به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر

# هوش مصنوعی

## پروژه چهارم - یادگیری ماشین

مهلت تحویل: یک‌شنبه، ۲۸ اردیبهشت
طراحان: علی الهی، بهزاد شایق

بهار ۹۹

## هدف پروژه:

هدف این پروژه آشنایی با روش‌های یادگیری‌ماشین[۱] به کمک کتابخانه SciKit-Learn است. این پروژه در سه فاز تعریف شده است. در فاز صفر، به بررسی مجموعه داده‌ها[۲] می‌پردازید. در فاز اول، با استفاده از چند classifier تعریف‌شده در کتابخانه SciKit-Learn مدل‌هایی را پیاده‌سازی و بهینه‌سازی خواهیدکرد. نهایتا در فاز دوم، با استفاده از مدل‌های بهینه فاز اول، به پیاده‌سازی چند روش یادگیری‌گروهی[۳] و تحلیل نتایج حاصل می‌پردازید.

## معرفی مجموعه داده:

مجموعه داده‌ای که در اختیار شما قرار دارد، شامل ویژگی[۴]های مشتریان چند فروشگاه می‌باشد و داده هدف[۵]، بازگشت یا عدم بازگشت مشتری به فروشگاه است.

## فاز صفر:

داده‌های خام ورودی را به مجموعه‌ای از ویژگی‌های قابل پردازش تبدیل‌کنید. information gain را برای ویژگی‌ها محاسبه کنید و نمودار gain بر حسب ویژگی ها را رسم و سپس توجیه کنید. ( برای محاسبه information gain می‌توانید از متد mutual_info_classif از کتابخانه SciKit-Learn استفاده کنید.)

## فاز اول:

در این فاز از پروژه، سه مدل بر پایه classifier های K Nearest Neighbours، Decision Tree و Logistic Classifier با کمک کتابخانه SciKit-Learn پیاده‌سازی کنید. برای بررسی دقت از معیارهای Accuracy، Precision و Recall استفاده‌کنید. نهایتا در این بخش، باید مدل‌های بهینه ای از این classifier ها داشته باشید. بهینه به این معنی است که پارامترها باید به گونه‌ای تنظیم شوند[۶] که هر مدل، به بیشترین دقت رسیده و همچنین، overfitting اتفاق نیفتد. تنظیم تعداد همسایه‌ها (n_neighbor) برای الگوریتم KNN و عمق ماکسیمم (max_depth) برای الگوریتم Decision Tree کفایت می‌کند.

۱. برای پیاده‌سازی classifier ها، ابتدا داده ها را به دو بخش test و train تقسیم کنید.

---

۱ Machine learning
۲ Data set
۳ Ensemble learning
۴ Features
۵ Target or Label
۶ Hyperparameter tuning

۲. دقت هر مدل را بر اساس معیارهای Accuracy، Precision و Recall برای داده‌های test و train اندازه‌گیری کنید.

۳. برای مدل‌های Decision Tree و KNN، نمودار دقت هر مدل (برای داده های test و train) را بر حسب hyperparameter ها رسم کنید و overfitting را برروی این نمودار بررسی و تحلیل کنید. (بهتر است نمودار های مربوط به train و test را در یک plot رسم کنید.)

۴. توضیح دهید، برای پیاده‌سازی classifier ها چه پیش‌پردازش[۷]هایی بر داده‌ها از جمله داده‌های طبقه‌ای[۸]، تاریخ‌ها و داده‌های پیوسته انجام دادید.

## فاز دوم:

یادگیری‌گروهی به این معناست که از تجمیع[۹] نتایج حاصل از تعدادی مدل، پیشبینی نهایی را انجام دهیم. در این فاز به پیاده‌سازی و تحلیل چند روش یادگیری‌گروهی می‌پردازیم.

۱. با استفاده از مدل های KNN و Decision Tree در فاز اول، روش Bagging را پیاده‌سازی کرده و معیارهای دقت را محاسبه کنید. (در این پیاده‌سازی، ماکسیمم تعداد پارامترها و تعداد ویژگی‌های مورد استفاده را برابر ۰.۵ قرار دهید.)

۲. به دلیل آنکه پیاده‌سازی Bagging (یا Bootstrap aggregating) بر مدل Decision Tree بطور معمول با افزایش دقت خوبی همراه است، از این پیاده‌سازی با عنوان Random forest یاد می‌شود. با کمک کتابخانه SciKit-Learn این مدل را پیاده‌سازی کنید. تاثیر حداقل دوتا از hyperparameter ها را بر مدل بررسی کرده و معیارهای دقت را محاسبه کنید.

۳. تاثیر Bagging را بر overfitting بررسی کنید و نتایج حاصل را ذکر کنید. (برای مشاهده این تاثیر می‌توانید از اجرای Bagging بر یک مدل با overfitting بالا کمک بگیرید.)

۴. Bootstrapping چیست و چه تاثیری بر واریانس و بایاس دارد؟ (اطلاعات بیشتر)

۵. یکی از روش‌های یادگیری‌گروهی Hard-Voting است. در این روش، پیشبینی نهایی، کلاسی است که بیشترین رای را بین مدل‌ها داشته باشد. با استفاده از سه مدل بهینه شده در فاز اول، یک مدل بر پایه Hard-Voting classifier پیاده‌سازی کنید و معیارهای دقت را اندازه‌گیری کنید.

۶. شباهت پاسخ مدل‌های فاز اول را با معیاری که از نظر شما مناسب است، بسنجید. (به عنوان مثال می‌توانید درصد تشابه دو به دو پاسخ مدل ها را محاسبه کنید.) نهایتا با تحلیل نتایج خود، دلیل موفقیت یا شکست مدل‌های یادگیری گروهی را بیان کنید

---

۷ Data preprocessing
۸ Categorical data
۹ Aggregating

## نکات پایانی:

- مقدار مطلوب برای Accuracy، Precision و Recall به ترتیب ۷۰، ۷۵ و ۸۰ درصد است. توجه داشته باشید که کسب این دقت، تنها توسط یکی از مدل‌هایی که در پروژه پیاده‌سازی کردید کافی است و سایر مدل‌ها باید صرفا دقت معقولی داشته باشند.
- در تمامی بخش‌های پروژه، استفاده از کتابخانه SciKit-Learn مجاز است. دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده بطور مثال نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را بطور **خلاصه و در عین حال مفید** در گزارش خود ذکر کنید.
- دقت کنید که مشاهده عدم پیشرفت چشمگیر در روش های یادگیری‌گروهی قابل انتظار است و هدف، تحلیل نتایج و بررسی دلایل موفقیت یا شکست مدل‌ها است.
- نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA4_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل تمامی پیاده‌سازی ها در فایلی با نام code.py و گزارش در فایلی با فرمت PDF باشد. در صورتی که از jupyter-notebook استفاده می‌کنید، نیازی به ارسال جداگانه کدها و گزارش نیست و هردو را می‌توانید در یک فایل Notebook به همراه خروجی html آن ارائه دهید.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت توسط ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.