دانشگاه تهران

پردیس دانشکده های فنی

دانشکده برق و کامپیوتر

پروژه نهایی درس یادگیری ماشین

دکتر اعرابی - دکتر ابوالقاسمی

پاییز ۱۴۰۰

۱ مقدمه

به طور کلی هدف از این پروژه انجام صفر تا صد یک پروژه واقعی یادگیری ماشین است. در این پروژه شما با چالش هایی روبه رو خواهید شد که بعضاً در دیگر پروژه های دانشگاهی یادگیری ماشین با آن روبه رو نبوده اید. به عنوان مثال در اکثر پروژه ها داده(بعضاً پیش پردازش شده) در اختیار شما قرار می گرفت اما شما در این پروژه بخش جمع آوری داده را خواهید داشت. امید است در پایان پس از انجام پروژه توانایی شما به عنوان یک مهندس در ارتباط با جمع آوری دادها، تحلیل و پردازش آن ها و طبقه بندی و خوشه بندی افزایش پیدا کرده و با دید وسیع تری نسبت به حوزه یادگیری ماشین در این راه قدم بگذارید.

۲ تعریف مسئله و شرح پروژه

۱.۲ تعریف کلی مسئله

هدف نهایی این پروژه طبقه بندی و خوشه بندی آهنگ های محلی در ۵ دسته بندی لری، کردی، ترکی، گیلکی و بندری است. ابتدا از شما خواسته شده است در زمان مشخص شده داده ها را برای هر گروه از آهنگ های محلی جمع آوری کنید. سپس توضیح نسبتا کوتاهی در ارتباط با چالش ها و تفاوت های این نوع داده گرد آوری می کنید. سپس پس از تمیز سازی داده ها آن ها را خوشه بندی و طبقه بندی می کنید. در انتها هم لازم است گزارشی کامل از تحلیل نتایج خود بنویسید.

۲.۲ جمع آوری داده

در مرحله اول لازم است هر نفر به صورت انفرادی برای هر گروه مشخص شده ۱۰ آهنگ جمع آوری کند. (یعنی هر نفر ۵۰ آهنگ) لازم به ذکر است که آهنگ ها به هیچ عنوان نباید تکراری باشند. برای جلوگیری ازین کار برای شما یک online sheet قرار داده می شود که هر فرد آهنگ های جمع آوری شده خود را پیش از قرار دادن چک کند که آیا فرد دیگری قبل از او آن را قرار داده است یا خیر. در صورت عدم توجه و ارسال داده های تکراری نمره ی بخش جمع آوری داده به طور کامل به شما تعلق نمی گیرد. برای همین منظور بهتر است زودتر به گردآوری داده اقدام کنید.:))

برای یکدست بودن داده ها همگی آهنگ ها را در فرمت mp3 128kb ارسال کنید. در صورتی که بیش از ۱۰ آهنگ برای هر گروه ارسال کردید نمره ی اضافه به شما تعلق می گیرد.

دقت داشته باشید مهلت ارسال داده ها حد اکثر تا پایان روز جمعه مورخ ۱۴ آبان ماه ساعت ۲۳:۵۵ می باشد.

۳.۲ گزارش اولیه

در این مرحله لازم است به صورت گروهی گزارشی تهیه کنید (حداقل دو صفحه) و در آن راجع به آهنگ های محلی اطلاعاتی جمع آوری کنید. در این گزارش شما باید نکاتی را در ارتباط با تفاوت های این آهنگ ها به همراه چالش ها و سختی های مسئله ذکر کنید.

مهم این است با جست جو قبل از انجام پروژه نسبت به داده هایی که قرار است با آن ها کار کنید، دید پیدا کنید. لزومی ندارد این بررسی ها فنی و از دید یادگیری ماشین انجام شود بلکه باید روان بوده و یک کاربر عادی هم بتواند با خواندن متن شما آن تفاوت هارا درک کرده و با مسئله آشنا شود.

مهلت ارسال این گزارش حداکثر تا پایان روز جمعه مورخ ۲۱ آبان ماه ساعت ۲۳:۵۵ می باشد.

۴.۲ گروه بندی

شما میتوانید به صورت انفرادی و یا گروه های حداکثر تا ۴ عضو فعالیت داشته باشید. فعالیت به صورت گروهی نمره مثبت دارد.

اسامی افراد گروه را یک نفر به نمایندگی در ابتدای گزارش اولیه نوشته و حداکثر تا پایان روز جمعه مورخ ۲۱ آبان ماه ساعت ۲۳:۵۵ ارسال کند.

۵.۲ تمیز سازی داده ها

حال پس از جمع آوری داده ها لازم است با پردازش اولیه داده ها آن ها را آماده برای مراحل بعدی پروژه کنید. در این مرحله توجه داشته باشید با توجه به نوع داده های خود بهترین روش ها را انتخاب کرده تا بتوانید دقت بالاتری در مراحل بعدی بدست آورید.

۶.۲ طبقه بندی

در این مرحله باید با استفاده از روش های یادگیری ماشین که در طول ترم آموخته اید طبقه بندی را انجام دهید. در این مرحله حداقل از ۳ روش جداگانه استفاده کرده و نتایج را باهم مقایسه و تحلیل کنید. دقت داشته باشید چیزی که در این جا اهمیت بسیار بالایی دارد صرفا کد نیست بلکه تحلیل نمودار ها داده ها و مقایسه ی روش های مختلف است.
توجه کنید در این قسمت برای استفاده از شبکه های عصبی فقط قادر به استفاده از شبکه ی MLP[1] هستید و از RNN[2] و CNN[3] نمی توانید استفاده کتید. دقت کنید در این مرحله باید داده ها را به دو گروه آموزش و تست تقسیم کنید و پس از آموزش شبکه بر روی داده های آموزش و سپس تست آن بر روی داده های تست نمودار های دقت و loss را رسم کرده و تحلیل کنید.

۷.۲ خوشه بندی

در اینجا هم همانند قسمت قبلی باید با استفاده از روش هایی که در درس آموخته اید با تعداد cluster های متفاوت از ۱ تا ۵، با انتخاب حداقل ۲ روش خوشه بندی داده ها را دسته بندی کنید. نکته مهم در این قسمت این است که به ازای هر تعداد خوشه تحلیل کنید که داده هایی که در هر خوشه قرار میگیرند چه شباهت هایی به همدیگر دارند و به چه دلیل در یک خوشه قرار گرفته اند. تحلیل داده ها در گزارش کار در این قسمت نقش عمده ای دارد. :))

---

[1] Multi Layer Perceptron
[2] Recurrent Neural Network
[3] Convolutional Neural Network

۳ بارم بندی

نمره دهی	از ۱۰۰
گروه بندی	۵
جمع آوری داده ها	۱۵
پیش پردازش داده ها	۲۰
خوشه بندی	۲۰
طبقه بندی	۲۰
گزارش کار نهایی	۲۰

۴ گزارش کار

همان طور که قبلا هم گفته شد، علاوه بر کد درست، گزارش کار مفصل و توضیح و تحلیل درست داده ها و نمودار ها 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 انتظار می رود دو گزارش کار تحویل داده شود. گزارش کار اولیه حداقل دو صفحه ای برای توضیحات تفاوت ها و چالش ها به همراه اسامی گروه که مهلت آن تا پایان جمعه ۲۱ آبان می باشد. و گزارش نهایی هم همراه با کد در انتهای هفته ۱۶ ام است. در زیر نکاتی را متذکر می شویم که حتما باید در گزارش کار نهایی ذکر شود:

- روش هایی که برای قسمت پیش پردازش استفاده کردید و توضیح مختصر نحوه کارکرد هر کدام از آن ها
- تحلیل هایی که در بخش های خوشه بندی با تعداد خوشه هایی متفاوت کردید
- مدل هایی که برای طبقه بندی استفاده کردید را مختصراً توضیح دهید و درباره چرایی انتخاب آن ها نیز توضیح دهید
- برای هر کدام از مدل های آموزش داده شده مقادیر , precision , recall , f1-score accuracy را گزارش دهید.

گزارش کار را مرتب بنویسید. تحلیل ها کامل و دقیق باشد.

نوشتن گزارش کار با LaTeX نمره امتیازی دارد. (تا ۱۰ درصد)

۵ نکات پایانی

- طبق تقویم آموزشی موعد تحویل نهایی تا انتهای هفته ۱۶ ام درس خواهد بود. با توجه به فشردگی در هفته های آخر با برنامه ریزی درست سعی به انجام پروژه به بهترین نحو ممکن بکنید.
- هیچگونه شباهتي در انجام این پروژه بین افراد مختلف پذیرفته نمي 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 بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کننده، در غیر این صورت یا در گروه تلگرامی مطرح کنید یا به طراحان پروژه ایمیل بزنید.

- در ارسال داده ها، آهنگ های یکسان با خواننده های متفاوت، داده های غیر مشابه در نظر گرفته می شود.

- برای قسمت امتیازی حداکثر تا ۱۲ آهنگ در هر بخش ارسال کنید.(اگر تا کنون تعداد بیشتری آهنگ ارسال کرده اید، به دلخواه حذف کنید)

Elahe.bvakili97@ut.ac.ir
Maryam.karimi7653@gmail.com
Setarehsoltanieh78@gmail.com