به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

هوش مصنوعی، ترم بهار ۹۷-۹۸

پروژه دسته‌بندی، مهلت: دوشنبه ۲۳ اردیبهشت

## مقدمه

در این پروژه هدف آشنایی با الگوریتم های مختلف طبقه بندی و تکنیک های پردازش تصویر است. در این پروژه شما مجازید از کتاب خانه های آماده استفاده کنید و تمرکز آن بر آشنایی با مفاهیم و تکنیک های گفته شده است. تضمینی وجود ندارد که چه بخشی از این موارد در کلاس درس گفته می شود. خودتان روی آن ها کار کنید و در صورت نیاز از استاد راجع به تکنیک های پردازش تصویر سوال کنید.

این پروژه شامل دو بخش است.در بخش اول پروژه مرحله به مرحله الگوریتم های زیر را انجام داده و کد را به همراه توضیح هر سوال (با ذکر شماره) در گزارش کار ثبت کنید. توجه کنید که نمودار ها هم به همراه برچسب هایشان ثبت شده باشند.

توجه کنید بخش زیادی از نمره پروژه مربوط به گزارش کار است و مابقی آن با توجه به رنکینگ دقت پروژه ها نسبت به یکدیگر محاسبه می شود. اگر بتوانید در قسمت دوم پروژه به دقت خیلی خوبی برسید تا مقدار قابل توجهی از پروژه نمره امتیازی می گیرید.

### توضیح مسئله

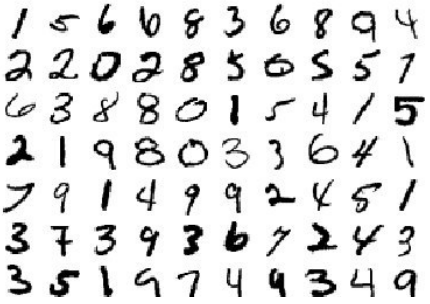
برای انجام این پروژه از کتابخانه scikit-learn استفاده کنید. این کتابخانه‌ی یادگیری ماشین برای زبان پایتون به صورت رایگان در دسترس است و دسته‌بندهای مورد نیاز شما در این کتابخانه موجود است.


- کلیه داده های مورد نیاز در یک پوشه کنار پروژه قرار گرفته است.
- در ادامه برای محاسبه دقت دسته‌دسته بندها از فرمول زیر استفاده کنید:

$$Accuracy = \frac{\text{num\_of\_correct\_predicted\_labels}}{\text{num\_of\_total\_labels}}$$

- **بخش اول(MNIST)**

مجموعه داده MNIST یک مجموعه‌ی بزرگ (شامل ۶۰۰۰۰ داده training و ۱۰۰۰۰ داده test) از ارقام انگلیسی دست نویس است. هم‌چنین در این مجموعه، برچسبی برای هر تصویر وجود دارد که بیانگر این است که هر تصویر نمایان گر چه رقمی است. اندازه ی هر تصویر در این مجموعه ۲۸*۲۸ پیکسل می باشد.



تصویر ۱. نمونه داده‌های MNIST

داده‌هایی که در این بخش در اختیار شما قرار گرفته برای هریک از مجموعه‌های train و test شامل دو فایل label و data است، که در فایل data در هر سطر ۷۸۴ = ۲۸*۲۸ عدد بین ۰ تا ۲۵۵ قرار دارد که هر عدد بیان کننده‌ی میزان روشنایی هر پیکسل در تصویر می‌باشد(۲۸ عدد اول هر سطر مربوط به پیکسل‌های سطر اول تصویر، اعداد ۲۹ تا ۵۶ هر سطر مربوط به پیکسل‌های سطر دوم تصویر و .. می‌باشد.). در فایل label برچسب نظیر هر یک از سطرهای فایل data (تصاویر) قرار گرفته است.

۰.یک نمونه از داده معادل رقم یکان شماره دانشجویی خود را در مجموعه داده train پیدا و آن را نمایش (visualize) دهید.

در این مسئله باید به کمک دسته‌بندهایی که در ادامه آمده است و مجموعه داده‌ی آموزش(train) که به شما داده شده است، بتوانید نمونه داده های جدید را تشخیص دهید.

**K-Nearest Neighbors**

۱. k-nearest neighbors را به اختصار توضیح دهید. (در ۳ تا ۴ خط)
۲. این الگوریتم را بر روی داده mnist پیاده سازی کنید و درصد دقت برای داده های train و test ثبت کنید.(در صورتی که اجرای این الگوریتم خیلی طولانی می شود، می توانید آن را بر روی بخش کوچکی از داده های تست (مثلا 50 مورد) انجام دهید.)
۳. نمودار دقت داده train و test را بر اساس بازه ای از مقادیر مختلف پارامتر neighbors_n را رسم کرده و نقطه بهینه آن را ثبت کنید.
۴. تغییرات خطا بر روی داده train در نمودار بالا چگونه است؟چرا؟
۵. یک عدد دلخواه را انتخاب کرده و با استفاده از این کتاب خانه دیتاهایی که در همسایگی آن قرار می گیرند را رسم کنید.
۶. یک مورد از مشکلات اصلی این الگوریتم را توضیح دهید؟

### درخت تصمیم (Decision Tree)

۷. درخت تصمیم را به اختصار توضیح دهید.
۸. مانند قسمت قبل الگوریتم درخت تصمیم را بر روی دیتاست اجرا کرده و دقت آن را ثبت کنید.
۹. پارامتر حداکثر ارتفاع درخت را tune کنید. یعنی درصد خطارا بر اساس مقادیر مختلف آن به دست آورده. نمودار آن را بکشید و نقطه بهینه آن را انتخاب کنید.(در هنگام tune کردن یک پارامتر، باید پارامترهای دیگر ثابت نگه داشته شوند.)
۱۰. درخت را به گونه ای مصور (visualize) کرده و آن را نمایش دهید.
۱۱. به نظر شما درخت تصمیم چه موقع اورفیت می کند؟ درستی ادعای خود را با تغییر دادن پارامتر ها بررسی کنید و نتایج آن را ثبت کنید.

### جنگل تصادفی (Random Forest)

۱۲. الگوریتم جنگل را به اختصار توضیح دهید.
۱۳. مساله را با جنگل تصادفی (Random Forest) حل کنید و دقت آن را برای داده های test و train ثبت کنید.
۱۴. مقدار پارامتر ارتفاع درخت را مانند قسمت قبل بر اساس مقادیر مختلف تحلیل کنید و نمودار آن را بکشید.
(مقدار سایر hyper parameter ها را روی یک مقدار مناسب ثابت نگهدارید.)
۱۵. سوال قبل را برای پارامترهای تعداد درختها، min_samples_split و max features تکرار کنید. همچنین هر یک از این پارامترها را به اختصار توضیح دهید.
۱۶. به نظر می رسد در حالتی که پارامتر estimators_n در رندوم فورست برابر یک است، رندوم فورست معادل درخت تصمیم می شود. آیا این ادعا درست است؟ آن را با مقایسه خطای آن دو بررسی کرده و علت آن را توضیح دهید.

### رگرسیون لجستیک

۱۷. رگرسیون لجستیک را به اختصار توضیح دهید.
۱۸. مسئله را با رگرسیون لجستیک حل و دقت آن بر روی داده های test و train را بدست آورید.

### K-Means

۱۹. K-Means را به اختصار توضیح دهید.
۲۰. داده های تست را با این روش دسته بندی (تعداد دسته‌ها را ۱۰ بگیرید) کنید و مرکز هر دسته را رسم (visualize) کنید.
۲۱. به کمک دسته های بدست آمده، عدد هر یک از داده های تست را تشخیص دهید و دقت آن را بدست آورید.
۲۲. تعداد دسته ها را افزایش دهید و مجددا مرکز هر دسته را رسم کرده و دقت را اندازه گیری کنید. تاثیر تغییر تعداد دسته‌ها بر دقت دسته‌بندی را بررسی کنید و نتایج بدست آمده را تحلیل کنید.
۲۳. به نظر می رسد مبنای عملکرد الگوریتم k-means و knn مشابه هم هست. به نظر شما کدام یک از نظر زمان پاسخ گویی به کوئری های جدید بهتر است، چرا؟

### جمع بندی انواع الگوریتم های دسته‌بندی

۲۴.بر روی داده های تست کدام الگوریتم بهترین تخمین را به شما می دهد؟ مشخصات آن را ثبت کرده و دقت را بیان کنید.
۲۵. یکی از معیارهای بررسی نتایج الگوریتم confusion matrix است. کانفیوژن ماتریکس را برای الگوریتمی که بهترین نتیجه را به شما داده است، رسم کنید.
۲۶. یک داده که در بعضی از دسته‌بندها درست و در برخی اشتباه دسته بندی شده است را پیدا کنید. و آن را رسم کنید.
(مشخص کنید در کدام دسته‌بندها درست و در کدام ها اشتباه دسته بندی شده است.)

**کاهش ابعاد بردار ویژگی‌ها**

یکی از روش‌های کاهش ابعاد بردار ویژگی‌ها و حذف ابعادی که اطلاعات کمتری در دسته‌بندی به ما می‌دهند استفاده از PCA است.

۲۷. در مورد PCA تحقیق کنید و به صورت مختصر آن را توضیح دهید و دلیل استفاده از آن را بیان کنید.(نیاز به بررسی روابط ریاضی آن نیست.)

۲۸. به کمک PCA بردار ویژگی های جدید را بدست آورید و الگوریتم K-means را روی آن اجرا کنید. نتایج را با حالت اولیه مقایسه کنید.

- **بخش دوم(CIFAR-10)**

مجموعه داده CIFAR-10 یک مجموعه ۶۰۰۰۰تایی از تصاویر رنگی با ابعاد ۳۲*۳۲ پیکسل است که در ۱۰ دسته، دسته‌بندی شده اند. در اینجا نیز مانند بخش قبل باید بتوانید با مشاهده‌ی داده های train، برچسب داده های تست را تشخیص دهید.


تصویر۲. نمونه داده‌های CIFAR-10

داده‌هایی که در این بخش در اختیار شما قرار گرفته برای هریک از مجموعه‌های train و test مانند بخش قبل شامل دو فایل label و data می‌باشد و برای real_test فقط شامل data است، که در فایل data در هر سطر 3072 = ۳*۳۲*۳۲ عدد بین ۰ تا ۲۵۵ قرار دارد که در آن هر سه عدد متوالی بیان کننده‌ی مقادیر RGB هر پیکسل در تصویر می‌باشد(۹۶ عدد اول هر سطر مربوط به پیکسل‌های سطر اول تصویر، اعداد ۹۷ تا ۱۹۲ هر سطر مربوط به پیکسل‌های سطر دوم تصویر و .. می‌باشد.). در فایل label برچسب نظیر هر یک از سطرهای فایل data (تصاویر) قرار گرفته است.

در این بخش شما باید به کمک یکی از دسته‌بندهای بالا که به دلخواه انتخاب می‌کنید، سعی کنید این مسئله را حل کنید و به دقت بالاتر برسید برای اینکار می توانید از تکنیک های feature engineering استفاده کنید. سوالات این قسمت بخشی از نمره‌ی شما را تشکیل می‌دهد اما بیشتر نمره‌ی این بخش براساس میزان دقت دسته‌بندی شما محاسبه می‌شود.

برای حل این مسئله از موارد زیر می توانید استفاده کنید.

- Turn the images to grayscale
- PCA
- Random projection
- Augmentation

۱. هر مورد را توضیح دهید و به طور خلاصه بگویید که هرکدام چرا احتمال دارد باعث افزایش دقت مسئله شود.
۲. تکنیک های بالا را استفاده کنید تا دقت بهتری از مسئله کسب کنید. هر کاری کردید - صرف نظر از بهبود بخشیدن یا نبخشیدن - نتیجه آن را در گزارش کار خود ثبت کنید.
۳. برچسب داده های CIFAR_real_test.csv را با دسته‌بندی که بهترین دقت را به شما داده است، پیشبینی کنید و در فایل <student_id>.csv (شماره دانشجویی) با فرمت زیر قرار دهید.(فایل سابمیشن نمونه در کنار پروژه قرار گرفته است.)

```
id, predict
1, car
2, bird
...
```

★ درصورتی که بتوانید به درصد بالایی از دقت در تشخیص تصاویر برسید، نمره‌ی امتیازی قابل توجه‌ای به شما تعلق می گیرد.

## گزارش کار

- درصد بالایی از نمره‌ی این فاز از پروژه برای گزارش کار است، سعی کنید گزارش کار مختصر و کاملی ارائه دهید.
- در گزارش کار خود به تمامی سوال هایی که در صورت پروژه آمده است پاسخ دهید.

## نکات پایانی

- ۷۰ درصد نمره را قسمت MNIST و ۳۰ درصد نمره را CIFAR-10 تشکیل می‌دهد.
- بیشتر نمره بخش اول با توجه به گزارش کار شما مشخص می شود. همچنین نمره‌ی بخش دوم بر اساس دقت دسته‌بند شما ارزیابی می شود.
- شدیدا توصیه می شود پروژه رو با ژوپیتر نوتبوک انجام داده و برای تهیه گزارش کار آن را annotate کنید. گزارش کار را به صورت یک فایل پی دی اف یا اچ تی ام ال به همراه کد ها و خروجی csv لازم اپلود کنید.
- مطمئن باشید فایل های ژوپیتر قابل اجرا شدن مجدد هستند.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه نیز از آن استفاده کنند، در صورت نیاز می توانید از طریق ایمیل های زیر با ما در ارتباط باشید.

emad.jabbarnk@gmail.com
armin.zirak97@gmail.com

- درصورتی که نیاز به پرسش سوالی به صورت حضوری دارید، می‌توانید با ما مطرح کنید.