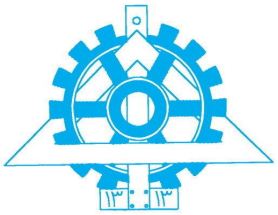
تمرین کامپیوتری شماره ۲



دانشکده مهندسی برق و کامپیوتر

سیستم‌های عامل - پاییز ۱۳۹۸

استاد :
مهدی کارگهی

مهلت تحویل :
شنبه ۲۵ آبان، ساعت ۲۳:۵۵

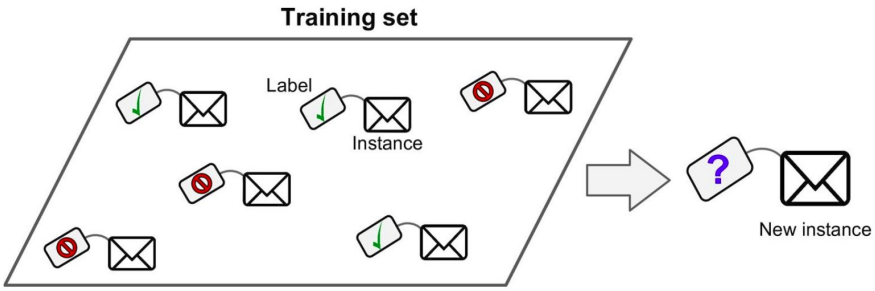
طراحان :
میلاد حکیمی، محمّد مریدی

مقدمه

هدف از این تمرین، آشنایی با نحوه مدیریت کردن پردازه[1]ها و راه‌های ارتباطی میان آن‌ها می‌باشد. در این تمرین به شبیه‌سازی یکی از روش‌های رایج در یادگیری ماشین[2] پرداخته می‌شود. به عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی، یادگیری ماشین به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آن‌ها رایانه‌ها و سامانه‌ها توانایی یادگیری و پیش‌بینی پیدا می‌کنند.

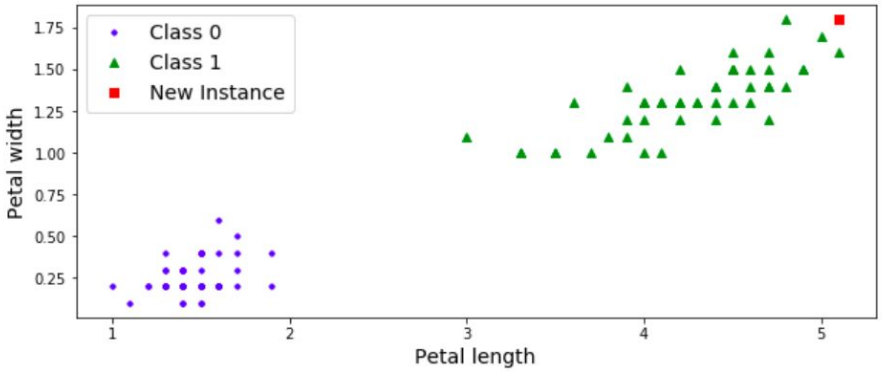
دسته‌بندی[3]

در حوزه یادگیری ماشین، دسته‌بندی نوعی یادگیری محسوب می‌شود که مجموعه‌ای از داده‌ها برای آموزش وجود دارند. در یادگیری ماشینی، دسته‌بندی مسئله شناسایی تعلق مشاهده جدید، به یکی از دسته‌ها بر اساس مجموعه‌ای از مشاهدات می‌باشد که عضویت در دسته‌هایشان مشخص می‌باشد.



[1] Process
[2] Machine Learning
[3] Classification

برای مثال تصور کنید که می‌خواهید نام یک گل را بر اساس طول و عرض گلبرگ‌های آن تشخیص دهید. بدین منظور لازم است که یک دسته‌بند[4] برای این منظور آموزش ببیند (توانایی تشخیص نوع گل را پیدا کند) و پس از آن بر اساس ویژگی‌هایی که یک گل را توصیف می‌کند (طول و عرض در این مثال) به دسته‌بند داده شود. این دسته‌بند براساس مشاهداتی که در گذشته داشته است (در مرحله آموزش) تعلق این گل را به یکی از دسته‌ها تشخیص می‌دهد.



دسته‌بندی خطی[5]

در حوزه یادگیری ماشین نمونه‌هایی که قصد پیش‌بینی نوع و یا یک ویژگی آن‌ها وجود دارد، با استفاده از تعدادی ویژگی عددی و قابل اندازه‌گیری در قالب بردار ویژگی[6] توصیف می‌شوند.

تعداد زیادی از الگوریتم‌هایی که برای دسته‌بندی وجود دارند، می‌توانند با استفاده از یک تابع خطی[7]، به هر یک از دسته‌ها امتیاز[8]ی اختصاص دهند. این امتیازدهی با استفاده از ضرب داخلی بردار ویژگی با بردار وزن هر یک از دسته‌ها صورت می‌گیرد. دسته‌ی پیش‌بینی شده، دسته‌ای می‌باشد که بالاترین امتیاز را بین سایر دسته‌ها به خود اختصاص دهد. این تابع در زیر توصیف شده است:

$$score(X_i,\ k)\ =\ \beta_k.X_i$$

بطوریکه X_i بردار ویژگی نمونه i ام، β_k بردار وزن دسته k ام و $score(X_i,\ k)$ امتیازی می‌باشد که دسته k ام با اختصاص یافتن به نمونه i ام بدست می‌آورد.

برای مثال تصور کنید که دسته‌بند توانایی تشخیص دو نوع گل از یکدیگر را دارد. بدین ترتیب این دسته‌بند دارای دو بردار وزن می‌باشد که هر دسته آن به ویژگی‌های مختلف نمونه وزن‌های مختلفی اختصاص می‌دهد. نمونه‌ای از بردارهای وزن یک دسته‌بند را در زیر مشاهده می‌کنید:

---

[4] Classifier
[5] Linear Classification
[6] Feature Vector
[7] Linear Function
[8] Score

	β_0	β_1	$Bias$
$Class_1$	31.18	-4.74	-8.00
$Class_2$	-31.18	4.74	8.00

حال این دسته‌بند با بردارهای وزن ذکر شده، قصد تشخیص نمونه‌ای که دارای بردار ویژگی زیر می‌باشد را دارد:

$Bias$	$Length$	$Width$
1	0.9	0.1

ستون‌های $Length$ و $Width$ همانطور که از نام آن‌ها برمی‌آید معرف طول و عرض گلبرگ مربوط به گل‌ها می‌باشد. پس از انجام ضرب داخلی دو بردار لازم است که امتیاز آن‌ها با مقداری ثابت برای هر دسته جمع شود. در این مثال برای این که امتیاز مربوط به هر دسته با محاسبه ضرب داخلی بدست آید، یک ویژگی به این نام و با مقدار ۱ به ویژگی‌های این نمونه اضافه شده است که با محاسبه ضرب داخلی آن با بردار وزن هر دسته، مقداری ثابت با امتیاز دسته برای نمونه مذکور جمع می‌شود.

برای محاسبه دسته مربوط به نمونه لازم است که ضرب داخلی بردار ویژگی نمونه در هر یک بردارهای وزن محاسبه شود.

$$score(X_i,\ k) = \beta_{k,0} \times Length_i + \beta_{k,1} \times Width_i + Bias_k \Rightarrow$$

$$score(X_i,\ 1) = 31.18 \times 0.9 + (-4.74) \times 0.1 + (-8.00) = 19.588$$

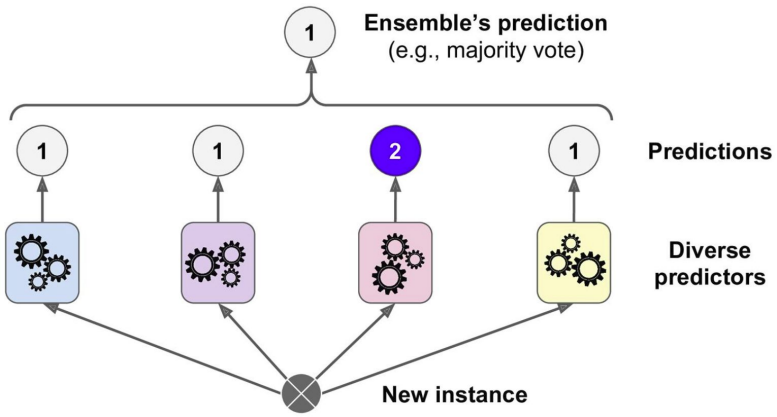
$$score(X_i,\ 2) = -31.18 \times 0.9 + 4.74 \times 0.1 + 8.00 = -19.588$$

با توجه به این که اولین دسته امتیاز بیشتری را کسب کرد، دسته مربوط به این نمونه دسته شماره یک می‌باشد.

دسته‌بندی ترکیبی[9]

در حوزه یادگیری ماشین، یکی از روش‌هایی که برای بدست آوردن بازدهی بهتر در دسته‌بندی مورد استفاده قرار می‌گیرد، ترکیب کردن نتیجه چندین دسته‌بند و پیش‌بینی دسته، برحسب بیشترین تعداد تکرار برای یک دسته می‌باشد.

[9] Ensemble Classification



مثال ذکر شده در قسمت قبل را در نظر بگیرید. دسته مربوط به این نمونه با استفاده از بردارهای وزن داده شده برای دسته‌بند مذکور، دسته شماره یک تعیین گردید. حال تصور کنید بردار ویژگی مربوط به این نمونه به دسته‌بندهای دیگری که دارای بردارهای وزن مخصوص به خود می‌باشند داده شده است و این دسته‌بندها به صورت فوق عمل کرده‌اند و هر کدام دسته‌ای را به نمونه اختصاص داده‌اند.

در این مرحله یک رأی‌دهنده[10]، خروجی‌های مربوط به دسته‌بندها را دریافت می‌کند و با توجه به این که چه دسته‌ای بیشتر از سایر دسته‌ها تکرار شده است، دسته نهایی را تعیین می‌کند. برای مثال در شکلی که در بالا آمده است دسته نهایی برای نمونه دسته شماره یک می‌باشد.

شرح تمرین

در این تمرین به شبیه‌سازی یک دسته‌بند ترکیبی می‌پردازید. این دسته‌بند شامل چندین دسته‌بند خطی می‌باشد که این دسته‌بندها آموزش دیده شده‌اند و بردارهای وزن هر یک از آن‌ها در پرونده[11]ای جداگانه در اختیار شما قرار داده شده است. وظیفه‌ای که برنامه شما بر عهده دارد، پیش‌بینی دسته مربوط به نمونه‌هایی می‌باشد که تحت عنوان مجموعه داده اعتبارسنجی[12] در اختیار شما قرار داده شده است.

معماری سامانه

به دلیل بالا بودن تعداد نمونه‌هایی که قصد دسته‌بندی آن‌ها وجود دارد و مستقل بودن نتیجه دسته‌بندهای خطی از یکدیگر، می‌توان عملیات دسته‌بندی توسط دسته‌بندها را بصورت موازی با یکدیگر انجام داد. به همین منظور یک راه‌حل مناسب برای این مسأله، استفاده از چندین پردازه جهت دسته‌بندی می‌باشد. برای این سامانه سه نوع پردازه در نظر گرفته شده است:

۱. پردازه دسته‌بند ترکیبی

۲. پردازه دسته‌بند خطی

[10] Voter
[11] File
[12] Validation Dataset

۳. پردازه رأی‌دهنده[13]

در ادامه وظایف هر یک از پردازه‌ها پرداخته می‌شود.

پردازه دسته‌بند ترکیبی

این پردازه، پردازه والد[14] سامانه محسوب می‌شود و وظیفه آن **بوجود آوردن** *پردازه‌های دسته‌بند خطی* و **معرفی** نام پرونده مربوط به بردارهای وزن هر دسته‌بند به آن از طریق یک **Unnamed Pipe** می‌باشد. همچنین این پردازه وظیفه **بوجود آوردن** *پردازه رأی‌دهنده* را نیز بر عهده دارد؛ پس از اتمام عملیات پردازه رأی‌دهنده، صحت عملکرد دسته‌بند ترکیبی سنجیده می‌شود. این عملیات از طریق مقایسه اطلاعات بدست آمده از پیش‌بینی با برچسب‌های داده‌های اعتبارسنجی صورت می‌گیرد.

پردازه دسته‌بند خطی

این پردازه، پردازه فرزند[15] برای پردازه‌های دسته‌بند ترکیبی محسوب می‌شود که پس از **دریافت** نام پرونده بردارهای وزن خود، دسته مربوط به هر نمونه‌ای که در پرونده مربوط به داده‌های اعتبارسنجی می‌باشد را با محاسبه ضرب داخلی بردارهای وزن خود با بردار ویژگی مربوط به نمونه، **محاسبه** کرده و از طریق یک **Named Pipe** به پردازه رأی‌دهنده می‌دهد.

پردازه رأی‌دهنده

این پردازه پس از اتمام عملیات تمام دسته‌بندهای خطی بر روی مجموعه داده‌های اعتبارسنجی، دسته مربوط به هر نمونه را از طریق یک **Named Pipe** در اختیار پردازه دسته‌بند ترکیبی قرار می‌دهد.

ورودی و خروجی برنامه

پردازه دسته‌بند ترکیبی، در قالب زیر آدرس مربوط به پوشه[16] بردارهای وزن دسته‌بندها و پوشه مربوط به داده‌های اعتبارسنجی را از طریق آرگومان‌هایی در رابط خط فرمان[17] از کاربر دریافت می‌کند. برنامه شما باید صحت عملکرد سامانه را تا دو رقم اعشار (با گرد کردن عدد اعشاری) نمایش دهد.

[13] Voter
[14] Parent
[15] Child
[16] Directory
[17] Command Line Interface

نمونه ورودی و خروجی سامانه (با فرض این که پوشه `Assets` بارگذاری در سایت درس، در کنار پرونده اجرایی شما قرار گرفته است) در ذیل آمده است:

- نمونه ورودی

```
./EnsembleClassifier.out Assets/validation Assets/weight_vectors
```

- نمونه خروجی

```
Accuracy: 97.20%
```

نکات تکمیلی

- داده‌های اعتبارسنجی، در پوشه‌ای به‌نام `validation` قرار داده شده‌اند. در این پوشه پرونده‌ای به‌نام `dataset.csv` که مجموعه‌داده‌های اعتبارسنجی می‌باشد و برچسب‌های مربوط به هریک از نمونه‌های موجود در این پرونده، در پرونده‌ای به‌نام `labels.csv` در اختیار شما قرار داده شده است.
- بردارهای وزن مربوط به هر دسته‌بند در پوشه‌ای به‌نام `weight_vectors`، تحت عنوان `classifier_<number>.csv` تهیه شده است که همواره از شماره **صفر** شروع می‌شوند.
- **توجه کنید،** *تعداد بردارهای وزنی* که در پوشه مربوطه وجود دارد و *تعداد دسته‌هایی* که هر دسته‌بند می‌تواند تشخیص دهد، **متغیر** می‌باشد و *تعداد ویژگی‌های* موجود در مجموعه‌داده **ثابت** است.
- **تأکید می‌شود،** هدف از پروژه طراحی و استفاده‌ی صحیح از مفاهیم موازی‌سازی پردازه‌ها می‌باشد و سایر پیاده‌سازی‌ها قابل قبول نمی‌باشد.
- **دقت کنید،** در صورتی که علاوه بر موارد ذکر شده در شرح تمرین، نیاز به ارسال اطلاعات بیشتری میان پردازه‌ها بود، ارتباط میان پردازه‌ها فقط از طریق **Pipe** صورت می‌گیرد و روش‌های دیگر ارتباط میان پردازه‌ها قابل قبول نمی‌باشد.
- در صورتی که تعداد رأی‌های مربوط به چند دسته با یکدیگر **برابر** شد، دسته‌ای بعنوان دسته نهایی انتخاب می‌شود که شماره کوچکتری را داراست.
- برای تجزیه[18] پرونده‌های CSV[19] می‌توانید از کتابخانه‌های رایج C++ برای تجزیه پرونده‌هایی که در اختیارتان قرار گرفته است، استفاده کنید.

[18] Parse

[19] Comma-separated values

نحوه‌ی تحویل

- تمام فایل‌های خود را در قالب یک پرونده‌ی زیپ با نام `A2-<SID>.zip` در صفحهٔ CECM درس بارگذاری کنید که SID شمارهٔ دانشجویی شماست؛ برای مثال اگر شماره‌ی دانشجویی شما ۸۱۰۱۹۸۹۹۹ است، نام پروندهٔ شما باید `A2-810198999.zip` باشد.
- برنامهٔ شما باید در سیستم عامل لینوکس و با مترجم[20] `g++` با استاندارد c++11 ترجمه و در زمان معقول برای ورودی‌های آزمون اجرا شود.
- **دقت کنید** که پروژه شما باید `Makefile` داشته باشید و در `Makefile` خود مشخص کنید که از استاندارد c++11 استفاده می‌کنید.
- درستی برنامه‌ی شما از طریق آزمون‌های خودکار سنجیده می‌شود. **دقت شود** که نام پرونده‌ی اجرایی شما باید `EnsembleClassifier.out` باشد.
- هدف این تمرین یادگیری شماست. لطفاً تمرین را خودتان انجام دهید. در صورت کشف تقلب مطابق قوانین درس با آن برخورد خواهد شد.

[20] Compiler