علی وزیری نژاد **گزارش فاز دوم** **هوش محاسباتی**
تهمینه توکلی

بخش اول (آموزش Digit Classifier):

(فایل Part1.ipynb)

در این بخش از RandomForestClassifier برای آموزش مدل طبقه بندی محتوا (digit) روی تمام داده ها استفاده کردیم.

پارامترها:

برای آموزش مدل تنها پارامترهای n_estimator و min_samples_split را تغییر دادیم که پارامتر های اصلی RandomForestClassifier بودند (پارامتر max_depth را تعیین نکردیم از آنجایی که با انتخاب درست min_sampels_split، خود به خود تعیین می شود)

پارامتر n_estimators:

مدل را با $n\_estimators > 0$ به صورت 50 تا 50 تا:

$$n\_estimators = [1, 50, 100, 150, \ldots]$$

آموزش دادیم. ماکزیمم مقدار n_estimators را هم مقداری تعیین کردیم که باعث می شود آموزش مدل بیش از ۵ دقیقه به طول بینجامد.

در نهایت $n\_estimators = 500$ را انتخاب کردیم که در مقدار $min\_samples\_split = default$ دقت برابر

$$Accuracy = 0.88$$

بر روی داده های Test را داشت و در حدود ۵ دقیقه قابل آموزش دادن بود.

پارامتر min_samples_split:

مدل را با $min\_samples\_split >= 2$ به صورت 100 تا 100 تا:

$$min\_samples\_split = [2, 100, 200, \ldots, 1000]$$

و $n\_estimators = 50$ (بهینه ترین مقدار n_estimators برای کوتاهی زمان و دقت کافی آموزش) آموزش دادیم.

بازه [2, 100] را بر اساس دقت مدل انتخاب کردیم و دوباره مدل را 10 تا 10:

$$\text{min\_samples\_split} = [2, 10, 20, \ldots, 100]$$

آموزش دادیم.

در نهایت $\text{min\_samples\_split} = 20$ را اتخاب کردیم که در مقدار $\text{n\_estimators} = 50$ دقت برابر:

$$\text{Accuracy} = 0.84$$

بر روی داده های Test را داشت.

مدل نهایی:

مدل نهایی را با پارامترهای انتخابی $\text{n\_estimators} = 200$ و $\text{min\_samples\_split} = 20$ آموزش دادیم. که دقت (f1_score) برابر بود با:

$$\text{Accuracy on Train Data} = 0.998$$

$$\text{Accuracy on Test Data} = 0.873$$

(از f1_score استفاده کردیم که بر اساس confusion_matrix است و بر اساس آن می توانستیم مدل ها را با هم مقایسه کنیم)

جدول confusion_matrix مدل هم در فایل مربوطه موجود است.

بخش دوم (آموزش Domain Classifier و Digit Classifier برای هر دامنه):

(فایل Part2.ipynb)

در این بخش از RandomForestClassifier برای آموزش مدل طبقه بندی دامنه (domain) روی تمام داده ها و سپس، آموزش مدل طبقه بندی محتوای (digit) هر دامنه (domain)، استفاده کردیم.

مدل domain_classifier:

مشابه روش بخش اول، مدل را با پارامتر ها مختلف آموزش دادیم و در نهایت به پارامتر های:

$$\text{n\_estimators} = 200$$

$$\text{min\_samples\_split} = 30$$

رسیدیم. که مدل دقت:

$$\text{Accuracy on Train Data} = 0.913$$

$$\text{Accuracy on Test Data} = 0.943$$

را دارا بود.

(از f1_score استفاده کردیم که بر اساس confusion_matrix است و بر اساس آن می توانستیم مدل ها را با هم مقایسه کنیم)

جدول confusion_matrix مدل هم در فایل مربوطه موجود است.

مدل های domain_digit_classifier:

مشابه روش بخش اول، مدل ها را با پارامتر ها مختلف آموزش دادیم (برای مدل های از پارامترهای مشترک استفاده کردیم) و در نهایت به پارامتر های:

$$\text{n\_estimators} = 200$$

$$\text{min\_samples\_split} = 20$$

رسیدیم. که مدل ها دقت وزن دار:

$$\text{Accuracy on Train Data (split based on true domain)} = 0.998$$

$$\text{Accuracy on Train Data (split based on predicted domain)} = 0.990$$

$$\text{Accuracy on Test Data (split based on true domain)} = 0.876$$

$$\text{Accuracy on Test Data (split based on predicted domain)} = 0.870$$

را دارا بودند. (جزئیات دقت خالص مدل digit_clf برای هر دامنه در فایل مربوطه موجود است)

بخش سوم (ترکیب داده های آموزشی Domain Classifier ها):

(فایل Part3.ipynb)

در این بخش، داده های آموزشی هر دامنه را با ضریب ترکیب:

$$\text{mix\_ratio} = [10\%, 20\%, \ldots, 100\%]$$

با دیگر داده های دیگر دامنه ها ترکیب کردیم.

(مثلا $mix\_ratio = 10\%$ برای دامنه $domain = 2$ یعنی: 10 درصد از داده های هر دامنه را با داده های دامنه $domain = 2$ ترکیب کردیم و مدل digit classifier را دوباره با این داده ها آموزش دادیم)

(نتایج تمام mix_ratio ها برای تمام دامنه ها در فایل مربوطه موجود است)

در نهایت با مقایسه دقت مدل های جدید با mix_ratio های مختلف، این mix_ratio ها را انتخاب کردیم:

$$mix\_ratio = 70\% \text{ برای } domain = 0$$

$$mix\_ratio = 20\% \text{ برای } domain = 1$$

$$mix\_ratio = 10\% \text{ برای } domain = 2$$

$$mix\_ratio = 0\% \text{ برای } domain = 3$$

$$mix\_ratio = 0\% \text{ برای } domain = 4$$

که به دقت وزن دار نهایی زیر برای مدل ها رسیدیم:

$$\text{Accuracy on Train Data (split based on true domain)} = 0.998$$

$$\text{Accuracy on Train Data (split based on predicted domain)} = 0.991$$

$$\text{Accuracy on Test Data (split based on true domain)} = 0.879$$

$$\text{Accuracy on Test Data (split based on predicted domain)} = 0.874$$

(جزئیات دقت خالص مدل $digit\_clf$ برای هر دامنه در فایل مربوطه موجود است)