هوش مصنوعی **تمرین کامپیوتری شماره‌ی ۴**

مدرس: **دکتر فدایی و دکتر یعقوب‌زاده** طراح: سید پارسا حسینی‌نژاد، آتیه آرمین، سروین بهمنی

مهلت تحویل: یکشنبه ۸ خرداد ۱۴۰۱، ساعت ۲۳:۵۹

هدف پروژه

هدف این پروژه آشنایی با روش‌های یادگیری ماشین با استفاده از کتابخانه Scikit-Learn است. این پروژه در چهار فاز تعریف شده است. در فاز صفر به بررسی مجموعه داده‌ها و تجزیه و تحلیل داده‌های اکتشافی می‌پردازید. در فاز اول با پیش‌پردازش آشنا خواهید شد. در فاز دوم با استفاده از چند Classifier تعریف شده در کتابخانه Scikit-Learn مدل‌هایی را پیاده‌سازی و بهینه‌سازی خواهیدکرد. نهایتا در فاز سوم با استفاده از مدل‌های بهینه فاز دوم، به پیاده‌سازی چند روش یادگیری گروهی و تحلیل نتایج حاصل می‌پردازید.

معرفی مجموعه داده

مجموعه داده‌ای که در اختیار شما قرار دارد شامل اطلاعات تعدادی قطعه‌ی موسیقی است که توسط Spotify گردآوری شده است. در این مجموعه داده هر موسیقی شامل ویژگی‌هایی نظیر طول موسیقی، میزان محبوبیت، انرژی و ... است. داده‌ی هدف در این پروژه، ژانر قطعه‌ی موسیقی است.

فاز صفر : EDA[1] and Visualization

اولین گام در هر پروژه یادگیری ماشین، مشاهده، شناخت و بررسی داده‌ها و ارتباط میان آنهاست. به این منظور قدم‌های زیر را انجام دهید و در هر مرحله، نتایج بدست آمده را در گزارشتان ذکر کنید.

1. با استفاده از متدهای describe و info از کتابخانه pandas، ساختار کلی داده‌ها را بررسی کنید.
2. درصد داده‌های از دست رفته هر ویژگی را پیدا کنید و نمایش دهید.
3. نمودار توزیع ویژگی‌های عددی و غیر عددی را رسم کنید (برای آشنایی می‌توانید از این لینک کمک بگیرید). ویژگی‌های عددی از چه توزیعی پیروی می‌کنند؟

[1] Exploratory Data Analysis

فاز اول : Preprocessing

عملیات پیش‌پردازش داده‌ها مرحله مهمی در هر پروژه یادگیری ماشین است. در این فاز شما باید داده‌های خام ورودی را به مجموعه‌ای از داده‌های قابل پردازش تبدیل کنید. به این منظور قدم های زیر را دنبال کنید و در گزارش خود توضیحات مربوط به هر مرحله را ذکر کنید.

1. برای رفع مشکل داده‌های گمشده، روش‌های زیادی وجود دارد؛ از جمله حذف کل ستون یا پر کردن مقادیر گمشده با آماره‌ها (مثلا میانگین). روش‌های موجود برای مدیریت داده‌های گمشده را مختصرا توضیح دهید و مزایا و معایب هر روش را به طور مختصر شرح دهید.
2. برای هر ستون با مقادیر گمشده، یکی از روش‌های مدیریت داده‌های گمشده را انتخاب کرده و آن را اعمال کنید.
3. برای ویژگی‌های عددی، Normalization یا Standardization به چه منظور استفاده می‌شود؟
4. با توجه به سوال ۳ فاز صفر، توضیح دهید که در این پروژه از چه روشی برای اسکیل داده‌ها استفاده می‌کنید؟ چرا؟
5. برای این که مدل ما بتواند با داده‌های دسته‌ای[2] کار کند، روش‌های زیادی وجود دارد. دو روش را توضیح دهید و بیان کنید از کدام روش استفاده کردید. چرا؟ (می‌توانید از این دو لینک کمک بگیرید؛ لینک1 و لینک2).
6. با توجه به این که نام خواننده به نظر ویژگی مفیدی در تشخیص ژانر موسیقی است (چون معمولا هر خواننده در یک ژانر مخصوص فعالیت می‌کند)، آیا می‌توان آن را به گونه‌ای تغییر داد تا قابل استفاده باشد؟ اگر نه، آیا راهی جز حذف کردن این ستون وجود دارد؟
7. برای ویژگی‌ها information gain را محاسبه کنید (برای محاسبه information gain می‌توانید از متد mutual_info_classif از کتابخانه Scikit-Learn استفاده کنید). سپس نمودار gain بر حسب ویژگی‌ها را رسم کنید.
8. بنظر شما آیا تمامی ستون‌ها به ما اطلاعات مفیدی در جهت شناخت ژانر موسیقی می‌دهند؟ نگه داشتن همه ویژگی‌ها چه مزایا و معایبی دارد؟ آیا می‌توانیم ستونی از مجموعه داده را حذف کنیم؟ حذف کردن یک یا چند ستون چه مزایا و معایبی دارد؟ به نظرتان کدام ویژگی‌ها در پیش‌بینی ژانر قطعه موسیقی می‌توانند مفید باشند؟ کدام ستون‌ها را می‌توانیم حذف کنیم؟ برای پاسخ به این سوال صرفا نظرتان را با توجه به نتایج بدست آمده برای سوالات قبلی، بیان کنید.

[2] Categorical Features

فاز دوم : Model Training, Evaluation and Hyper Parameter Tuning

در این فاز از پروژه دو مدل فردی بر پایه K-Nearest-Neighbours و Decision Tree به کمک کتابخانه SciKit-Learn پیاده‌سازی می‌کنید. نهایتا در این بخش باید مدل‌های بهینه‌ای از این Classifier ها داشته باشید. بهینه به این معناست که هایپر پارامترها را به گونه‌ای تنظیم کنید که هر مدل به بیشترین دقت برسد در حالی که overfitting اتفاق نیفتد (برای آشنایی با overfitting از این لینک کمک بگیرید).

1. اولین قدم تقسیم داده‌ها به دو دسته train و test است. یک روش این است که P درصد اول داده‌ها را برای train و مابقی را برای test در نظر بگیریم. شما چه عددی را برای P انتخاب می‌کنید؟ آیا نیاز است تقسیم داده‌ها به صورت تصادفی باشد؟ چرا؟ آیا لازم است نسبت تعداد داده‌های مربوط به هر ژانر موسیقی به کل داده در داده‌ی آموزش و تست برابر باشد؟ در مورد پارامتر stratify در تابع train_test_split در کتابخانه SciKit-Learn تحقیق کنید.
2. با توجه به پاسختان به سوال قبل، داده‌ها را به دو دسته train و test تقسیم کنید.
3. برای مدل KNN نمودار دقت مدل (برای داده‌های test و train) را بر حسب هایپر پارامتر تعداد همسایه‌ها (n_neighbor) رسم کنید و overfitting را بر روی این نمودار بررسی و تحلیل کنید. بهتر است نمودارهای مربوط به train و test را در یک plot رسم کنید (از این لینک کمک بگیرید).
4. برای مدل Decision Tree پارامترهای max_depth و min_samples_leaf را تنظیم کنید. سپس، مانند سوال قبل نمودار دقت را بر حسب هایپر پارامترها رسم نمایید.
5. در مورد underfitting و overfitting تحقیق کنید و بررسی کنید آیا در مدل‌های شما underfitting یا overfitting اتفاق افتاده است؟
6. معیارهای Recall، Precision، Accuracy و F1 Score را توضیح دهید و دقت هر مدل را بر اساس این معیارها برای داده‌های test و train اندازه‌گیری کنید. دقت[3] مطلوب برای هر دو مدل حداقل ۶۰٪ است.
7. تاثیر پیش‌پردازش‌هایی که روی داده‌ها انجام دادید را به طور کامل بررسی کنید (مثلا تاثیر روش‌های مختلف مدیریت مقادیر گمشده روی معیارهای نهایی).

[3] Accuracy

فاز سوم : Ensemble Methods

یادگیری گروهی به این معناست که از تجمیع نتایج حاصل از تعدادی مدل، پیش‌بینی نهایی را انجام دهیم. در این فاز به پیاده‌سازی و تحلیل نتایج مدل Random Forest می‌پردازیم.
در مدل Random Forest تعدادی Decision Tree با ویژگی‌ها و داده‌های مختلف را در کنار هم قرار می‌گیرند و هر کدام از این درخت‌ها یادگیری را جداگانه انجام می‌دهند. خروجی جنگل تصادفی کلاسی است که توسط اکثر درختان انتخاب شده است.

1. با کمک کتابخانه Scikit-Learn این مدل را پیاده‌سازی کنید.
2. تاثیر هایپرپارامترهای max_depth، n_estimators و min_samples_leaf را بر این مدل بررسی کرده و هر کدام را توضیح دهید (از این لینک کمک بگیرید).
3. با استفاده از معیارهای Recall، Precision، Accuracy و F1 Score دقت مدل خود را بسنجید. نتیجه مطلوب در این فاز دستیابی به دقت حداقل ۷۰٪ است.
4. پس از پیدا کردن هایپر پارامترهای بهینه، Confusion Matrix مدل خود را نشان دهید.

نکات پایانی

- دقت کنید **قسمت اعظمی از نمره‌ی شما در این پروژه مربوط به گزارش شماست.** پس حتما گزارش خود را کامل بنویسید و موارد خواسته شده را در گزارش خود لحاظ کنید.
- هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده بطور مثال نمودارها استفاده کنید.
- در همه بخش‌ها مجاز هستید از متدهای کتابخانه Scikit-Learn استفاده کنید.
- نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA4_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی های خواسته شده بخشی از نمره این تمرین را تشکیل می‌دهد. از نمایش درست خروجی های مورد نیاز در فایل html مطمئن شوید.
- هیچگونه شباهتی در انجام این پروژه بین افراد مختلف پذیرفته نمیشود. در صورت کشف هرگونه تقلب برای همه افراد متقلب نمره ۱۰۰- در نظر گرفته میشود.
- استفاده از مراجع با ارجاع به آنها بلامانع است. اما در صورتی که گزارش شما ترجمه عینی از آنها باشد یا از گزارش افراد دیگر استفاده کرده باشید کار شما تقلب محسوب میشود.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم یا گروه تلگرامی درس مطرح کنید تا بقیه از آن استفاده کنند، در غیر این صورت می‌توانید به طراحان پروژه ایمیل بزنید و سوالتان را از یکی از آنها بپرسید. ایمیل طراحان نیز در ابتدای تمرین مشخص شده‌است.

موفق باشید!