پروژه درس یادگیری ماشین

دکتر محمدزاده

علیرضا سخایی راد - ۹۸۱۰۱۷۱۴

۴ بهمن ۱۴۰۰

فهرست مطالب

۱ مقدمه

یکی از راه‌های بررسی فرآیندهای رخ داده در مغز، استفاده کردن از سیگنال‌های EEG است. این سیگنال‌ها توسط الکترودهای مخصوصی بر روی سر اندازه‌گیری می‌شوند و سپس با استفاده از ویژگی‌های زمانی، فرکانسی و زمان‌فرکانسی آنها تحلیل انجام می‌گردد. در این پروژه تلاش شده با به کمک مدل‌های یادگیرنده ساده، پردازشی از این دست سیگنال‌ها انجام دهیم.

نکته اینکه در این گزارش به توضیح تمامی توابع نپرداخته شده و بخش زیادی از اطلاعات ریز به صورت کامنت و تکست در خود فایل نوتبوک موجود هستند.

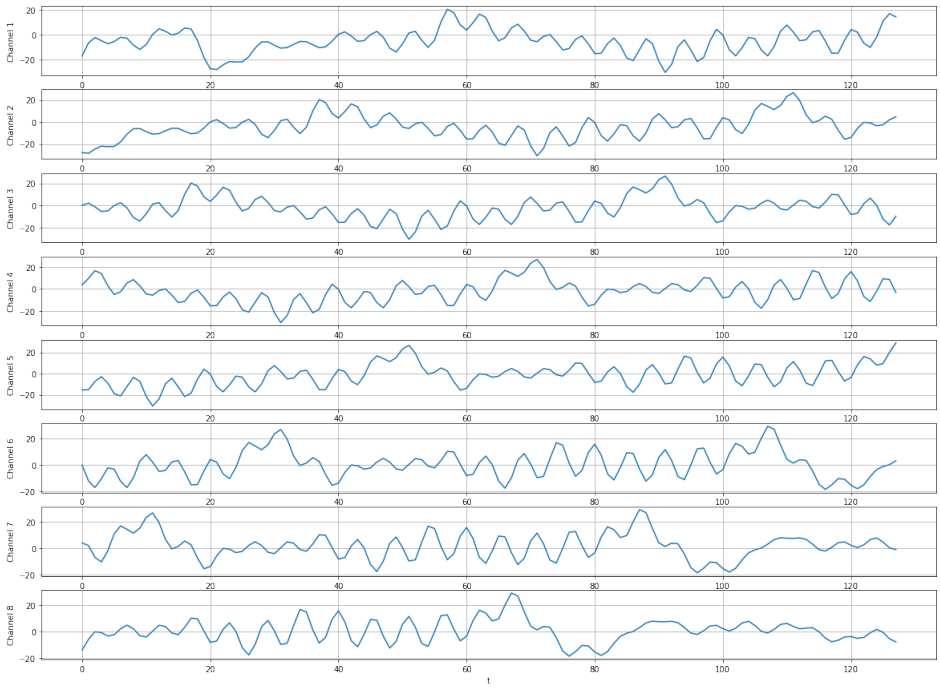
۲ نگاهی به دیتا

دیتا به صورت کانال های ۸ تایی داده شده است که برای هر شخص ۸ کانال برای ترین و ۸ کانال برای تست داریم.

نمونه برداری به دو شکل مختلف تک‌کاراکتر و سطرستونی انجام شده است. در حالت تک‌کاراکتر هر حرف ۱۵ بار روشن شده است و این فرآیند ۵ بار به ازای هر حرف کلمه LUKAS انجام شده است. با جداسازی سیگنال‌ها و لیبل‌های این زمان‌ها، نهایتا ۲۷۰۰ دیتای ۸ بعدی خواهیم داشت که فقط $\frac{1}{35}$ آنها تارگت هستند. در بخش چالش‌ها راجع به غلبه به این مشکل توضیح کافی داده شده است.

ذر افرادی که به صورت سطرستونی نمونه برداری شده‌اند، مجموع سطرها و ستون‌ها (۱۲ تا) هر کدام ۱۵ بار و ۵ بار به ازای هر حرف روشن شده‌اند که ۹۰۰ دیتای ۸ بعدی را تشکیل خواهد داد. به ازای هر روشن شدن، تعداد مشخصی سمپل از هر کانال را استفاده خواهیم کرد.

نمونه‌ای از این دیتا در ۸ کانال برای نفر اول به شکل زیر است:



شکل ۱: نمونه دیتا

درباره نحوه ذخیره این داده‌ها در بخش کلاس Person توضیح داده خواهد شد.

Person ۳

بخش اصلی کار در اینجا انجام خواهد شد. به جز قسمت فیت کردن مدل، اینجا این کلاس و نحوه استخراج دیتا را شرح خواهیم داد و توضیح بعضی توابع نیز به بخش‌های مخصوص موکول می‌گردد.

این کلاس درحقیقت نماینده تمامی افراد خواهد بود و هر فرد یک شی از این کلاس به شمار خواهد آمد. جهت جلوگیری از شلوغی گزارش، از آوردن کدها خودداری شده اند و دستورات در فایل نوتبوک ارسالی در دسترس می‌باشند.

در هنگام ساخته شدن یک شی از این کلاس، دیتاهای ترین و تست به آن داده می‌شود و در تابع init مقادیر مخصوص هر شخص مثل نحوه نمونه برداری (تک‌کاراکتر یا سطرستونی)، تعداد سمپل از سیگنال‌ها به ازای هر روشن شدن و ... ذخیره و ست می‌شوند. سپس تابع استخراج دیتا صدا زده می‌شود. این تابع با بررسی کانال دهم که نشان دهنده چراغ‌های روشن در هر دو حالت است، به استخراج دیتا می‌پردازد.

هرکجا که این سیگنال از صفر به مقدار بالاتری می‌رود، نشاندهنده یک تحریک است. این زمان‌ها پیدا شده‌اند و سیگنال‌های نظیر (از سطرهای ۲ تا ۹ دیتاست) و لیبل (سطر ۱۱ برای دادگان ترین) استخراج می‌شوند و در متغیرهای مناسبی ذخیره می‌گردند.

۴ تعیین هایپرپارامترها

۱.۴ ویژگی ها و نوع طبقه بند

کلاسی به اسم Feature برای این کار نوشته شده‌است. با صدا زدن تابع مناسب از کلاس Person یک شی از کلاس Feature ساخته می‌شود که شخص را به عنوان یک ویژگی دریافت می‌کند.

این کلاس، شامل تعداد زیادی تابع برای استخراج ویژگی و همچنین تعدادی تابع کمکی است. این توابع با گرفتن دیتا، ویژگی‌های مختلف زمانی، فرکانسی و بین‌کانالی (مثل همبستگی سیگنال‌ها) را استخراج می‌کنند. حال با صدا زده شدن تابعی مناسب، به ازای هر کدام از طبقه‌بندها، تک‌ویژگی‌ای که بهترین نتیجه را بدهد با چندین بار آموزش و استفاده از Cross Validation پیدا می‌شود. حال این ویژگی ثابت فرض شده و با همین روند دومین و سومین ویژگی خوب نیز پیدا می‌شوند و نهایتا بین طبقه‌بندها و هایپرپارامترهای آن‌ها و ویژگی‌ها بهترین دقت استخراج می‌شود و به شخص نسبت داده می‌شوند.

ضمنا استفاده از ۳ ویژگی برتر خودش نیز یک هایپرپارامتر بود که با آزمایش و خطا انتخاب شد. تغییر این تعداد با یک خط کد نیز ممکن است، اما افزایش آن به نیاز به محاسبات زیاد خواهد داشت.

برای سنجش مدل‌ها از معیار Accuracy استفاده شده است. یکی از مشکلات در این زمینه آنبالانس بودن دیتا است که در بخش چالش‌ها نحوه غلبه بر آن توضیح داده خواهد شد.

با استفاده از این کلاس، دیگر نیازی به بررسی کردن طولانی‌مدت مدل‌های مختلف با فیچرهای مختلف نیست و به طور خودکار این بررسی انجام می‌گردد. طبیعتا همه انتخاب‌ها برای هایپرپارامترها و همه مدل‌های ممکن با توجه به محدودیت‌های پردازشی و زمانی قابل بررسی نیستند و تعداد محدودی از آنها با آزمون و خطای بسیار انتخاب شده‌اند و سپس انتخاب بین آن‌ها به چند حلقه for سپرده شده است.

در این پروژه از طبقه‌بندهای رگرسیون خطی، بردارهای پشتیبان و جنگل تصادفی استفاده شده است. همچنین ویژگی‌های انتخاب شده شامل میانگین و واریانس زمانی و فرکانسی، خود سیگنال‌ها در حوزه زمان و فرکانس، انرژیِ باندهای فرکانسی، فرکانس بیشینه، میانگین و میانه، همبستگی کانال‌ها و تبدیل فوریه‌های آنها می باشند.

۲.۴ سایز پنجره

یکی دیگر از عوامل موثر سایز پنجره بود. یعنی اینکه به ازای هر تحریک چه تعدادی سمپل را نگه داریم. موارد مختلفی از جمله ۶۴، ۱۲۸، ۲۵۶، ۵۱۲ و ۱۰۲۴ تست شدند و نهایتا پردازش‌ها همه با سایز ۱۲۸ و ۲۵۶ انجام شدند. سپس با توجه به نتایج روی داده ولیدیشن برای هر فرد بین این دو سایز انتخاب انجام می‌شود و همه اشیا در یک لیست قرار می‌گیرند. به علت اینکه برنامه‌نویسی به صورت شی‌گرا انجام شده است، این انتخاب‌ها به راحتی انجام می‌گیرند.

نکته اینکه با توجه به نرخ نمونه برداری، ۲۵۶ سمپل معادل ۱ ثانیه می‌باشد که بخش اعظم آن بعد از تحریک و بخش کوچیکی ($\frac{1}{8}$) از قبل از تحریک هستند.

۵ نتایج

۱.۵ خلاصه

برای هر شی یک تابع summary پیاده سازی شده است. با صدا زده شدن این تابع، ویژگی‌های منتخب و طبقه‌بند بهینه نمایش داده می‌شوند. به عنوان نمونه:

```
Person Number 4: Using weighted accuracy as the metric, top 3 features for this person has been extracted
These Features are:
1) stack_time
2) bandpower
3) ft_median
3 different Classifiers were tested and the best is: LogisticRegression(class_weight={0: 1, 1: 5}, max_iter=5000)
The max CV accuracy is 0.76
```

شکل ۲: خلاصه

۲.۵ دادگان اعتبارسنجی

در هربار آموزش در این قسمت، یک پنجم دیتا (معادل یک حرف) کنار گذاشته شدند و روی ۴ بخش دیگر یادگیری روی یک مدل تازه انجام شد. سپس بر روی دادگان اعتبارسنجی پیش‌بینی انجام شد و با استفاده از ترکیب نتایج و یک دیکشنری آماده و توابع bincount و argmax نامپای، از نتایج به پیش بینی حرف رسیدیم. دقت شود که این نتایج مطابق اصول ولیدیشن، بر روی دادگانی هستند که شبکه هیچگاه ندیده است و اعتبار خوبی دارند.

نتیجه به ازای تمامی افراد در جای مناسب کد داده شده است که در اینجا نیز قرار داده می‌شود. دقت شود که در کد به ازای پنجره ۱۲۸ و ۲۵۶ همه این نتایج پرینت شدند و از روی آن‌ها انتخاب انجام شد. سپس در بخش Results نتایج اصلی و نهایی ارائه شده‌اند.

```
Person Number 1: Predicted Word using cross validation is (L4KAY) With accuracy of 0.59
Person Number 2: Predicted Word using cross validation is (LUKAS) With accuracy of 0.68
Person Number 3: Predicted Word using cross validation is (HKKAS) With accuracy of 0.61
Person Number 4: Predicted Word using cross validation is (HUKAS) With accuracy of 0.76
Person Number 5: Predicted Word using cross validation is (IOKAS) With accuracy of 0.66
Person Number 6: Predicted Word using cross validation is (DCKAA) With accuracy of 0.57
Person Number 7: Predicted Word using cross validation is (JUKAS) With accuracy of 0.64
Person Number 8: Predicted Word using cross validation is (HIKAS) With accuracy of 0.73
Person Number 9: Predicted Word using cross validation is (KUKAS) With accuracy of 0.68
```

شکل ۳: نتایج اعتبارسنجی

۳.۵ دادگان تست

پس از اینکه برای هر فرد براساس دادگان ولیدیشن بین سایز پنجره ۱۲۸ و ۲۵۶ انتخاب شد، نتایج تست آنها نیز گرفته می‌شود. فرآیند تست اینگونه است که با صدا زده شدن توابع مناسب، ابتدا یک مدل کامل روی تمام دیتای ترین، آموزش داده می‌شود.

در فرآیند تنظیم فراپارامترها، این مقادیر برای هر فرد در attribute مناسب ذخیره شدند. حال در هر فرآیند آموزشی (چه دادگان اعتبارسنجی و چه تست) از این فراپارامترها و مدل انتخاب شده و فیچرها استفاده می‌گردد.

پس از آموزش دیدن مدل، هر بار روی یک پنجم دیتا پیش‌بینی انجام می‌شود و بازهم به کمک یک دیکشنری که پیشتر توصیف شد، حرف حدس زده می‌شود. نتایج در بخش Results کد موجودند و اینجا نیز قرار می‌گیرند.

```
Person Number 1: Predicted Word is LDNHA
Person Number 2: Predicted Word is LUKAS
Person Number 3: Predicted Word is KUKAS
Person Number 4: Predicted Word is KUKAS
Person Number 5: Predicted Word is WATEQ
Person Number 6: Predicted Word is WMUAA
Person Number 7: Predicted Word is 2AZEM
Person Number 8: Predicted Word is WATEM
Person Number 9: Predicted Word is WATEP
```

شکل ۴: نتایج تست

۶ چالش‌ها

۱.۶ استخراج ویژگی‌ها

در این بخش با آزمون و خطا ویژگی‌های بسیار مختلفی امتحان شدند و سپس بهترین‌ها در کلاس فیچر پیاده سازی شدند و انتخاب از بین این ۱۴ منتخب به کامپیوتر سپرده شد.

۲.۶ آنبالانس بودن دیتا

۱.۲.۶ آموزش

با توجه به اینکه نمونه‌های ۰ بسیار بیشتر از نمونه‌های ۱ هستند، شبکه این نمونه‌ها را بیشتر یاد خواهد گرفت و هیچگاه خروجی ۱ برنمیگرداند. برای حل این مشکل، در تمامی طبقه‌بندهای کتابخانه sklearn یک ویژگی برای تنظیم کردن وزن کلاس‌ها قرار داده شده است. با این کار، به هر نمونه مطابق تعدادش در دادگان در تابع هزینه یک ضریب نسبت داده می‌شود و آموزش به صورت بالانس شده انجام می‌شود.

۲.۲.۶ معیار ارزیابی

با استفاده از تابع دقت‌سنج عادی، یک خروجی ثابت صفر دقت بسیار بالایی خواهد داد که طبیعتا مطلوب نیست. برای اینکار به هر نمونه ۱ با توجه به نسبتش با تعداد صفرها (که در تک کاراکتری‌ها صفرها ۳۵ برابر و در سطرستونی‌ها تعداد صفرها ۵ برابر است)، یک ضریب نسبت داده می‌شود، به طوری که خروجی ثابت صفر دقت وزن‌دار ۵۰ درصد بگیرد که معادل حدس کاملا تصادفی می‌شود. اینگونه این معیار قابل استناد می‌شود و در تمامی بخش‌هایی که نیاز به سنجش و مقایسه است از آن استفاده شده است.

۷ پیشنهاداتی برای بهبود

۱.۷ مدل

می‌توان برای تحلیل این دادگان که به صورت سری زمانی هستند، از مدل‌های مخصوص تحلیل سری زمانی مثل شبکه‌های عمیق بازگشتی (مانند LSTM) استفاده کرد.

۲.۷ ویژگی‌ها

۱.۲.۷ استخراج دستی

علاوه بر ویژگی‌های زمانی و فرکانسی، می‌توان ویژگی‌های زمان فرکانسی را نیز به این مساله افزود. نشان داده شده است که در تحلیل سیگنال‌های مغزی این ویژگی‌ها منجر به نتایج خوبی می‌شوند.

۲.۲.۷ استخراج خودکار

می‌توان از یک شبکه عمیق (Dense و یا کانولوشنی) برای استخراج بهترین ویژگی‌ها استفاده کرد.