به نام خداوند جان و خرد

نیمسال اول 1402-1401

Machine Learning
Final Exam

استاد درس: دکتر منصور رزقی آهق

دانشجو: بهار مهدوی
40152521337

Tarbiat Modares University

پرسش:

داده های لینک زیر را در نظر بگیرید
https://www.kaggle.com/datasets/wanghaohan/confused-eeg

این مسله را با روش‌های کاهش ابعاد و کلاس بندی مطرح شده در کلاس حل نمایید. داده‌ها را به 80 درصد ترین و 20 درصد تست با کی فلد 5 حل کنید:

فهم توضیحات و هدف مسئله جزو وظایف دانشجو است البته توضیحات کمی پایین داده شده است.
نکته 1: از ابزار شبکه های عمیق استفاده نشود
نکته 2: از کدهای این چالش کگل استفاده نگردد

راهنمایی:
این داده ها برای افراد مختلف ویدوهای مختلف را نشـــان داده اند و در طول زمان ویژگی‌های مختلفی از آن‌ها را ثبت نموده‌اند: لذا برای 10 نفر اگر 10 تا ویدیو نشان داده باشند کلا 100 تا داده است. که هر داده شامل سیگنال‌های ثبت شده در زمان‌های مختلف است. هدف یک مسله دو کلاسه است اینکه فرد از دیدن ویدیو تحت تاثیر قرار گرفته یا نه

تحویل: کد به همراه یک گزارش از مشخصات داده + همراه تحلیل نتایج

مشخصات داده:
داده‌های سیگنال EEG از 10 دانشجو در حین تماشای 10 کلیپ ویدئویی با اندازه‌گیری سیگنال‌هایی از فعالیت لوپ پیشانی این دانشجویان جمع آوری شده است. بنابراین، 100 نقطه داده برای 12811 ردیف مشاهده می‌گردد که به طور متوسط حدود 1200 ردیف داده، به ازای هر دانشجو و در حدود 120 ردیف داده، از هر ویدیوی 1 دقیقه‌ای مشاهده شده توسط هر دانشجو (که در بازه های زمانی 0.5 ثانیه 0.5 ثانیه نمونه برداری شده) بدست آمده است.

داده های ویدیویی: هر ویدیو تقریباً دو دقیقه طول می کشـد، که در ایجاد این دیتا فریم 30 ثانیه اول و 30 ثانیه آخر حذف شـده و فقط داده های EEG، در 1 دقیقه وسط جمع آوری شده است.

EEG_data.csv: حاوی داده های EEG ثبت شـــده از 10 دانشـــجو (SubjectID) در هنگام مشـــاهده 10 ویدیو (VideoID) اســـت (بنابراین ما 2 لیبل SubjectID و VideoID برای نمونه های سمپل‌مان داریم). با اندازه‌گیری سیگنال‌های فعالیت لوپ پیشانی دانشجویان ویژگی‌های (Attention، Mediation، Raw، Delta، Theta، Alpha1، Alpha2، Beta1، Beta2، Gamma1، Gamma2) بدست آمده‌اند. (تحقیقات گذشته نشان داده است که سیگنال Theta با سطح سردرگمی در ارتباط است.)

دانشجویان پس از مشاهده هر ویدیو سطح سردرگمی خود را در مقیاس 1-7 رتبه بندی کردند، که در ستون ویژگی "Attention " 1 با کمترین سطح گیج کنندگی و 7 با گیج کننده ترین سطح مطابقت دارد. (سیگنال‌های با فرکانس بالاتر به عنوان مقدار میانگین در هر 0.5 ثانیه گزارش می‌شوند.)

اگر این ویژگی‌ها به دو کلاس، نرمالایز گردد که آیا دانشجو با دیدن ویدیو سردرگم شده است یا خیر، در ستون‌های آخر 2 لیبل تحت عنوان "predefinedlabel" و "user-definedlabeln" خواهیم داشت که اولی لیبل از پیش تعریف شده‌ی سردرگمی توسط محققان آزمایش است و دومی که متغیر هدف ما می‌باشد لیبلی است که خود دانشجو از سطح سردرگمی خود ثبت کرده و نشان دهنده‌ی این است که فریم 0.5 ثانیه‌ای ویدیو برای آن دانشجو سردرگم کننده بوده یا خیر

demographic.csv: حاوی اطلاعات جمعیت شناختی برای هر دانشجو شامل (age، ethnicity، gender) است.

پس از ترکیب دو دیتا فریم EEG_data و demographic، مشخصات داده ای به شرح جدول زیر است:

Number	Features (Column)	Data Type	Non-Null Count	Missing Values
1	SubjectID	Integers	12811 non-null	0
2	VideoID	Float	12811 non-null	0
3	Attention	Float	12811 non-null	0
4	Mediation	Float	12811 non-null	0
5	Raw	Float	12811 non-null	0
6	Delta	Float	12811 non-null	0
7	Theta	Float	12811 non-null	0
8	Alpha1	Float	12811 non-null	0
9	Alpha2	Float	12811 non-null	0
10	Beta1	Float	12811 non-null	0
11	Beta2	Float	12811 non-null	0
12	Gamma1	Float	12811 non-null	0
13	Gamma2	Float	12811 non-null	0
14	predefinedlabel	Float	12811 non-null	0
15	user-definedlabeln	Float	12811 non-null	0
16	age	Integers	12811 non-null	0
17	ethnicity	Object	12811 non-null	0
18	gender	Object	12811 non-null	0

مراحل کار و تحلیل نتایج:

1. دو دیتا فریم df1= EEG_data و df2= demographic فراخوانی و با هم ترکیب شـــدند (df). برای این کار لازم بود "subject ID" از دیتاست demographic با "subjectID" از دیتاست EEG_data هم نام گردند.

2. داده ها طی مراحلی پیش پردازش شدند. ابتدا تعداد null ها و حضور یا عدم حضور Missing value ها بررسی گردید که صفر بودند و نیازی به حذف یا جایگذاری آن‌ها وجود نداشت.

   ویژگی‌های age، ethnicity و gender اسپیس اضافی در شروع حرف خود داشتند که باعث ایجاد خطا در ادامه مراحل می‌شد به این ترتیب با عنوانی دیگر تغییر نام داده شدند.

   داده های ویژگی Age و Ethnicity که non-numerical بودند به numerical تبدیل گردید.

   ویژگی‌های SubjectID، VideoID و predefinedlabel که به کار مدل ما نمی‌آمدند حذف شدند.

3. ویژگی user-definedlabeln به عنوان متغیر هدف (y) تعیین گردید و مابقی ویژگی‌ها به عنوان (X) در نظر گرفته شد.

4. داده ها طی یک مرحله استانداردسازی، نرمال و هم اسکیل شدند.

5. برای کاهش بعد داده های هم اسـکیل شـده‌ی مرحله قبل از تکنیک PCA اسـتفاده گردید. 7 ویژگی اول از دید من (که ممکنه دید صحیحی نباشد) مناسب به نظر آمد به این ترتیب تعداد پارامتر component آن 7 در نظر گرفته شد. ratio variance explained

6. داده ها از طریق روش 5-Fold Cross Validation به 80 درصد داده آموزش و 20 درصد داده تست تقسیم بندی شدند.

7. برای کلاسه بندی داده‌ها ابتدا از روش logistic regression استفاده گردید و Accuracy 0.585 بدست آمد که دقت خوبی به حساب نمی‌آید.

8. در مرحله‌ی بعد از روش درخت تصمیم برای کلاسه بندی داده‌ها بهره گرفته شد. همچنین در این بخش به مقایسه روش تقسیم داده‌ها به روش train_test_split، 5-Fold Cross Validation و 10-Fold Cross Validation پرداخته شد و به ترتیب Accuracy های 0.579، 0.551 و 0.553 بدست اومد که به نسبت روش train_test_split از مابقی نتیجه بهتری داشت ولی در کل دقت خوبی بدست نیامد.

9. روش 5-Fold Cross Validation با logistic regression 0.585 Accuracy نشاندهنده‌ی این است که برای کلاسه بندی این داده ها مدل مناسبتری نسبت به مدل decision tree با 0.55 Accuracy به حساب می‌آید.

10. در مرحله بعد از روش SVM برای کلاسه بندی داده ها استفاده شد و 0.562 Accuracy بدست آمد که کمی از مدل درخت تصمیم بهتر ولی همچنان از روش logistic regression ضعیف‌تر عمل کرده است.

کد:

- Final Exam - Bahar Mahdavi.ipynb