به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

پروژه امتیازی یادگیری ماشین

نام و نام خانوادگی اعضای گروه

فاطمه نائینیان ۸۱۰۱۹۸۴۷۹

رومینا امیدی ۸۱۰۱۹۸۵۳۹

تیر ماه ۱۴۰۱

فهرست

چکیده

در این پروژه می خواهیم داده های EEG که از مغز یک فرد هنگام مشاهده تصویر پیانو و عدم نگاه به آن را که طی تکرار ها متعدد گرفته شده را تحلیل، ویژگی هایی را از آنها استخراج و نهایتا مدل هایی را بر روی آنها برای طبقه بندی پیادی سازی کنیم تا لیبیل داده ها را پیش بینی کند و نهایتا میخواهیم مدل های طبقه بند را با هم مقایسه و بررسی کنیم که کدام طبقه بند عملکرد بهتری دارد.

اطلاعات اولیه دیتاست

در این پروژه دیتاستی حاوی دو ازمایش داریم. در ازمایش اول شخص به چهره انسان نگاه میکند و در ازمایش دوم شخص به پیانو نگاه میکند.

در این ازمایش ۱۲۸ الکترود به مغز متصل شده و سیگنال های eeg را استخراج میکند. این سیگنال ها در ۷ ثانیه با فرکاس ۵۰۰ هرتز نمونه برداری شده است. همچنین برای هر ازمایش ۴۵ بار نمونه برداری شده است.

پس در نهایت ۹۰ ازمایش با ۱۲۸ الکترود در ۳۵۰۰ بازه زمانی خواهیم داشت.

حال میخواهیم ببینیم چگونه میتوان با کمک این سیگنال های eeg بهترین بازه ای را پیدا کنیم که میتوان در ان تشخیص داد شخص به چه چیزی نگاه میکند.

تقسیم بندی دیتا

ابتدا دیتاست ها را لود میکنیم. سپس دیتا ها به تعدادی بازه تقسیم میکنیم.

در این پروژه داده ها را به ۵۰ بازه زمانی تقسیم میکنیم. سپس برای هر بازه ویژگی استخراج میکنیم و با کمک آن ویژگی ها یک طبقه بند اموزش میدهیم تا دقت ان را به دست اوریم.

هر بازه زمانی ۰.۱۴ ثانیه خواهد بود. در نتیجه در هر بازه ۷۰ نمونه در ۱۲۶ کانال الکترود خواهیم داشت.

استخراج ویژگی

از سیگنال های eeg به طرق مختلف میتوان ویژگی استخراج کرد.

برای مثال ویژگی های DFA و HFD و LZC و PDF و Count Sign Changes و Hjorth Activity و Hjorth mobility و Samp En را می توان استخراج کرد.

که ویژگی های Hjorth از ویژگی های مبتنی بر واریانس و مشتق سیگنال هستند.

در نهایت دو ویژگی از میان این ها انتخاب می کنیم که بهترین دقت را داشته باشد : DFA و HFD

برای انتخاب ویژگی از کتابخانه ای-ای-جی استفاده کردیم این کتابخانه ویژگی های مختلفی برای داده های ای-ای-جی تعریف می کند و ما با استفاده از ترکیبی از این ویژگی ها می توانیم از این داده ها ویژگی استخراج کنیم و سپس به کمک آنها مدل را اجرا کنیم. ویژگی هایی که انتخاب کردیم به شرح زیر هستند:

detrended fluctuation analysis(DFA)

detrended fluctuation analysis (DFA) is a method for determining the statistical self-affinity of a signal. It is useful for analysing time series that appear to be long-memory processes (diverging correlation time, e.g. power-law decaying autocorrelation function) or 1/f noise.

Higuchi fractal dimension(HFD)

In fractal geometry, the Higuchi dimension (or Higuchi fractal dimension (HFD)) is an approximate value for the box-counting dimension of the graph of a real-valued function or time series. This value is obtained via an algorithmic approximation so one also talks about the Higuchi method.

با توجه به اینکه دیتاست یک محیط باینری است و برای تشخیص دیدن یا عدم مشاهده تصویر پیانو می باشد، معیار های نام برده شده به سبب آنکه در رابطه با تعداد بعد های مربوط به هر سیگنال ای-ای-جی می باشد می تواند متفاوت شوند و معیار مناسبی برای تشخیص کلاس باشند همچنین معیار دیگر که دی-اف-ای هست، همانطور که در توضیحات نیز آورده شده برای سیگنال هایی که به حافظه بلند مدت مربوط می شود ربط دارد می تواند در این تسک ویژگی مناسبی باشد چرا که داده ها در این قسمت نیز در بازه ی ۷ ثانیه ای محاسبه شده اند و به دلیل ممتد بودن و وابسته بودن سیگنال به ثانیه های قبلی، می تواند ویژگی مناسبی باشد.

با محاسبه هریک از این ویژگی ها برای هر سیگنال الکترود نهاتا ۱۲۶*۲ فیچر برای داده بدست می آید که می توان به کمک آن مدل را آموزش داد.

تقسیم بندی به داده ها آموزش و آزمون

همانطور که مشاهده کردیم داده ها را به چندین بازه تقسیم کردیم و ویژگی هایی برای هر بازه استخراج کردیم.

حال برای اینکه بتوانیم بک مدل برای هر بازه به دست اوریم، لازم است تا به داده های آموزش و آزمون تقسیم بندی کنیم. این تکنیک کمک میکند تا ارزیابی بهتری از مدل داشته باشیم. به گونه ای که مدل را روی داده های آموزش اجرا میکنیم و سپس با کمک داده های آزمون، صحت عملکرد مدل را میسنجیم. داده های آزمون داده هایی هستند که مدل از قبل از آنها اطلاعاتی ندارد و برای مدل داده هایی جدید هستند. پس میتواند معیار خوبی برای سنجش مدل باشد.

در تقسیم بندی باید توجه کنیم که مدل بر روی داده ها overfit نشود. در صورتی که overfit شود دقت مدل برای داده های آموزش زیاد می شود و دقت برای داده های آزمون کم می شود. اگر مدل underfit باشد نیز دقت خوبی نخواهیم گرفت.

بنابراین ابتدا داده ها را shuffle میکنیم و سپس ۷۰ درصد آن ها را به عنوان داده آموزش و ۳۰ درصد آن را به عنوان داده آزمون در نظر میگیریم.

طبقه بندی

حال میخواهیم برای هر بازه با طبقه بند های مختلف دقت را به دست اوریم.

ابتدا توضیح مختصری از هر یک از طبقه بند ها خواهیم داد.

SVM : در این طبقه بند اگر داده ها جدایی پذیر باشند، بهترین مرز را پیدا میکند که بیشترین میزان حاشیه و فاصله را از داده ها داشته باشد. اگر داده ها جدایی پذیر نباشند، مصالحه ای بین میزان خطا و حاشیه خواهیم داشت و در نهایت یک مرز، با دو حاشیه خواهیم داشت. داده تست با کمک این مرز، طبقه بندی می شود.

Gaussian Naïve Bayes : در این طبقه بند فرض می شود که ویژگی ها نرمال و از هم مستقل هستند و با این فرض، داده تست به کلاسی تعلق میگیرد که احتمال اینکه داده از ان کلاس باشد، بیشتر باشد.

Decision Tree : در این طبقه بند با کمک ویژگی هایی که داریم در هر مرحله یک ویژگی را ملاک قرار می دهد و داده ها را با شاخه هایی که ناشی از ان ویژگی است، به یک برگ میبرد و در ان برگ با ویژگی

بعدی، شاخه و برگ جدید را پیدا میکند. حال داده تست از شاخه اصلی شروع به مقایسه شدن میکند تا در نهایت به یک برگ تعلق پیدا کند.

KNN : در این طبقه بند مبنای تصمیم گیری، تعداد همسایه های نزدیک به ان داده است. به صورتی که برای هر داده تعداد مشخصی از داده های نزدیک به آن را بررسی میکنیم و داده فعلی به پرتکرار ترین برچسب تعلق خواهد داشت.

MLP : از انواع شبکه های عصبی محسوب می شود. به صورتی که در شبکه عصبی تعدادی لایه مخفی داریم که در هر لایه وزن هایی خواهیم داشت. هنگام آموزش این شبکه ها، این وزن ها تعیین می شود و وقتی داده های تست را به عنوان ورودی به این طبقه بند بدهیم، وزن ها سبب می شود تا خروجی نهایی تعیین شود و برچسب به داده تعلق گیرد.

استاندارد آماری و بازه های اطمینان

فرض کنید تعدادی مشاهده داریم، میخواهیم این مشاهدات را با کمک استاندارد های آماری بیان کنیم. یکی از ساده ترین راه ها استفاده از میانگین است. اما میانگین به تنهایی معیاری خوبی برای توصیف همه مشاهدات نیست. یکی از روش های خوب، استفاده از بازه های اطمینان است. به صورتی که میتوان تضمین کرد چه درصدی از مشاهدات در بازه به دست امده قرار میگیرند.

$$(\overline{X} - t_{(n-1,1-\frac{\alpha}{2})}\frac{S}{\sqrt{n}} \quad , \quad \overline{X} + t_{(n-1,1-\frac{\alpha}{2})}\frac{S}{\sqrt{n}})$$

بازه اطمینان به روش بالا به دست می آید.

$\bar{x}$: میانگین

S : انحراف معیار

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

n : تعداد عناصر

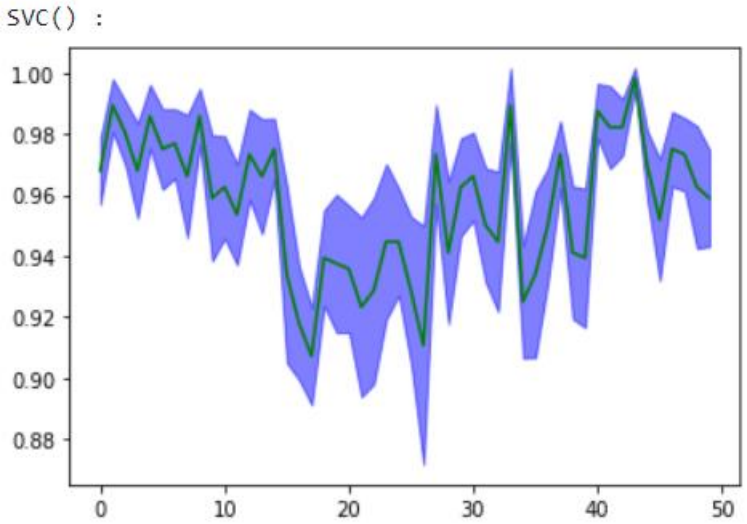
$t(n-1,1-\alpha 2)$: به ازای میزان اطمنان از جدول توزیع استاندارد به دست می اید که برای بازه ۹۵ درصد، مقدار ۱.۹۶ را دارد.

حال برای random seed های مختلف، همه طبقه بند های بالا را برای هر بازه اجرا میکنیم و دقت انها را به دست می اوریم. سپس با کمک دقت ها ، میانگین و بازه های اطمینان را برای هر بازه به ازای هر طبقه بند نمایش می دهیم.

تحلیل نتایج مدل ها

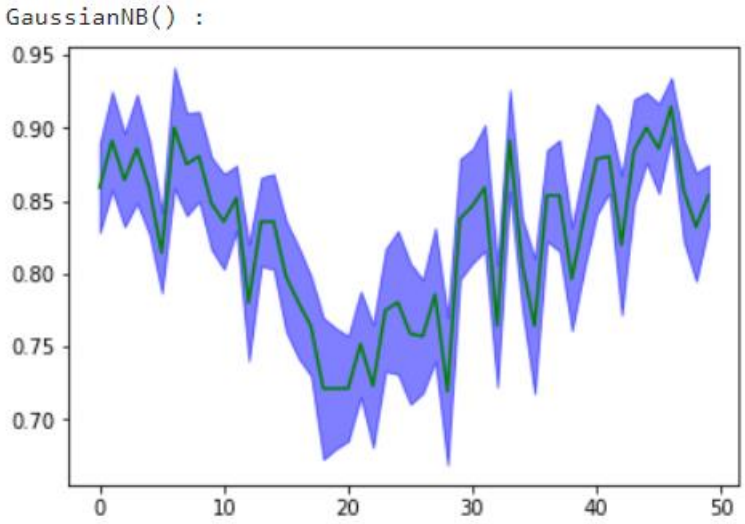
دقت شود که محور افقی در این نمودار ها شماره ی بازه است و از آنجایی که ما سیگنال های EEG را به ۵۰ قسمت تقسیم کردیم محور افقی از ۱ تا ۵۰ می باشد. همچنین محور عمودی مربوط به دقت است:

SVC:



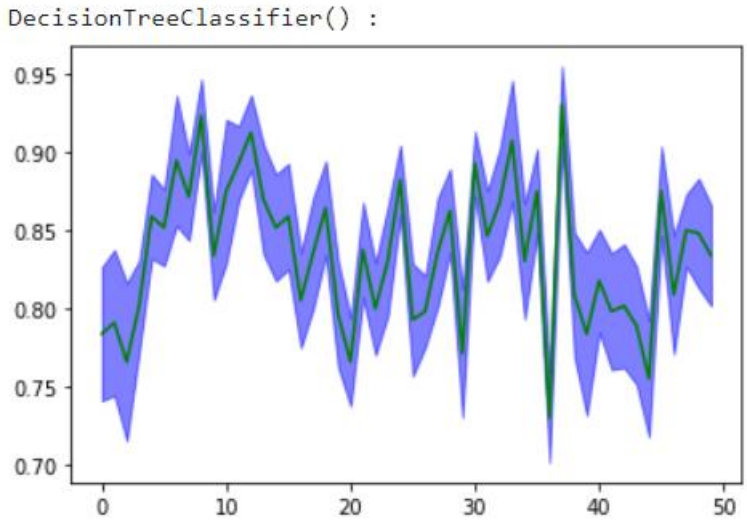
همانطور که مشاهده می شود برای زمان های وسط میزان دقت بسیار کم شده است و بیشترین دقت مربوط به کمی پس ار ابتدای بازه است همچنین به طور میانگین دقت در حدود 94%می باشد که دقت بسیار مطلوبی است.

Gaussian Naïve bayas:



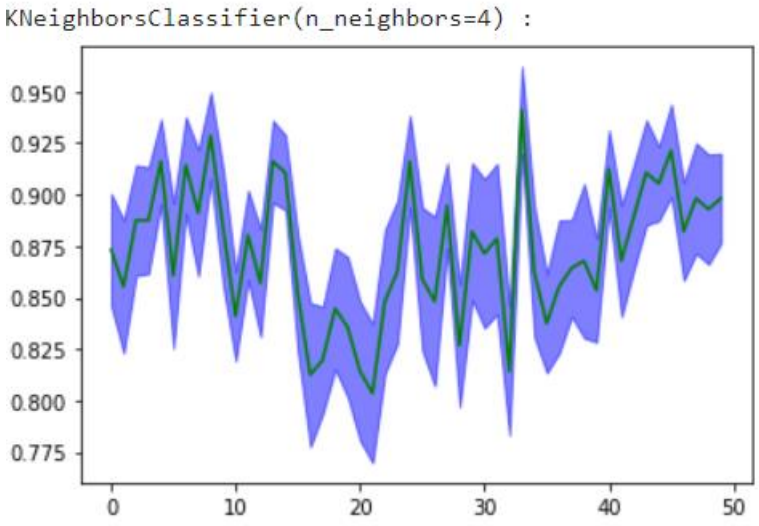
دقت در روش گاوسین همانطور که مشاهده می شود نسبت به svm پایین تر است و مجددا همانند svmدر وسط سیگنال دقت پایین تر و در سر و ته بازه دقت بالاتری دارد. اما مجددا بیشترین دقت مربوط به کمی پس از ابتدای بازه داده سیگنال است. همچنین میانگین دقت نیز 83% است.

Decision Tree:



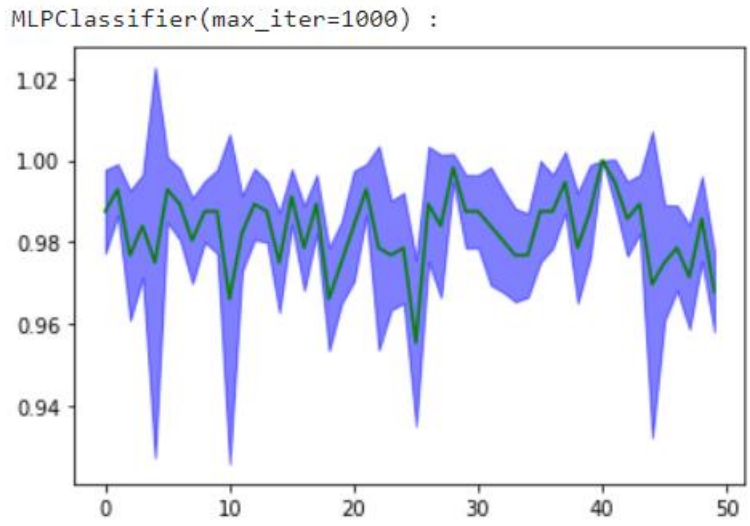
این الگوریتم بر خلاف دو مدل قبل به طور تقریبی دقت یکسانی در بازه های مختلف دارد و برعکس دو مدل قبل در ابتدای سیگنال دقت کمتری دارد. این الگوریتم نیز به طور میانگین دقت ۸۵٪ دارد. اما بازه اطمینان مربوط به درخت تصمیم از مدل بیز کوچکتر و بنابراین مدل دقیق تر است.

KNN:



در این حالت نیز همانند دو مدل اول دقت در وسط کمینه و در ابتدا و انتها بیشتر است همچنین بازه اطمینان نسبت به دو مدل دیگر کمتر است و دقت بهتری دارد که در حدود ۸۷٪ می باشد.

MLP:



در MLPکه نوعی شبکه عصبی است همانطور که مشاهده می شود دقت بسیار بسیار بالاتر از مدل های قبل است ولی بازه اطمینان گسترده تر ایت بدین معنا که واریانس درصد دقت در بازه ی نسبتا بزرگی قرار می گیرد اما نسبت به مدل های دیگر استیبل تر است بدین معنا که در تمام بازه دقت به طور تقریبی یکسان و یکنواخت است و دقت نیز از همه ی مدل های پیشین بالا تر و در حدود ۹۹٪ است.

ننیجه گیری اینکه بهترین مدل طبقه بند از این ۵ مدل MLPاست زیرا هم مستقل از این است که از چه بازه ای از سیگنال نمونه برداری کنیم و هم دقت آن نسبت به سایر مدل ها بالاست.

سوال آخر: همانطور که مشاهده کردیم در میانه سیگنال دقت کم و در ابتدا و انتها دقت بالاتر است و به طور میانگین نیز در کمی پس ار ابتدای بازه مدل دقت بالاتر. حدس زده می شود علت این امر آن است که اولا تا قبل از مشاهده ی تصویر فرد احساس خاصی ندارد و سیگنال ها مشابه هم هستند. در ادامه و لحظاتی پس از مشاهده ی عکس ها سیگنال ها بسیار از هم متمایز می شوند و لذا تقسیم بندی آنها راحت تر و دقت مدل ها در این نقاط بالاتر است.