تمرین سوم درس هوش محاسباتی – بهار ۱۴۰۱

در این تمرین شما قصد دارید با استفاده از طبقه‌بندی[1] رندم فارست[2] یک دیتاست تشکیل شده از ۵ دامین با کلاس‌های ارقام ۰ تا ۹ را در ابتدا به صورت مستقیم ارقام آن را طبقه بندی کنید. سپس در بخش بعدی باید دامین داده و سپس برای هر دامین به صورت جدا مجددا یک طبقه‌بندی دیگر ایجاد و برای تشخیص نوع رقم آن دامین از آن طبقه‌بندی استفاده کنید. در ادامه باید بررسی کنید که در صورتی که در آموزش طبقه‌بندی های رقم از دیتاهای دامین های دیگر استفاده کنید به دقت بالاتری می رسید یا خیر (در گزارش خود باید به روند نتیجه گیری خود و دلیل آن اشاره کنید)

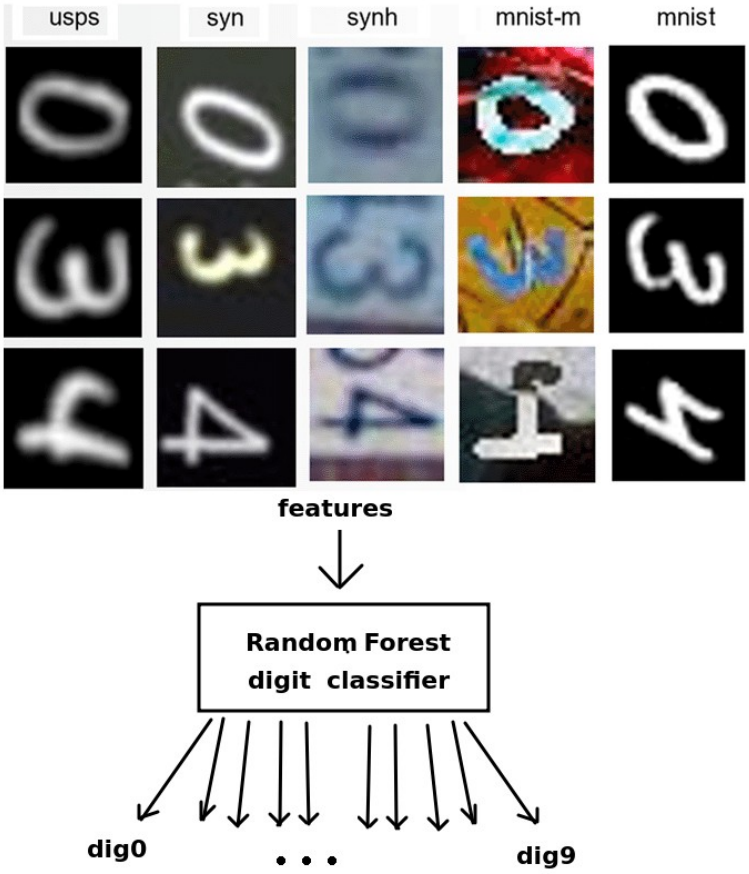
برای این منظور، یک دیتاست تشکیل شده از ۵ دامین مختلف به شما داده شده است. دیتاست داده شده، به این صورت می باشد که به ازای هر تصویر یک بردار ویژگی استخراج شده ودو برچسب برای آن در اختیار شما قرار داده شده است. یکی از این برچسب‌ها ، نشان دهنده برچسب محتوای تصاویر(۰-۹) است و برچسب دیگر مربوط به برچسب دامین(۰-۴) می باشد. در داده های آموزشی که در اختیار شما قرار گرفته (داده های این فولدر)، بردار های ویژگی استخراج شده از ۵۰ هزار تصویرآورده شده که ۲۵ هزارتای آن برای آموزش و ۲۵ هزارتای باقی مانده برای تست استفاده خواهند شد، به طوری که بردار ویژگی هر تصویر ۱۰۲۴ بعدی است.

همچنین کد لود دیتا در همان فولدر قرار دارد و می توانید برای لود کردن فیچر ها از آن استفاده کنید.

1 Classification
2 Random Forest

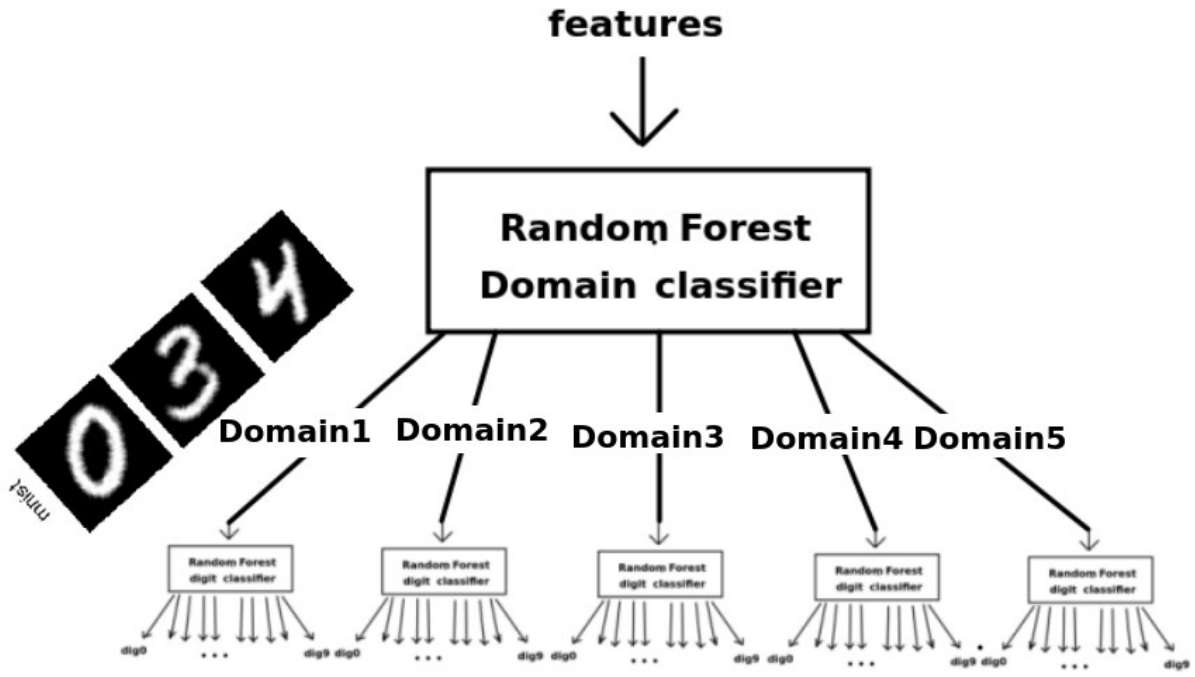
بخش اول – طبقه‌بندی رقم

در بخش اول، شما باید از یک الگوریتم طبقه‌بندی رندم فارست استفاده کنید که بر روی لیبل رقم‌داده‌ها بـه خـوبی عمل کند. انتخاب هایپرپارامترهای مناسب برای رسیدن به دقت مطلوب بر عهده شماست امـا بایـد پس از انتخـاب این هایپر پارامترها، نشان دهید که انتخاب مناسبی بوده‌اند. همچنین در هر مرحله تاثیر تغییر پارامتر های مختلـف الگوریتم را تست کرده و گزارش کنید. همچنین ماتریس گمراهی را نیز محاسبه کنید.



بخش دوم – طبقه‌بندی دامین و رقم

در بخش دوم، شما باید در ابتدا از یک الگوریتم طبقه‌بندی رندم فارست استفاده کنید که برای تشخیص دامین داده‌ها به خوبی عمل کند. در ادامه باید به ازای هر دامین یک الگوریتم رندم فارست دیگر با داده های همان دامین آموزش دهید. حال برای ارزیابی دقت مدل خود باید دیتا تست را خود را ابتدا از رندم فارست دامین عبور داده و سپس با توجه به پیشبینی طبقه‌بندی دامین آن را به طبقه‌بندی رقم همان دامین داده و درستی آن را بررسی کنید. انتخاب هایپرپارامترهای مناسب برای رسیدن به دقت مطلوب بر عهده شماست اما باید پس از انتخاب این هایپر پارامترها، نشان دهید که انتخاب مناسبی بوده اند.

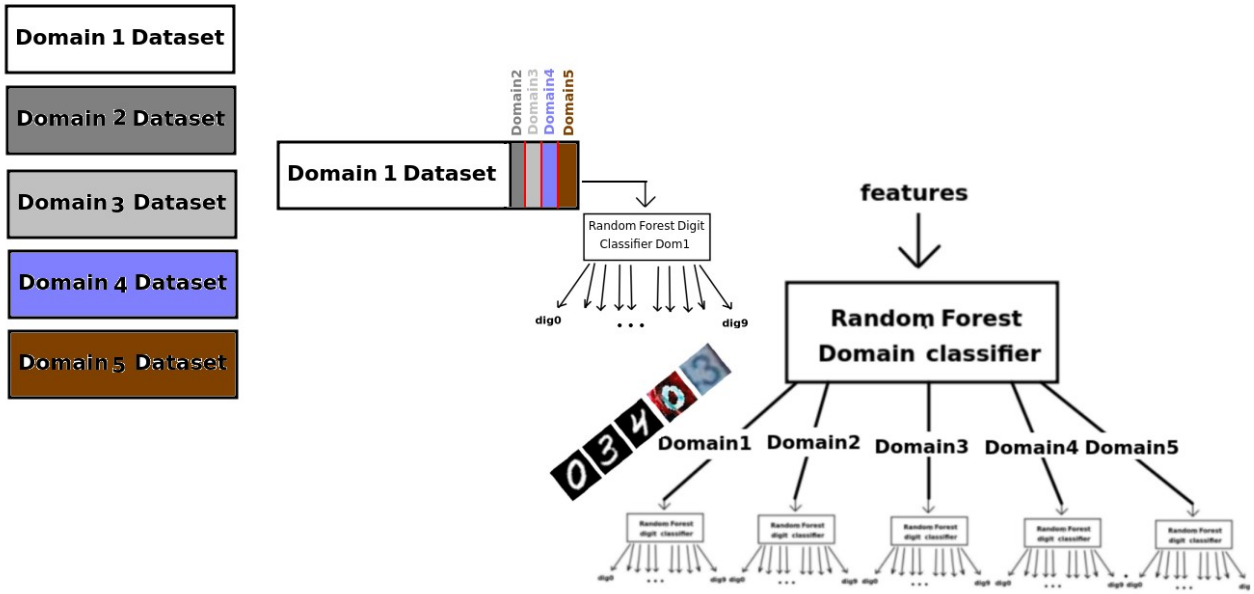


بخش سوم – استفاده از داده های دامین دیگر در طبقه‌بندی رقم

در دنیا واقعی ممکن است که داده های ما دارای نـویز یـا تحت تـاثیر عوامـل دیگـر بـوده و در نتیجـه ممکن اسـت علی‌رغم آموزش یک مدل خوب همچنان نتوانیم به درسـتی دامین یـک داده را پیشـبینی کـنیم. نتیجـه آن بـاعث پیشبینی اشتباه[3] در طبقه‌بند دامین شده و باعث می شود دیتاهای تست از یک دامین به طبقه بند دامین دیگـر بروند. از آنجایی که در پروسه آموزش، طبقه‌بند رقمِ هر دامین تنها دیتاهای همان دامین را دیده است، در نتیجه احتمالا نمی تواند داده هایی که اشتباهی در این دامین پیشبینی شده اند را تشخیص دهد.

در بخش سوم، شما باید بررسی کنید که در صورتی کـه از داده هـای دامین دیگـر در پروسـه آمـوزش طبقه‌بنـدی رقم‌ها استفاده کنید آیا به نتایج بهتری دست پیدا می کنید یا خیر؟ برای اینکار ابتدا باید ماتریس گمراهی طبقه‌بند دامین را پلات کرده و با توجه به پیشبینی هایی که طبقه بند انجام داده تصمیم بگیرید که با اضافه کردن دیتاهای دامین دیگر به دیتا آموزشِ کدام طبقه‌بند رقم (ها) باعث افزایش دقت آن می شود. سپس تغییرات را با چند ضریب (از هر دامین به چه مقدار) اعمال کرده و مدل را تست کنید و گزارش بهبود یا عدم بهبود مدل را در گـزارش خـود به همراه استدلال خود برای آن نتیجه را مکتوب کنید.

در نهایت با استفاده از معیار(جمع وزندار یا ...) مناسب خروجی هر سه بخش را بـا هم مقایسـه کـرده دقت بدسـت آمده در هر بخش را مورد بررسی قرار دهید.



3 Miss-classified