دانشگاه تهران

پردیس دانشکده های فنی

دانشکده برق و کامپیوتر

پروژه نهایی یادگیری ماشین

اساتید درس:

دکتر اعرابی

دکتر ابوالقاسمی

دکتر توسلی پور

پاییز ۱۴۰۲

فهرست مطالب

۱) اهداف آموزشی

در پروژه نهایی درس یادگیری ماشین، قصد داریم شما را با استفاده از تجربه ای عملی، با چالش ها و جذابیت های این حوزه در دنیای واقعی آشنا کنیم. در طول این پروژه، با چالش های مختلفی نظیر جمع آوری و تمیز کردن داده، استخراج هدفمند ویژگی، استفاده از مدل های طبقه بند و خوشه بند و تحلیل و بررسی نقادانه نتایج آن ها روبرو خواهید شد. گاها برای حل این چالش ها، نیازمند در پی گرفتن رویکردی چند وجهی و خلاقانه و همینطور جستجو در منابع مربوطه خواهید بود. همچنین مشابه با هر پروژه دیگری در این حوزه، برای حل مسائل نیاز خواهید داشت تا با مطالعه، دانش حوزه[1] مورد نیاز برای استخراج ویژگی های مناسب را بدست آورید.

امید است این پروژه به شما این امکان را دهد که کاربردهای عملی تکنیک‌های یادگیری ماشین که در طول درس فراگرفته اید را مشاهده کنید و مهارت[2] های ضروری برای بکارگیری آن ها در شرایط غیر آیده آل دنیای واقعی را بدست آورید.

[1] Field knowledge
[2] Skill

۲) تعریف مسئله و شرح پروژه

۲.۱– تعریف کلی مسئله

این پروژه در چهار قسمت تعریف می شود:

۱. ابتدا شما باید در بازه زمانی مشخص شده، داده های صوتی لازم را جمع آوری کنید. از این داده های صوتی در قسمت های بعدی پروژه استفاده خواهد شد.

۲. در ادامه لازم است یک گزارش اولیه آماده کنید که تحقیقی در مورد مباحث زیر خواهد بود:

- توضیحی مختصر راجع به نحوه کار با داده های صوتی در الگوریتم های یادگیری ماشین
- توضیح مختصر راجع به روش های پیاده سازی تسک ASR[3] (تبدیل اتوماتیک صوت به متن)

۳. پس از استخراج ویژگی های مناسب، با استفاده از الگوریتم های طبقه بندی و خوشه بندی بایستی اطلاعات خواسته شده را از داده های صوتی ای که جمع‌آوری کرده اید استخراج کنید.

۴. در این قسمت به کمک یک مدل آماده و داده هایی که قبلا جمع آوری کرده اید، ASR را به روش انتها-به-انتها برای زبان فارسی انجام خواهید داد. سپس نتایج را ارزیابی خواهید کرد و حساسیت مدل را نسبت به ویژگی هایی همچون جنسیت و لهجه را بررسی می کنید.

[3] Automatic speech recognition

۲.۲– جمع‌آوری داده

در بخش جمع‌آوری داده لازم است هر نفر به صورت **انفرادی**، تعدادی جمله مناسب را تهیه کند و آن ها را تک به تک بخواند. ضمنا برای انجام قسمت خوشه بندی و طبقه بندی و همینطور پیاده سازی تسک ASR، علاوه بر فایل های صوتی بایستی اطلاعات دیگری را نیز در کنار آن ها ضمیمه کنید.

- حجم کل متن خوانده شده توسط هر فرد باید بین ۷۰۰ تا ۸۰۰ کلمه باشد. هر جمله از متن را بصورت جداگانه خوانده و هر یک از آن ها را در یک فایل جداگانه با فرمت mp3. ذخیره نمایید. لذا دقت کنید که طول جملات متن شما نباید خیلی زیاد یا خیلی کم باشد، تا در نهایت تعداد مناسبی داده داشته باشید.
- نیازی نیست که جملات به هم ارتباطی داشته باشند و یا از یک منبع گردآوری شده باشند. منبع این جملات می‌تواند یک کتاب باشد و یا می‌توانید از ویکی‌پدیا فارسی استفاده کنید. ولی برای بالا بردن تنوع داده ها، از ساختن جملات توسط خودتان بپرهیزید.
- دیتاست ای که هر فرد باید ارائه دهد علاوه بر فایل های صوتی باید شامل یک فایل csv به فرمت زیر باشد:

voice_filename	transcript	accent	gender	tone

  - voice_filename : نام فایل mp3. مربوط به جمله مورد نظر(به همراه پسوند mp3.)
  - transcript : متن جمله. با حروف فارسی و فرمت یونیکد
  - accent : لهجه گوینده. مقادیر ممکن: فارسی / ترکی / لری / کردی / عربی / بلوچ / لاری / خراسانی / گیلکی / مازنی / یزدی / اصفهانی

    در صورتی که مقدار مورد نظر شما برای لهجه در لیست بالا از قلم افتاده بود، می توانید به به دستیاران آموزشی اطلاع دهید تا در دیتاست نهایی در نظر گرفته شود.

  - gender : جنسیت گوینده. مقادیر ممکن: male/female
  - tone : لحن گوینده. مقادیر ممکن: normal/question/exclamatory/imperative/incomplete
- در نام گذاری فایل ها دقت زیر را داشته باشید:

- ابتدای نام فایل های صوتی باید شماره دانشجویی گوینده باشد. مثلا SID_voice1.mp3 که SID شماره دانشجویی شماست.
- فایل csv باید کنار فایل های صوتی قرار گرفته باشد و نام transcripts.csv داشته باشد.

در نهایت فایل های صوتی و فایل csv را در کنار هم zip کرده و در مکان قرار شده در ایلرن درس بارگذاری کنید.

- توجه داشته باشید که کیفیت صوت و متنی که جمع‌آوری می‌کنید، در ادامه برکیفیت مدل‌سازی شما تاثیر خواهد داشت. مثلا تعدادی از نکاتی که هنگام نگارش متن جملات باید مد نظر قرار دهید به این شرح است:
  - رعایت ملاحظات در نگارش متن، نظیر نوشتن آن ها به جای آنها (ترجیح جدا نویسی بر سر هم نویسی)
  - عدم استفاده از ی و ک عربی، تنوین و نیم فاصله
  - نکات دیگری که به ذهنتان می رسد.

از آن جا که داده جمع آوری شده توسط دانشجویان با هم ادغام خواهد شد، در صورت عدم رعایت برخی استاندارد های مشترک نظیر موارد ذکر شده، کار تمیز کردن داده اندکی سخت تر خواهد شد و دقت مدل ASR می تواند پایین بیاید.

مهلت ارسال داده ها **حداکثر تا پایان آذر ماه** است.

۲.۳– گروه‌بندی

شما می‌توانید به صورت انفرادی و یا گروه های حداکثر چهارنفره فعالیت داشته باشید. اسامی افراد گروه را یک نفر به نمایندگی **حداکثر تا پایان آذرماه** در محل تحویل گروه‌بندی در ایلرن ارسال کند. دقت کنید که با وجود گروه‌بندی بخش جمع‌آوری داده به صورت انفرادی باید انجام شود، ولی باقی قسمت ها به صورت گروهی انجام می شود.

۲.۴– گزارش اولیه

در این مرحله لازم است به صورت گروهی گزارشی تهیه کنید (حداقل دو صفحه) و در آن به موارد زیر بپردازید:

- توضیحی مختصر راجع به نحوه کار با داده های صوتی در الگوریتم یادگیری ماشین مانند نحوه تمیز کردن داده ها، ویژگی های معمول استخراج شده و نحوه normalization

- روش های پیاده سازی تسک ASR[4] (تبدیل اتوماتیک صوت به متن). گزارش شما باید هم روش های سنتی تر یعنی روش های آماری[5] و همینطور روش های جدیدتر انتها-به-انتها[6] را در بر داشته باشد. این دو مدل پیاده سازی ASR را از جنبه های میزان داده مورد نیاز، منابع پردازشی و همینطور دقت مقایسه کنید.
- به اختصار راجع به مفهوم و دلیل استفاده از fine tuning در آموزش شبکه های عصبی عمیق توضیح دهید. شما در بخش های بعدی پروژه از این تکنیک برای آموزش دادن مدل ASR خود استفاده خواهید کرد.

هدف این قسمت اینست که قبل از انجام پروژه، نسبت به داده ها و روش هایی که قرار است در طول پروژه از آن ها استفاده کنید، دید بهتری پیدا کنید. مهلت ارسال این گزارش **حداکثر تا پایان آذرماه** است.

۲.۵– تمیز کردن داده و استخراج ویژگی

حال پس از جمع‌آوری داده، لازم است با پردازش اولیه، آن ها را آماده برای مراحل بعدی پروژه کنید. در این مرحله دقت داشته باشید با توجه به نوع داده ی خود بهترین روش ها را برای استخراج و تمیز کردن انتخاب کرده تا بتوانید دقت بالاتری در مراحل بعدی بدست آورید. به طور مثال می توانید نویز داده های صوتی خود را کم کنید، و لازم خواهد بود برای بخش ASR پیش پردازش هایی روی نرخ فایل های صوتی و transcript ها انجام دهید.

به منظور استفاده از داده صوتی خام مراحل قبل، نیاز به استخراج ویژگی از داده و تبدیل آن‌ها به بردار ورودی مناسب برای مدل ها خواهید داشت. برای استخراج ویژگی می‌توانید از کتابخانه‌های معروفی که ویژگی‌های زمانی یا فرکانسی سیگنال‌های صوتی را در اختیار قرار می‌دهند و یا گزینه های ابتکاری ای که با الهام از روش های موجود به ذهن شما خطور می کند استفاده کنید. دقت کنید که ویژگی های استخراج شده، در قسمت طبقه بندی و خوشه بندی بکار خواهند رفت، فلذا با این دید به انتخاب ویژگی های مناسب اقدام نمایید. ضمنا در گزارش کار خود دلیل انتخاب ویژگی های استخراج شده را توضیح دهید.

دقت داشته باشید که اگرچه شبکه های عصبی عمیق امروزه به طور گسترده برای استخراج ویژگی و طبقه بندی به کار می روند، اما در این پروژه انتظار می رود که شما از روش های دیگری که فراگرفته اید استفاده کنید.

[4] Automatic speech recognition
[5] Statistical ASR methods
[6] End-to-end ASR pipeline

۲.۶– طبقه‌بندی

هدف ما در این قسمت، پیش بینی جنسیت گوینده با استفاده از ویژگی های استخراج شده از فایل های صوتی در قسمت قبلی است. در مراحل مختلف طبقه‌بندی، نکات زیر را در نظر داشته باشید:

۱. داده ها 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 جنسیت در هر دو دسته از داده ها، به یک میزان باشد. (اصطلاحا دیتاست شما نسبت به کلاس های مختلف نباید بایاس باشد)

۲. از دو روش جداگانه برای طبقه بندی استفاده کرده و نتایج را با هم مقایسه و تحلیل کنید. در این بخش کدها به تنهایی حائز اهمیت نیستند بلکه در کنار آن ها تحلیل نتایج وهمچنین مقایسه‌ی روش‌های مختلف، اهمیت ویژه ای دارد. می‌توانید برای طبقه‌بندی از هر یک از روش هایی که در درس فراگرفته اید استفاده کنید.

۳. در هرکدام از روش ها، برای بهبود نتایج خود می‌توانید از انواع تکنیک های Normalization، Dimensionality Reduction و Ensemble Learning نیز استفاده کنید. هنگام استفاده از هر یک از این تکنیک ها، نتایج را برای قبل و بعد از استفاده از آن ها به اختصار ذکر کنید، بگونه ای که تاثیر مثبت استفاده از آن ها مشهود باشد.

۴. برای هر طبقه‌بندی که مورد استفاده قرار می‌دهید، ماتریس آشفتگی، ROC Curve ، تحلیل میزان خطای هر کلاس و سایر مواردی که فکر می کنید می تواند بیانگر عمکرد مدل شما باشد را در گزارش خود بیاورید.

۲.۷– خوشه‌بندی

در اینجا همانند قسمت قبلی باید با استفاده از روش‌هایی که در درس آموخته‌اید، با انتخاب دو روش جداگانه، داده ها را خوشه‌بندی کنید. با استفاده از روش هایی مثل silhouette score، نمودار پراکندگی خوشه ها را به ازای تعداد خوشه های مختلف رسم کرده و تعداد خوشه مناسب را پیدا کنید. سپس خوشه‌بندی را به ازای تعداد خوشه مناسب که بدست آوردید و همچنین به ازای ۲ مقدار دلخواه دیگر برای تعداد خوشه، انجام دهید و نتایج حاصل را تحلیل و بررسی کنید. برای هر یک از تعداد خوشه های انتخاب شده، شباهت داده های درون یک خوشه و تفاوت بین خوشه ها و دلایلی که فکر می کنید برخی داده ها در یک خوشه قرار گرفته‌اند باید بررسی دقیق شوند. تحلیل و گزارش در این بخش از اهمیت بالایی برخوردار است.

۲.۸– پیاده‌سازی ASR

برای این قسمت، ابتدا باید کد مربوط به یک مدل ASR که در زمان مناسب در اختیار شما قرار می گیرد را کامل کنید. سپس با استفاده از داده‌های جمع‌آوری شده در قسمت قبل، مدل را آموزش دهید.

در این قسمت، انتظار می‌رود در گزارش خود میزان خطای مدل را بر اساس آماره WER[7] را ذکر کرده و نیز با استفاده از فرمت مناسب (نمودار یا جدول)، همبستگی احتمالی میزان خطا با ویژگی های هر صوت (جنسیت، لهجه و لحن) را تحلیل کنید. در صورت تمایل می توانید علاوه بر موارد ذکر شده، همبستگی میزان خطا را با ویژگی هایی نظیر سرعت صحبت کردن و نویز محیطی را نیز به صورت کیفی بررسی کرده و گزارش دهید.

[7] Word error rate

۳) فرمت گزارش کار و کد ها

۳.۱– گزارش کار

همانطور که قبلا هم گفته شد، علاوه بر کد درست، گزارش کار کامل و توضیح و تحلیل درست داده و نمودارها از اهمیت بسیار بالایی برخوردار است. گزارش کار منبع اصلی برای صحت سنجی روش های بکار رفته و سنجش تسلط شما به شمار می آید فلذا کد درست بدون گزارش کار مناسب معنایی ندارد. بنابراین سعی کنید تمام نکات قابل ذکر در انجام پروژه را در گزارش کار ذکر کنید.

دقت داشته باشید که گزارش کار اولیه که پیشتر ذکر شد، به عنوان مقدمه ای بر گزارش کار نهایی است و مهلت تحویل آن قبل از گزارش کار نهایی است. گزارش کار نهایی، شامل تمامی قسمت هایی می شود که پس از تحویل گزارش اولیه انجام می شوند و مهلت تحویل آن به همراه کد ها **حداکثر تا ۲۷ دی ماه است.**

در زیر تعدادی از نکاتی را متذکر می‌شویم که حتما باید در گزارش کار نهایی ذکر شود:

- توضیح مختصر درباره نحوه کارکرد و علت استفاده از روش‌هایی که برای قسمت پیش‌پردازش استفاده کردید
- توضیح مختصر درباره علت انتخاب مدل‌هایی که برای طبقه‌بندی و خوشه بندی استفاده کردید
- توضیح مختصر درباره علت و نتیجه استفاده از روش‌های تدریس شده در درس مانند روش‌های کاهش بعد و normalization در جهت بهبود کارایی مدل های طبقه بند و خوشه بند
- ذکر و تحلیل معیار هایی نظیر precision، recall، F1 score و غیره، برای هر کدام از مدل‌های طبقه بند که آموزش داده‌اید
- برای مدل های خوشه بند و به ازای هر تعداد خوشه، ذکر میزان پراکندگی درون خوشه ای و میان خوشه ای و همچنین تحلیل اینکه هر خوشه نماینده چه دسته ای از داده هاست (ویژگی های مشترک داده های درون هر خوشه)
- میزان خطای مدل ASR و بررسی همبستگی این خطا به ویژگی هایی که قبلا در قسمت مربوطه ذکر شد

گزارش کار میبایست مرتب بوده، منطبق با سوالات مطرح شده در شرح پروژه بخش بندی شده و تحلیل های دقیق داشته باشد. همچنین نوشتن گزارش کار با Latex تا پنج درصد نمره امتیازی دارد.

۳.۲– فرمت کد های تحویلی

کد ها باید در Jupyter notebook (.ipynb) و از پیش اجرا شده باشند. جدا سازی کد بخش های مختلف در نوتبوک با سلول ها و ترجیحا درج title، در فهم کد شما تاثیر بسیار مثبتی خواهد داشت. ضمنا تقسیم بندی زیر برای کد های تحویلی رعایت کنید:

- Cleaning_and_feature_extraction.ipynb

  تمیز کردن داده و استخراج ویژگی ها. این نوتبوک باید در انتها ویژگی های استخراج شده را در یک فایل csv ذخیره کند. آنالیز های دیگر این فایل را لود کرده و از ویژگی ها استفاده خواهند کرد.

- Classification.ipynb

  این نوتبوک ویژگی ها را لود کرده و طبقه بندی را انجام می دهد.

- Clustering.ipynb

  این نوتبوک ویژگی ها را لود کرده و خوشه بندی را انجام می دهد.

- ASR.ipynb

  یک نوتبوک که به صورت نیمه آماده به شما داده می شود و شما برای انجام ASR و همینطور تحلیل نتایج آن را کامل می کنید.

نوتبوک های هر بخش، باید کد تولید تمامی آنالیز های انجام شده در گزارش کار را دارا باشند.

۴) بارم بندی و تاریخ تحویل

نمره دهی	از ۱۰۰
گزارش اولیه	۱۰
جمع‌آوری داده	۱۰
پیش‌پردازش داده	۱۵
طبقه‌بندی	۱۵
خوشه‌بندی	۱۵
پیاده سازی ASR	۱۵
گزارش کار نهایی	۲۰

بخش	تاریخ تحویل
گزارش اولیه، جمع آوری داده، گروه بندی	پایان آذرماه
کد ها و گزارش نهایی	۲۷ دی ماه

۵) نکات پایانی

- هیچگونه شباهتی در انجام پروژه بین افراد در بخش های فردی و گروه های مختلف در بخش گروهی پذیرفته نمی شود. در صورت کشف هرگونه تقلب، مطابق قوانین درس با افراد خاطی برخورد خواهد شد.
- استفاده از مراجع در صورت ارجاع به آن ها 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 کامل و دقیق این توضیحات، در صورتی که سوالی در مورد پروژه داشتید بهتر است **در فروم درس در ایلرن مطرح کنید** تا بقیه نیز بتوانند از آن استفاده کنند. در غیر این صورت یا در گروه تلگرامی مطرح کنید یا به طراحان پروژه ایمیل بزنید.
- راه ارتباطی با دستیاران آموزشی مسئول پروژه:

رامین طوسی: ramin.tsi@gmail.com

مریم دادخواه: m.dadkhah99@gmail.com

تابان سلیمانی: taabansoleymani@gmail.com

سبحان نیلی: s.nili80@outlook.com