به نام خداوند بخشنده مهربان

طراحی سیستم‌های یادگیری ماشین

دانشکده مهندسی کامپیوتر
علی زارع‌زاده
بهار ۱۴۰۲

تمرین اول | **کار با داده** | تاریخ انتشار: ۲۰ اسفند ۱۴۰۱

---

۱. پرسش‌های خود در مورد این تمرین را در سامانه کوئرا مطرح کنید.

۲. **سیاست ارسال با تاخیر پاسخ:** به مدت ۵ روز پس از ددلاین اولیه، می‌توانید پاسخ خود را با تاخیر ارسال نمایید. در مجموع در طول ترم ۱۵ روز تاخیر مجاز دارید. به ازای هر ساعت تاخیر پس از اتمام ۱۵ روز مجاز نیم درصد از نمره شما کسر خواهد شد.

۳. **سیاست مشارکت دانشجویان در حل تمارین:** رعایت آداب‌نامه‌ی انجام تمرین‌های درسی در انجام تمارین الزامی است. در صورت مشاهده تخلف مطابق قوانین دانشکده و دانشگاه برخورد خواهد شد. در بار اول منفی صد درصد نمره تمرین لحاظ خواهد شد و در مرحله دوم ضمن اعلام به کمیته آموزشی دانشکده و انضباطی دانشگاه، نمره مردودی برای درس ثبت خواهد شد.

پرسش‌ها (۱۰۰ نمره) | تاریخ تحویل: ۲۰ فروردین ۱۴۰۲

مقدمه

یک مساله معروف در حوزه یادگیری ماشین مساله دسته‌بندی[۱] است که کاربرد بسیار فراوانی در زندگی امروزی دارد. یکی از کاربردهای مساله دسته‌بندی، بازیابی راحت‌تر از بانک اطلاعاتی بزرگ با استفاده از پرسمان بر روی دسته‌بندی است. به عنوان مثال در یک وب‌سایت فروشگاه اینترنتی، دسته‌بندی یک ویژگی الزامی خواهد بود تا که کاربران بتوانند به راحتی و با تجربه کاربری مناسب به کالاهای مورد نظر خود دسترسی پیدا کنند.

در این تمرین ما می‌خواهیم بر روی کالاهایی که بر روی وب‌سایت دیوار وجود دارند، مساله دسته‌بندی را انجام بدهیم. وب‌سایت دیوار یک وب‌سایت ایرانی است که مانند آگهی در روزنامه نیازمندی‌ها عمل می‌کند. کاربری که چیزی برای فروش دارد می‌تواند برای کالای خود اگهی ایجاد کند. در حین ایجاد آگهی، از کاربر خواسته می‌شود که دسته‌بندی مناسب برای کالای خود را وارد کند. در این‌جا دو مساله می‌تواند ایجاد شود. اولی نبود مقدار[۲] است. به این معنی که ممکن است کاربر مقداری برای دسته‌بندی مورد نظر خود وارد نکند. مورد دوم عدم صحت دسته‌بندی وارد شده است. به این معنی که ممکن است کاربر دسته‌بندی خود را به صورت دقیق وارد نکرده باشد و نیاز به تصحیح آن باشد. لذا وجود سیستمی که بتواند به اصلاح این دسته‌بندی بپردازد الزامی است.

هدف این تمرین این است که به ابعاد داده‌ای این مساله بپردازیم و تکنیک‌های مختلف برای اعمال تغییر بر داده، از این جهت که در نهایت نتیجه بهتری حاصل آید را بررسی کنیم. در ادامه در هر بخش خواسته‌هایی بر روی داده مطرح شده است که شما باید کد مربوط به آنها را پیاده‌سازی نموده و در گزارش خود به بیان توضیحات در مورد نتایج بپردازید.

آماده‌سازی داده

در این بخش می‌خواهیم به آماده‌سازی داده مورد استفاده بپردازیم. همان طور که می‌دانید، دادگان و رویکرد ما نسبت به آنها در یک مساله به مرور زمان دستخوش تغییر می‌شوند. بنابرین الزامی است که ما این تغییرات را به صورت مستمر رهگیری کنیم و سابقه آن را به صورت منسجم ذخیره‌سازی کنیم. لذا ابزارهایی جهت کنترل نسخه[۳] داده توسعه داده شده اند. یکی از این ابزارها DVC است. تمام مراحل این تمرین را توسط DVC ورژن‌گذاری کنید تا در نهایت سابقه‌ای از فرآیند توسعه شما وجود داشته باشد.

برای بخش آماده‌سازی داده شما باید ETL انجام دهید. یعنی در ابتدا باید داده را از طریق خزش[۴] در وب‌سایت دیوار دانلود کنید. پس از آن باید داده دریافت شده را به یک ساختار جدولی تبدیل کنید و در نهایت این داده را در یک پایگاه داده SQL ذخیره کنید. پیشنهاد ما برای پایگاه داده PostgreSQL است که هم متن‌باز است هم از توان پردازشی بالایی بهره می‌برد[۵] داده جمع‌آوری شده توسط شما باید شامل داده‌های مرتبط به هر آگهی در این وب‌سایت باشد. توصیه می‌شود که آگهی‌ها را از دسته‌های مختلف جمع‌آوری کنید تا پراکندگی در داده وجود داشته باشد. در این لینک می‌توانید نمونه داده جمع‌آوری شده را مشاهده کنید.

---

[۱] Classification
[۲] Value Missing
[۳] Version Control
[۴] Crawl
[۵] توصیه می‌شود که پایگاه داده PostgreSQL را از طریق Docker راه‌اندازی کنید.

تحلیل اکتشافی داده

یک مرحله مهم پیش از اجرای پروژه یادگیری ماشین تحلیل اکتشافی داده است. در این مرحله باید ویژگی‌های مختلف داده، روابط آنها با یک دیگر، صحت آنها از طریق محاسبه آماره‌ها و رسم نمودارها بررسی شود. شما پس از انجام این مرحله، باید نسبت به پاک‌سازی داده نیز اقدام فرمایید.

مهندسی ویژگی

مهندسی ویژگی یک موضوع مهم در مسائل یادگیری ماشین است. دادگانی که با آنها کار می‌کنیم معمولا از فضای حل مساله فاصله زیادی دارند. مدل‌های یادگیری ماشین سعی بر این می‌کنند تا فضای ورودی را به فضای خروجی تبدیل کنند لکن اینکه بخشی از این تبدیل توسط انسان و قواعد مشخص شده توسط انسان با توجه به مساله انجام شود، می‌تواند به افزایش دقت منتهی شود. همچنین روش‌های دیگر در کاهش ابعاد که بعضا به عنوان مهندسی ویژگی استفاده می‌شوند، می‌توانند به افزایش دقت مدل کمک کنند. در این بخش شما باید مهندسی مناسب بر روی ویژگی را انجام داده و در نهایت داده را پاکسازی کنید.

۱. در بخش اول شما باید به صورت انسانی ویژگی‌های مناسب را انتخاب کنید. می‌توانید در این بخش ویژگی‌های ترکیبی ایجاد نمایید.

۲. در مرحله بعد شما باید از روش‌های کاهش ابعاد مانند PCA، TSVD، TNSE و مواردی از این دست استفاده کنید.

۳. یک کارکرد شبکه‌های عصبی که موجب موفقیت چشم‌گیر آنها شده است، توانایی آنها در استخراج ویژگی است. در شبکه‌های عصبی ویژگی‌های سطح بالاتر در لایه‌های عمیق‌تر لایه به لایه استخراج می‌شوند. برای این منظور شما باید یک خودرمزنگار پیاده‌سازی کنید و به وسیله آن از طریق بازنمایی تولید شده در گلوگاه کاهش ابعاد را انجام دهید.

لطفا توجه کنید که خروجی مدل خود را باید بر روی تمام حالت‌ها تست کنید. یعنی سه حالت فوق و حالت عدم استفاده از مهندسی ویژگی را خروجی دریافت نموده و آنها را تحلیل کنید.

کدگذاری داده

همان‌طور که می‌دانید، برای استفاده از داده‌های غیرعددی، نیاز به تبدیل آنها به یک نمایش عددی_ برداری وجود دارد. برای بازنمایی داده‌های متنی روش‌های بسیاری وجود دارد. این مدل‌ها که برای تولید بازنمایی متنی استفاده می‌شوند، از جهات مختلفی با یک‌دیگر تفاوت دارند. به عنوان مثال برخی از آنها به بعد معنایی و برخی دیگر به بعد ظاهری حساس‌تر هستند.
برای این بخش از حداقل سه مدل برای تولید بازنمایی استفاده کنید. دو مورد از آنها باید روش‌های TF-IDF و Pre-trained BERT باشند. برای مورد سوم می‌توانید خودتان انتخاب کنید.

ناهمگنی توزیع داده

یک مشکل شایع در این حوزه ناهمگنی دادگان این است که تعداد داده‌های در دسترس در طبقه‌های مختلف متفاوت باشد. این مشکل می‌تواند باعث عدم آموزش درست مدل شود به نحوی که مدل به سمت کلاس اکثریت جهت‌گیری می‌کند. در این بخش می‌خواهیم با استفاده از سه روش بر این مشکل فائق آییم.

- راه کار اول کاهش داده است. در زمان نمونه‌برداری از داده‌ها به نحوی نمونه‌برداری را انجام می‌دهیم که تعداد کلاس‌های مختلف در آن داده از توزیع یکنواخت باشد.
- روش دیگر برای این مشکل تغییر تابع خطا در آموزش مدل است. در این حالت مدلی را باید آموزش دهید و تابع خطای آن را به نحوی تغییر دهید که نسبت به ناهنگنی داده مقاوم باشد.
- در نهایت شما باید با استفاده از راهکار افزایش داده[۶] نسبت به تولید داده جدید بپردازید و با استفاده از داده جدید تولید شده، توزیع داده بر روی کلاس‌های مختلف را یکنواخت کنید.

حالت عدم استفاده از هیچ‌یک از راهکارهای فوق را نیز بررسی کنید.

نکته مهم

توجه کنید که در نهایت شما باید تمام ترکیب‌های مختلف از حالت‌هایی که در بخش‌های بالا معرفی شد را اجرا کنید و نتایج آنها را گزارش و تحلیل کنید.

[۶] Data Augmentation