



دانشگاه تهران

پردیس دانشکده های فنی

دانشکده برق و کامپیوتر



پروژه نهایی یادگیری ماشین

دکتر ابوالقاسمی

دکتر توسلی پور

اردیبهشت 1403

فصل 1	3
مقدمه	3
فصل 2	4
تعریف مسئله و شرح پروژه	4
2.1 تعریف کلی مسئله	4
1. 2.2 فرمت داده ها	7
2.3 گروه بندی	7
2.4 گزارش اولیه	8
2.5 پیش پردازش داده و استخراج ویژگی	9
2.7 خوشه بندی	12
فصل 3	13
گزارش کار	13
فصل 4	14
بارم بندی	14
فصل 5	15
نکات پایانی	15

فصل 1

مقدمه

سیگنال‌های الکتروانسفالوگرام (EEG) نقش مهمی در فهم فعالیت‌های مغزی و فرایندهای شناختی دارند. امروزه پردازش و طبقه‌بندی این سیگنال‌ها با استفاده از روش‌های یادگیری ماشین، کاربرد گسترده‌ای در واسطه‌های مغز و رایانه، تشخیص‌های بالینی و به طور کلی در علوم اعصاب دارد. در این پروژه، شما ابتدا با داده‌های سیگنال EEG آشنا می‌شوید و روش‌های مختلف پیش‌پردازش، تمیز کردن داده و حذف نویز مزاحم که معمولاً در ارتباط با سیگنال‌های دنیای واقعی وجود دارد را بررسی خواهید کرد. سپس با استفاده از روش‌هایی که در درس یادگرفته‌اید یا با مطالعه‌ی روش‌های جدید، به استخراج ویژگی‌هایی که در سیگنال‌های EEG می‌توانند مفید باشند خواهید پرداخت. در ادامه نیز انتظار می‌رود با استفاده از ویژگی‌های استخراج‌شده در مرحله‌ی قبل، داده‌ها را به طور مناسب طبقه‌بندی و خوشه‌بندی نمایید. امیدواریم پس از انجام این پروژه، توانایی شما در به کارگیری عملی الگوریتم‌های یادگیری ماشین برای حل مسائل دنیای واقعی افزایش یابد.

فصل 2

تعریف مسئله و شرح پروژه

2.1 تعریف کلی مسئله

تصویرسازی حرکت¹، فرایندی است که طی آن از فرد خواسته می‌شود تا بدون انجام حرکتی مانند حرکت دادن انگشتان دست یا پا، انجام آن را در ذهن خود تصور کند. در حین انجام این عمل از فرد سیگنال EEG ضبط می‌شود. سیگنال‌های EEG، فعالیت الکتریکی مغز را با قرار دادن الکترودهایی که بر روی نواحی مختلف مجمله قرار داده می‌شود ضبط می‌کنند. این سیگنال‌ها، پتانسیل الکتریکی را به صورت تابعی از زمان نشان می‌دهند. انتظار می‌رود در هنگام تصور فرد از فعالیت حرکتی، بتوان الگوهایی مرتبط با فعالیت بخش‌های مرتبط مغز را با استفاده از سیگنال‌های EEG تشخیص داد. هدف از انجام این پروژه، پردازش سیگنال EEG مرتبط با این تسک و تشخیص این الگوها می‌باشد.

این پروژه در دو فاز تعریف می‌شود:

فاز اول: به طور کلی هدف از این فاز، آشنایی با سیگنال‌های EEG به خصوص در تصویرسازی حرکتی و چالش‌های آن می‌باشد. در این فاز از شما انتظار می‌رود گزارش کوتاهی در مورد نوع داده‌های سیگنال‌های EEG و سپس چالش‌هایی که برای تشخیص الگو از سیگنال‌های تصویرسازی حرکتی وجود دارد آماده کنید. در هر مرحله، لازم است پس از توضیح چالش مدنظر، راهکار(های) مرتبط با حل آن چالش را به صورت مختصر توضیح داده و مزایا و معایب آن را با ارجاع به منابع معتبر شرح دهید. برای مثال یکی از روش‌های استخراج ویژگی از سیگنال‌های EEG، استفاده از الگوریتم‌های ICA² و CSP³

¹ Motor Imagery

² Independent Component Analysis

³ common spatial patterns

است. ابتدا این الگوریتم ها را توضیح داده و سپس چالشی که این روش برای حل آن مناسب است و مزایا و معایب آن در حل این چالش را معرفی کنید.

فاز دوم: در این فاز، به صورت عملی با استفاده از داده‌هایی که به صورت آماده در اختیار شما قرار داده خواهد شد، به پیاده‌سازی مراحل که در ادامه اشاره خواهد شد می‌پردازید. دیتای مورد نظر، سیگنال‌های EEG است که از تعدادی فرد که در تسک تصویرسازی حرکتی شرکت کرده‌اند، گرفته شده است. در این تسک، بعد از نمایش دادن علامت سمت راست، سمت چپ یا پایین برای مدت مشخصی به فرد مورد نظر، از فرد خواسته شده است تا حرکت متناسب با این علامت را در ذهن خود تصویر کند. علامت سمت راست، متناظر با تصور حرکت دادن دست راست، علامت چپ متناظر با تصور حرکت دادن دست چپ و علامت پایین متناظر با حرکت دادن پا(ها) است. البته برای هر فرد، تنها دو کلاس از این موارد انتخاب شده است.

در این فاز لازم است تا ابتدا داده‌ها را به صورت مناسب پیش‌پردازش کرده تا آن‌ها را به فرمت مناسب تبدیل کرده و همچنین نویز و یا تداخلات اضافی که ممکن است به مراحل بعد منتقل شود را حذف کنید. سپس با استفاده از روش‌های استخراج ویژگی که در ادامه به آن‌ها اشاره شده است یا با توجه به روش‌هایی که خودتان در نظر دارید و در فاز قبل راجع به آن‌ها اطلاعاتی جمع‌آوری کرده‌اید، به استخراج ویژگی از داده‌ها بپردازید. در نهایت اگر ویژگی‌های استخراجی مناسب باشند، باید به طبقه‌بندی و خوشه‌بندی داده‌ها بپردازید. در صورت عدم موفقیت در این مرحله، ممکن است به ویژگی‌های بهتری احتیاج داشته باشید تا بتوانید به دقت مناسبی برسید.

2.2 دیتاست

در این پروژه، ما از مجموعه داده‌ای استفاده می‌کنیم که به طور خاص برای به چالش کشیدن و پیشرفت توسعه فناوری های رابط مغز و کامپیوتر⁴ طراحی شده است⁵. این مجموعه داده ابزاری برای آزمایش الگوریتم‌هایی است که آزمایش‌های EEG را

⁴ BCI

⁵ [BCI Competition IV](#)

طبقه‌بندی می‌کنند، که در آن هر آزمایش یک سیگنال EEG پنجره‌دار مربوط به طول ثابت است و وضعیت ذهنی خاصی را نشان می‌دهد. برخلاف مجموعه داده‌های معمولی، که در آن تغییرات حالت ذهنی مشخص می‌شود، این مجموعه داده بر برنامه‌های BCI ناهمزمان متمرکز است. در اینجا، طبقه‌بندی‌کننده‌ها باید به‌طور مداوم بر روی داده‌های EEG ورودی بدون نشانه‌های صریح که نشان می‌دهد کاربر قصد خود را تغییر می‌دهد، کار کنند. این پیچیدگی قابل توجهی را معرفی می‌کند، زیرا الگوریتم باید این تغییرات را به‌طور مستقل تشخیص دهد. علاوه بر این، مجموعه داده شامل دوره‌هایی است که کاربر قصد کنترل خاصی ندارد. در طول این فواصل، برای طبقه‌بندی‌کننده مهم است که به درستی عدم وجود قصد کنترل را شناسایی کرده و صفر را برگرداند، که نشان دهنده عدم وابستگی به حالت‌های ذهنی هدف است. این ویژگی منحصر به فرد مجموعه داده تضمین می‌کند که الگوریتم‌های ما می‌توانند به‌طور موثر سناریوهای دنیای واقعی را که در آن حالت ذهنی کاربر به‌طور مداوم با سیستم BCI درگیر نیست، کنترل کند.

برای دانلود این مجموعه داده، به سایت زیر مراجعه کنید و Data Set 1 را دانلود کنید:

(داده‌های 100Hz: این سیگنال‌ها قبل از زیر نمونه برداری⁶ با فرکانس 49 هرتز فیلتر شده‌اند)

https://www.bbc.de/competition/download/competition_iv/BCICIV_1_mat.zip

https://www.bbc.de/competition/iv/desc_1.html

لینک مقاله:

<http://dx.doi.org/10.1016/j.neuroimage.2007.01.051>

⁶ subsampling

2.2.1 فرمت داده ها

سیگنال‌های پیوسته EEG دارای 59 کانال هستند. برای داده‌های کالیبراسیون، نشانگرهایی وجود دارد که نقاط زمانی ارائه ی نشانه و کلاس‌های هدف مربوطه را نشان می‌دهند.

داده ها در قالب *.mat (Matlab) حاوی متغیرهای زیر ارائه می شوند:

- cnt: سیگنال های EEG پیوسته، اندازه [time x channels]. آرایه در نوع داده INT16 ذخیره می شود. برای تبدیل آن به مقادیر uV، از دستور زیر استفاده کنید.

$$\text{cnt} = 0.1 * \text{double}(\text{cnt})$$

- mrk: ساختار اطلاعات نشانه هدف با فیلدها (فایل داده های ارزیابی شامل این متغیر نیست)
- pos: بردار موقعیت های نشانه در سیگنال های EEG داده شده در نمونه واحد،
- y: بردار کلاس های هدف (1- برای کلاس یک یا 1 برای کلاس دو)،
- nfo: ساختار ارائه اطلاعات اضافی با فیلدها
- fs: نرخ نمونه برداری،
- clab: آرایه سلولی از برچسب های کانال،
- classes: آرایه سلولی از نام کلاس های motor imagery،
- xpos: موقعیت x الکترودها در یک پروجکشن 2 بعدی،
- ypos: موقعیت y الکترودها در یک پروجکشن 2 بعدی.

شما می‌بایست دو فایل calibration از فایل‌های موجود در پوشه دانلود را که کلاس‌های مختلفی دارند برای کار خود انتخاب کنید و تمام مراحل بعد را به طور مستقل روی این دو فایل اعمال کرده و در نهایت نتایج را باهم مقایسه کنید. در هر فایل انتخابی، از بردار pos برای نقطه شروع هر پنجره زمانی می‌توانید استفاده کنید.

2.3 گروه‌بندی

شما می‌توانید به صورت انفرادی و یا گروه‌های حداکثر سه نفره فعالیت داشته باشید. اسامی افراد گروه را یک نفر به نمایندگی **تا آخر اردیبهشت** در محل تحویل گروه‌بندی ارسال کند.

2.4 گزارش اولیه

در این مرحله لازم است به صورت گروهی گزارشی تهیه کنید.

گزارش اولیه با هدف بررسی مفهوم تصویرسازی حرکتی، چالش‌های مرتبط با آن و روش‌های به کار گرفته شده در پیش‌پردازش و استخراج ویژگی‌های سیگنال‌های EEG طراحی شده است. هدف این است که درک عمیقی نسبت به پیشرفت‌ها و تکنیک‌های موجود در این زمینه به دست آید.

این گزارش باید شامل موارد زیر باشد :

1. مقدمه‌ای بر تصویرسازی حرکتی

- تعریف تصویرسازی حرکتی و اهمیت آن

- بررسی کاربردهای آن با تأکید بر رابط‌های مغز و کامپیوتر.

2. چالش‌های تصویرسازی حرکتی

- شناسایی و توضیح چالش‌های اصلی که در تحقیقات و کاربردهای تصویرسازی حرکتی وجود دارد.

- بررسی راه‌حل‌های بالقوه و تحقیقات جاری برای غلبه بر این چالش‌ها.

3. پیش‌پردازش سیگنال‌های EEG:

- بحث در مورد اهمیت پیش‌پردازش سیگنال‌های EEG در زمینه تصویرسازی حرکتی.

- توضیح مراحل معمول پیش‌پردازش با تمرکز بر:

- فیلترهای میان‌گذر (۸-۳۰ هرتز): نقش آن در جداسازی باندهای فرکانسی خاص (فرکانس‌های مو و بتا)

- فیلترهای فضایی^۷: ارائه اطلاعاتی در مورد تکنیک‌های زیر، با توضیح کاربردها و مزایای آن‌ها در سیگنال‌های

EEG

- Common Average Reference (CAR)
- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Minimum Norm Estimation (MNE)
- Laplacian Filter

4. تکنیک‌های استخراج ویژگی

- توضیح تکنیک‌های مختلف استخراج ویژگی که در تجزیه و تحلیل سیگنال‌های EEG به کار می‌روند با تمرکز ویژه بر

الگوریتم CSP

مهلت تحویل گزارش اولیه تا ۴ ام خرداد می‌باشد.

⁷ Spatial Filtering

2.5 پیش پردازش داده و استخراج ویژگی

در این بخش شما یاد می گیرید که چگونه فیلترهای میان گذر را اعمال کنید و روش های مختلف استخراج ویژگی ها را برای آماده سازی داده ها برای کاربردهای یادگیری ماشین به کار ببرید.

این بخش شامل مراحل زیر است.

1. بارگذاری داده

2. فیلترهای میان گذر:

- فیلتر میان گذر را بر روی داده های EEG اعمال کنید تا باندهای فرکانسی خاصی که برای تجزیه و تحلیل شما مهم هستند را جدا کنید (مانند باندهای بتا و مو). برای ایجاد فیلترهای از ماژول `scipy.signal` استفاده کنید. (باترورت مرتبه ۳ در باند فرکانسی ۸ تا ۳۰ هرتز)

3. Spatial Filtering:

- پس از فیلتر کردن فرکانس، از فیلتر فضایی برای کاهش نویز و بهبود کیفیت سیگنال استفاده کنید. تکنیک هایی مانند CAR و فیلترهای لاپلا سین توصیه می شود.

- تکنیک های دیگر پیش پردازش مانند تجزیه مؤلفه های اصلی (PCA) و تجزیه مؤلفه های مستقل (ICA) را برای پاک سازی بیشتر داده ها بررسی کنید.

4. تصویرسازی داده ها:

- داده ها را قبل و بعد از اعمال فیلترهای فضایی تصویرسازی کنید. از نمودارها برای مقایسه سیگنال های خام و فیلتر شده استفاده کنید.

- برای داده‌های با بعد بالا، برای درک بهتر ساختار داده‌ها از t-SNE برای تصویرسازی استفاده کنید.

5. استخراج ویژگی‌ها:

- الگوریتم الگوی فضایی مشترک (CSP) را برای استخراج ویژگی‌ها از داده‌های پیش‌پردازش شده پیاده‌سازی کنید. CSP در بهبود واریانس سیگنال بین دو کلاس در داده‌های EEG مؤثر است. (از روش‌های دیگر استخراج ویژگی نیز برای مقایسه استفاده کنید)

- پس از استخراج ویژگی‌ها، ویژگی‌های تبدیل شده را ترسیم کنید تا تغییرات را تجزیه و تحلیل کنید و کیفیت ایجاد تمایز بین کلاس‌های مختلف را بررسی کنید.

- مقایسه‌های بصری بین روش‌های مختلف استخراج ویژگی را ارائه دهید تا عملکرد آن‌ها را ارزیابی کنید.

2.6 طبقه‌بندی

همانطور که پیشتر ذکر شد، هدف ما در این قسمت، طبقه‌بندی استفاده از ویژگی‌های استخراجی مرحله قبل است. در مراحل مختلف طبقه‌بندی، مراحل زیر را در نظر داشته باشید.

1. داده‌ها را به دو دسته آموزش و تست تقسیم کنید. داده تست باید حداقل ۲۵ درصد کل داده باشد. برای جداسازی

داده تست و آموزش دقت داشته باشید که نسبت دو کلاس در هر دو دسته از داده‌ها، به یک میزان باشد.

2. در این مرحله باید با استفاده از روش‌های یادگیری ماشین که در طول ترم آموخته‌اید، طبقه‌بندی را انجام دهید. در این

مرحله **حداقل از سه** روش جداگانه استفاده کرده و نتایج را با هم مقایسه و تحلیل کنید. در این بخش صرفاً کدها به

تنهایی حائز اهمیت نیستند بلکه در کنار آن‌ها تحلیل نمودارهای داده و همچنین مقایسه‌ی روش‌های مختلف، اهمیت

ویژه ای دارد . می‌توانید برای طبقه‌بندی از روش های $XGBoost$, $AdaBoost$, SVM^8 , KNN^9 , MLP^{10} ،

$Logistic Regression$ یا هر روش دیگری استفاده کنید.

3. در هر کدام از روش ها، می‌توانید از روش های $Normalization$ نیز استفاده کنید.

4. برای هر طبقه‌بندی که مورد استفاده قرار می‌دهید، ماتریس آشفتگی، $ROC Curve$ ، تحلیل میزان خطای هر کلاس

و سایر موارد مورد نیاز را در گزارش خود بیاورید.

2.7 خوشه‌بندی

در اینجا همانند قسمت قبلی باید با استفاده از روش‌هایی که در درس آموخته‌اید، با انتخاب حداقل دو روش، داده‌ها را خوشه‌بندی کنید. با استفاده از روش‌هایی مثل $silhouette score$ تعداد خوشه‌های مناسب را پیدا کرده و نمودار آن را رسم کنید. سپس خوشه‌بندی را به ازای تعداد خوشه مناسب که بدست آوردید و همچنین به ازای 2 مقدار دیگر برای تعداد خوشه انجام دهید و نتایج حاصل را تحلیل و بررسی کنید. همچنین شباهت داده‌ی درون یک خوشه و تفاوت بین خوشه‌ها و دلایلی که برخی داده‌ها در یک خوشه قرار گرفته‌اند باید بررسی دقیق شوند. تحلیل و گزارش در این بخش از اهمیت بالایی برخوردار است.

⁸ Support vector machines

⁹ K-nearest neighbors

¹⁰ Multilayer perceptron

فصل 3

گزارش کار

همانطور که قبلا هم گفته شد، علاوه بر کد درست، گزارش کار مفصل و توضیح و تحلیل درست داده و نمودارها از اهمیت بسیار بسیار بالایی برخوردار است. سعی کنید تمام نکات قابل ذکر در انجام پروژه را در گزارش کار ذکر کنید. دقت داشته باشید که گزارش کار اولیه، که مهلت آن **تا ۴ ام خرداد** است، به عنوان مقدمه گزارش کار نهایی است. گزارش نهایی هم همراه با کد و تحلیل موارد خواسته شده **تا ۱۹ تیرماه** فرصت دارد. در زیر نکاتی را متذکر می شویم که حتما باید در گزارش کار نهایی ذکر شود.

- توضیح کامل نحوه کارکرد روش هایی که برای قسمت پیش پردازش استفاده کردید.
- تحلیل های لازم برای بخش خوشه بندی با تعداد خوشه هایی متفاوت.
- توضیح درباره علت انتخاب مدل هایی که برای طبقه بندی استفاده کردید.
- ذکر معیار هایی نظیر precision، recall، F1 score و غیره، برای هر کدام از مدل هایی که آموزش داده اید.
- معیار نهایی نمره دهی گزارش کار بوده و **گزارش کار باید شامل تمامی بخش های مهم کد و توضیحات آنها** باشد و در صورت نبودن توضیحات کد در گزارش کار نصف نمره آن بخش را نمی گیرید.
- گزارش کار میبایست مرتب و بخش بندی شده باشد و تحلیل های دقیق داشته باشد.
- کدهای خود را به صورت عکس در داخل گزارش کپی نکنید و با فرمتی مناسب متنی در گزارش قرار دهید.
- گزارش اولیه و نهایی میتواند به زبان فارسی و یا انگلیسی باشد.

فصل 4

بارم بندی

از ۱۰۰	نمره دهی
۲۰	گروه بندی و گزارش اولیه
۳۰	پیش پردازش و استخراج فیچر
۳۰	طبقه بندی
۲۰	خوشه بندی

فصل 5

نکات پایانی

- هیچگونه شباهتی در انجام این پروژه بین افراد مختلف پذیرفته نمی شود. در صورت کشف هرگونه تقلب، مطابق قوانین درس با افراد خاطی برخورد خواهد شد.
- استفاده از مراجع با ارجاع به آنها بلامانع است. اما در صورتی که گزارش شما ترجمه عینی از آن ها باشد، یا از گزارش افراد دیگر استفاده کرده باشید کار شما تقلب محسوب می شود.
- بعد از مطالعه ی کامل و دقیق این توضیحات، در صورتی که سوالی در مورد پروژه داشتید به طراحان پروژه ایمیل بزنید.
- ایمیل دستیاران آموزش:

علیرضا حسینی: arhosseini77@ut.ac.ir

کیانا هوشانفر: k.hooshanfar@ut.ac.ir

رضا رضائیان: r.rezaeian@ut.ac.ir

پرهام سازدار: p.sazdar@gmail.com