تنظیم‌هایپر‌پارامتر‌ها و دقت مدل‌ها

# تنظیم هایپر‌پارامتر‌ها

در این قسمت ما بایستی دقت هر یک از مدل‌های خود بر روی داده‌ها را به دست آوریم. از آنجایی که بعضی مدل‌های ما نظیر شبکه عصبی و یا svm می‌توانند خود دارای هایپر‌پارامتر‌های با مقادیر متفاوتی باشند در نتیجه نیاز داریم تا مقادیر بهینه این هایپرپارامتر‌ها را برای هر مدل بنابر داده‌‌های آموزش تعیین کنیم. برای نیل به این هدف ما از متد RandomizedSearchCV ارائه‌شده توسط scikit بهره می‌گیریم. اساس کار این متد این است که در ورودی خود یک مدل و یک داده آموزش و یک فضای حالت از پارامتر‌ها را دریافت می‌کند و سپس به میزان دفعات مشخص شده یک مدل با پارامتر‌های تصادفی از آن مجموعه فضای حالت پارامتر‌ها را تولید کرده و دقت cross validation روی آن‌ها را محاسبه کرده و در نهایت بهترین مدل از میان مدل‌های جستجوشده را بر‌می‌گرداند. از آنجا که برای هر مدل فضای حالت پارامتر‌ها بسیار بزرگ بوده و توان محاسباتی و زمانی ما کمتر از میزان حد لازم برای جستجوی کل فضای بهینه است و بنابر توصیه گفته شده در توضیحات پروژه ما به جستجوی ۱۰ حالت از میان فضای حالت محدود شده پارامتر‌ها اکتفا می‌کنیم.

لازم به ذکر است که هر از یک مدل‌های استفاده‌شده در این پروژه فایلی به نام خود در پوشه models دارا می‌باشند که در آنها متد find\_hyperparameter توسط ما پیاده‌سازی شده است که در خروجی خود بهترین دقت مدل را چاپ کرده و همچنین مدل (بدون پنالتی منظم‌ساز) با بهترین هایپرپارامتر‌ها را نیز بر‌می‌گرداند. برای راحتی در مشاهده نتایج نیز برای هر یک از مدل‌ها این متد‌ها در فایل ipython مربوط به jupyter notebook واقع در پوشه notebook صدا زده‌شده اند.

# مدل adaboost

ما برای این مدل که نیاز به یک دسته‌بند پایه دارد از یک دسته‌بند درخت با حداکثر عمق ۴ استفاده کردیم و ابرپارامتر‌های تعداد ارزیابی‌گرها و نرخ یادگیری Adaboost را مورد جستجو قرار دادیم. در نهایت نتیجه حاصل این بود که بهترین تعداد ارزیابی‌گرها برابر ۱۰۴ و بهترین نرخ یادگیری برابر ۰.۶ است. دقت این مدل با این هایپرپارامتر‌های بهینه نیز برابر با ۰.۸۵۳ است.

# مدل logistic regression

از آنجایی که این مدل دارای هایپرپارامتر‌ خاصی نبود در نتیجه در آن به جستجوی پارامتری نپرداختیم. دقت این مدل نیز برابر با ۰.۹۶۳ است.

# مدل neural network

بنابر توضیحات گفته شده در صورت پروژه ما بایستی از یک شبکه عصبی تماما متصل با دو لایه استفاده کنیم. پارامتر‌های مورد جستجوی ما در این مدل عبارتند از سایز لایه پنهان و همچنین تابع فعالسازی نورون‌ها. در نهایت برای سایز بهینه لایه پنهان به ۵۶ و برای تابع بهینه فعالسازی نورون‌ها به tanh رسیدیم. با این پارامتر‌های بهینه دقت این شبکه عصبی برابر است با ۰.۹۷۵

# مدل random forest

ما برای این مدل هایپرپارامترهای تعداد ارزیابگر‌ها و حداکثر عمق درخت و حداقل تعداد داده‌های لازم برای جداسازی را مورد جستجو قرار دادیم که بنابر بر نتیجه، تعداد ارزیابگر‌ها برابر ۲۱۳ و حداکثر عمق درخت برابر با ۳۸ و حداقل تعداد داده لازم برای جداسازی ۲ می‌باشد. همچنین دقت نهایی این مدل با پارامتر‌های بهینه برابر با ۰.۹۶۳ می‌باشد.

# مدل SVM

برای این مدل هایپر‌پارامتر‌های میزان خطا برای داده‌های به اشتباه دسته‌بندی شده، تابع کرنل، درجه تابع کرنل (در صورت خطی بودن کرنل) و پارامتر گاما را مورد جستجو قرار دادیم. که در نهایت بهترین حالت وجود یک تابع کرنل خطی درجه دو به همراه ضریب خطا برابر با ۱ می‌باشد. در نهایت دقت مدل svm با بهترین پارامتر‌ها برابر با ۰.۹۷۱ می‌باشد.