یادگیری ماشین ۷۱۷-۴۰ (گروه دوم) **نیم‌سال اول ۹۵-۹۴**

فاز اول پروژه **مدرس: دکتر سلیمانی**

موعد تحویل: ۱۰ آذر ۹۴ **طراح پروژه: شجاعی، اوصیا**

هدف پروژه

در سال‌های اخیر تحلیل داده اهمیت بسیاری پیدا کرده است و بسیاری از شرکت‌های نوپا نیز بر همین اساس بنا شده‌اند. در واقع با رشد روزافزون سخت‌افزار، امکان پردازش و ذخیره‌سازی حجم زیادی از اطلاعات فراهم شده است و انسان به تنهایی نمی‌تواند اطلاعات مهم این داده‌ها را استخراج کند. به همین دلیل است که تحلیل‌گران داده سعی می‌کنند با استفاده از این امکانات به بهترین نحو از داده‌ها استفاده کنند و اطلاعات نهفته در آن‌ها را کشف کنند. هدف یادگیری ماشین نیز استخراج اطلاعات و الگوها از تجربیاتی است که در قالب داده در اختیار قرار گرفته است. شما تاکنون با تعدادی از روش‌های این کار آشنا شده‌اید و تئوری آن‌ها را فراگرفته‌اید. در این پروژه قرار است همین مطالب را در عمل ببینید تا شهود بهتری نسبت به آن‌ها داشته باشید و علاوه بر این با چالش‌هایی که در عمل با آن‌ها مواجهید آشنا شوید. مثلا یکی از چالش‌ها در آوردن داده‌ها به فرمی قابل قبول برای الگوریتم‌های یادگیری است؛ زیرا در بسیاری از موارد داده‌ها به شکلی نیستند که بتوانیم به راحتی از آن‌ها استفاده کنیم.

پیاده‌سازی

برای پیاده‌سازی بخش‌های مختلف می‌توانید از هر کتاب‌خانه‌ای استفاده کنید. قسمتی از نمره‌ی شما به انتخاب محیط پیاده‌سازی مربوط می‌شود. بنابراین پیش از این‌که اجرای پروژه را شروع کنید باید یک تحقیق نسبتا جامع از کتابخانه‌های معروف داشته باشید، خوبی‌ها و بدی‌های کتاب‌خانه‌هایی که دیده‌اید را در گزارشتان ذکر کرده و یکی را برای اجرای این فاز پروژه انتخاب کنید.

مجموعه دادگان

مجموعه‌ی دادگان آموزش شامل ۲۲۰۰۰ داده‌ی ۴۰۰ بعدی است که از ۵۰۰۰ تای آن‌ها به عنوان داده‌ی validation استفاده می‌شود. تقسیم‌بندی داده‌ها قبلاً انجام شده و در آدرس‌های xtrain و xval در اختیار شما قرار گرفته است. همچنین مجموعه‌ی دادگان تستی شامل ۳۰۰۰ وجود دارد که در آدرس xtest موجود است. مقادیر درخواستی در قسمت‌های بعد، برای xtrain و xval در آدرس‌هایی با نام‌های با معنی وجود دارد. برای xtest شما باید جوابهایتان را در فایل‌های متنی با نام مشابه ارسال کنید؛ به این صورت که هر سطر حاوی مقدار پیش‌بینی شده توسط شما برای هر یک از داده‌های درون xtest خواهد بود.

معرفی پروژه

در هر قسمت تعدادی روش اولیه برای پیاده‌سازی انتخاب شده است. در این روش‌ها شما باید با پیدا کردن تنظمیات بهینه بهترین نتیجه را برای هر روش به دست آورید. نتیجه‌ی شما در هر روش با نتیجه‌ی دیگر افراد مقایسه می‌شود.

پس از به دست آمدن نتایج برای روش‌های گفته شده، نوبت به شما می‌رسد که روشی برای حل مساله پیشنهاد کنید. روش پیشنهادی شما می‌تواند اصلا ربطی به روش‌های گفته شده در صورت سوال نداشته باشد و یا یکی از آن‌ها باشد. در این بخش تلاشتان را بکنید تا به بهترین نتیجه برای حل مساله برسید؛ زیرا در نهایت نتایج نهایی روش‌های افراد مختلف با هم مقایسه خواهند شد و قسمتی از نمره‌ی شما بر این اساس خواهد بود.

جزئیات نمره‌دهی پروژه در فایل دیگری آماده شده است و در اختیارتان قرار خواهد گرفت.

پیش‌پردازش و تنظیمات

در تمامی قسمت‌ها پیشنهاد می‌شود، تنظیمات مختص آن بخش روی داده‌ها داشته باشید. پیش‌پردازش خوب ممکن است تاثیر خیلی زیادی روی نتیجه‌ی نهایی داشته باشد. سعی کنید مواردی مشابه موارد زیر را در بخش تنظیمات درنظر داشته باشید:

- انتخاب توابع پایه مناسب $\boldsymbol{\phi}(\boldsymbol{x})$ (یا تابع کرنل مناسب)
- استفاده از روش‌های انتخاب ویژگی[1]
- استفاده از انواع مختلف منظم‌ساز مثل l1 و l2 و تنظیم وزن جمله‌ی منظم‌ساز

می‌توانید با استفاده از روش‌هایی مثل cross validation بهترین مدل انتخابی را برای هر قسمت پیدا کنید.

تحویل‌دادنی‌ها

- تمامی کدهایی که توسط خودتان پیاده‌سازی شده‌اند.
- موارد پیش‌بینی شده توسط شما برای داده‌های تست، در قالبی که در هر قسمت گفته شده است.
- مستند پروژه شامل
  - معرفی کتاب‌خانه‌های مورد استفاده، دلیل انتخاب آن‌ها و نحوه‌ی نصب آن‌ها
  - مستند اجرای کدها
  - شرح نحوه‌ی انجام تنظیمات برای هر قسمت
  - شرح معیارهای ارزیابی انتخاب شده و جواب‌های گرفته شده بر اساس آن‌ها در هر قسمت روی داده‌های آموزش و validation.

[1] Feature Selection

بخش‌های مختلف این فاز از پروژه به صورت زیر هستند.

۱. دسته‌بندی دو کلاسه[۲]

در این قسمت، هدف تشخیص وجود یا عدم وجود یک ویژگی درون داده‌هاست. برچسب درست نشان‌دهنده این موضوع در مجموعه داده‌ها در آدرس bin_train و bin_val وجود دارد. آن‌چه در برای این بخش باید تحویل دهید عبارت است از کدها و جوابتان روی داده‌ی تست در فایلی با نام bin_test.

دسته‌بندهای اولیه که برای این قسمت در نظر گرفته شده‌اند به ترتیب زیر می‌باشند:

- دسته‌بندهای غیر احتمالی:
  - SVM
  - KNN
- دسته‌بندهای احتمالی:
  - Logistic Regression
  - Naïve Bayes

از معیارهای دقت[۳] و F[۴] برای ارزیابی روش‌ها استفاده کنید.

۲. دسته‌بندی چند کلاسه[۵]

هرکدام از داده‌هایی که در اختیار شما قرار گرفته است متعلق به یکی از ۴۳ کلاس ممکن است. برچسب‌های صحیح در آدرس‌های y_val و y_train در اختیار شما قرار گرفته است و شما نیز باید جوابتان را روی داده تست در فایلی با نام y_test تحویل دهید.

دسته‌بندهای اولیه که برای این قسمت در نظر گرفته شده‌اند به ترتیب زیر می‌باشند:

- دسته‌بندهای غیر احتمالی:
  - SVM
  - KNN
- دسته‌بندهای احتمالی:
  - Multiclass Logistic Regression
  - Naïve Bayes

از معیارهای دقت، ROC-AUC و Macro-F1 و Micro-F1 برای ارزیابی استفاده کنید.

---

2 Binary Classification
3 Accuracy
4 F-Measure
5 Multiclass Classification

۳. رگرسیون[۶]

برای داده‌های ما یک خروجی عددی پیوسته نیز مدنظر است که مقادیر آن در آدرس‌های reg_train و reg_val موجود است و شما نیز پیش‌بینی خود را برای داده‌های تست در فایلی با نام reg_test ارسال خواهید کرد.

از روش‌های زیر برای رگرسیون استفاده کنید:

- Linear Regression
- KNN Regression

همچنین از معیارهای RMSE و MAE برای ارزیابی روش‌تان استفاده کنید.

۴. دسته‌بندی چند برچسبی[۷] (اختیاری)

مساله‌ی دسته‌بندی چندبرچسبی با مسائلی که تاکنون دیده‌اید کمی فرق می‌کند. در این مساله تعدادی برچسب موجود است و هر داده می‌تواند تعدادی از این برچسب‌ها را داشته باشد. در واقع هر داده می‌تواند هر برچسب را داشته باشد یا نداشته باشد. مثالی از این مساله می‌تواند برچسب زدن به عکس‌های مختلف باشد. یک عکس می‌تواند برچسب‌های *دریا*، *خورشید* و *قایق* را هم‌زمان داشته باشد و عکس دیگری می‌تواند برچسب‌های *خیابان*، *ساختمان* و *ماشین* را دارا باشد؛ ولی بعید است که عکسی هم‌زمان برچسب‌های *قایق* و *خیابان* را داشته باشد.

داده‌های این پروژه نیز ۴۰ برچسب برایشان متصور است که مقادیر صحیح برای داده‌های آموزش و validation در آدرس‌های mult_val و mult_train در اختیار شما قرار گرفته است. بدیهی است که شما باید جوابتان را روی داده تست در فایلی با نام mult_test تحویل دهید.

مطالعه‌ای روی روش‌های دسته‌بندی چندبرچسبی و روش‌های ارزیابی آن‌ها داشته باشید و برای این مساله یک روش دسته‌بندی چندبرچسبی ارائه دهید.

از معیارهای دقت مبتنی بر نمونه[۸] و Micro-F1 و Macro-F1 برای ارزیابی روش‌هایتان استفاده کنید.

---

6 Regression

7 Multi Label Classification

8 Example Accuracy