌دی‌های `MJAHMADEE` و `AliBagheriNejad` را به‌عنوان Collaborator به مخزن اضافه کنید. پروژه‌های گیت‌هاب می‌بایست در انتهای ترم پابلیک شوند. درمقابل، لینک گوگل‌کولب را در حالتی که دسترسی عمومی دارد به اشتراک بگذارید. دفترچه‌کد گوگل‌کولب باید به‌صورت منظم و با بخش‌بندی مشخص تنظیم شده باشد و خروجی سلول‌های اجراشده قابل مشاهده باشد. در گیت‌هاب نیز یک مخزن برای درس و یک پوشهٔ مجزا برای هر مینی‌پروژه ایجاد کنید.
(آموزش پرایوت‌کردن مخزن گیتهاب و آموزش افزودن Collaborator به مخزن گیت‌هاب)
- هرجا از دفترچه‌کد گوگل‌کولب شما نیاز به فراخوانی فایلی خارج از محیط داشت، مطابق آموزش‌های ارائه‌شده ملزم هستید از دستور gdown استفاده کنید و مسیرهای فایل‌ها را طوری تنظیم کنید که صرفاً با اجرای سلول‌های کد، امکان فراخوانی و خواندن فایل‌ها توسط هر کاربری وجود داشته باشد.
- در تمامی مراحل تعریف داده و مدل و هرجای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقم آخر شمارهٔ دانشجویی خود در نظر بگیرید.
- استفاده از ابزارهای هوشمند (مانند ChatGPT) در کمک‌گرفتن برای بهبود کدها مجاز است؛ اما لازم است تمام جزئیات مواردی که در خروجی‌های مختلف گزارش خود عنوان می‌کنید را به خوبی خوانده، درک و تحلیل کرده باشید. استفاده از این ابزارهای هوشمند در نوشتن گزارش و تحلیل‌ها ممنوع است.
- در جاهایی که با توجه به دو رقم آخر شمارهٔ دانشجویی خود محدود به انتخاب عدد، متغیر و یا داده‌ای خاص شده‌اید، برای تست‌های اضافه‌تر و نمایش بهبود در نتایج خود، مجاز هستید از مقادیر دیگر هم استفاده کنید.
- رعایت نکات بالا به حرفه‌ای‌ترشدن شما کمک خواهد کرد و اهمیتی معادل مطالب درسی فراگرفته‌شده دارد؛ بنابراین، درصورت عدم رعایت هریک از این نکات، از نمره تمرین شما **کسر خواهد شد**.
- آی‌دی پرسش هرگونه سوال درخصوص مینی‌پروژهٔ شمارهٔ 1