تمرین کامپیوتری سوم

دانشکده مهندسی برق و کامپیوتر

سیستم‌های عامل - پاییز ۹۹

استاد:

دکتر مهدی کارگهی

پیش‌زمینه

مسئولان تمرین:

محمّد مریدی، مبینا شاه‌بنده و غزل مینایی

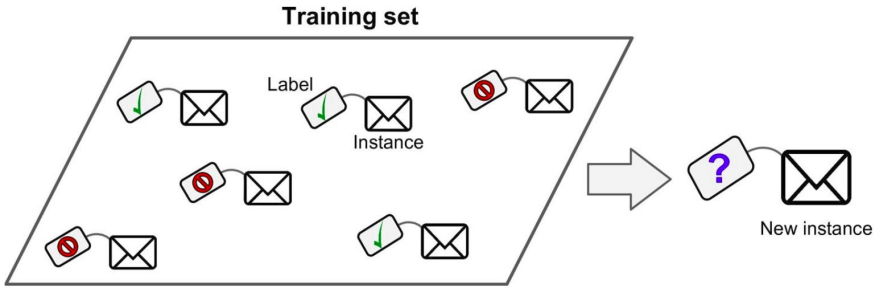
مقدمه

در این تمرین شما به تحلیل داده‌هایی که از مشخصات و قیمت فروش گوشی‌های موبایل جمع‌آوری شده‌است می‌پردازید. در این تمرین به شبیه‌سازی یکی از روش‌های رایج در یادگیری ماشین[1] 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.

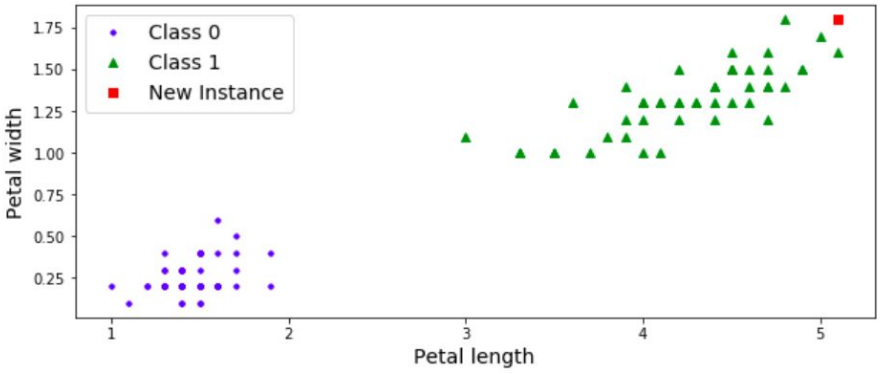
[1] Machine Learning

طبقه‌بندی[2]

در حوزه یادگیری ماشین، طبقه‌بندی نوعی یادگیری محسوب می‌شود و طبقه‌بندی مسئله شناسایی تعلق مشاهده جدید، به یکی از دسته‌ها بر اساس مجموعه‌ای از مشاهدات می‌باشد که عضویت در دسته‌هایشان مشخص می‌باشد.



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 طبقه‌بند[3] 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 طبقه‌بند داده شود. این طبقه‌بند براساس مشاهداتی که در گذشته داشته است (در مرحله آموزش) تعلق این گل را به یکی از دسته‌ها تشخیص می‌دهد.



[2] Classification

[3] Classifier

طبقه‌بندی خطی[4]

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[5] توصیف می‌شوند.

تعداد زیادی از الگوریتم‌هایی که برای طبقه‌بندی وجود دارند، می‌توانند با استفاده از یک تابع خطی[6]، به هر یک از دسته‌ها امتیاز[7]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 طبقه‌بند توانایی تشخیص دو نوع گل از یکدیگر را دارد. بدین ترتیب این طبقه‌بند دارای دو بردار وزن می‌باشد که هر دسته آن به ویژگی‌های مختلف نمونه وزن‌های مختلفی اختصاص می‌دهد. نمونه‌ای از بردارهای وزن یک طبقه‌بند را در زیر مشاهده می‌کنید:

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

[4] Linear Classification

[5] Feature Vector

[6] Linear Function

[7] Score

حال این طبقه‌بند 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.

مجموعه داده[8]

مجموعه داده‌ای که در این تمرین به شما داده شده‌است در قالب CSV[9] است. CSV نام یک قالب برای پرونده‌های متنی است که در آن مقادیر با استفاده از نماد کاما (,) از یکدیگر جدا می‌شوند. این قالب یکی از روش‌های پرطرفدار برای تبادل اطلاعات است.

اطلاعات گوشی‌های موبایل

اطلاعات گوشی‌های موبایل در پرونده train.csv در اختیار شما قرار داده شده‌است. در ادامه درباره‌ی هر ویژگی و نوع داده[10]ی مربوط به آن، توضیح مختصری آمده‌است.

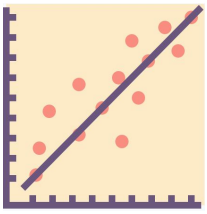
نام ویژگی	توضیح	نوع داده
battery_power	ظرفیت باتری در واحد میلی آمپر ساعت	عدد صحیح
blue	بلوتوث دارد یا خیر	عدد صحیح (۰ یا ۱)
clock_speed	سرعت پردازنده	عدد اعشاری
dual_sim	دو سیم کارته است یا خیر	عدد صحیح (۰ یا ۱)
fc	دقت دوربین جلو در واحد مگاپیکسل	عدد صحیح
four_g	از 4G پشتیبانی می‌کند یا خیر	عدد صحیح (۰ یا ۱)
int_memory	ظرفیت حافظه داخلی در واحد گیگابایت	عدد صحیح
m_dep	ضخامت گوشی در واحد سانتی‌متر	عدد اعشاری

[8] Dataset

[9] Comma-Separated Values

[10] Data Type

mobile_wt	وزن گوشی در واحد گرم	عدد صحیح
n_cores	تعداد هسته‌های پردازنده	عدد صحیح
pc	دقت دوربین پشت در واحد مگاپیکسل	عدد صحیح
px_height	رزولوشن صفحه نمایش (طول)	عدد صحیح
px_width	رزولوشن صفحه نمایش (عرض)	عدد صحیح
ram	ظرفیت حافظه موقت	عدد صحیح
sc_h	طول صفحه نمایش	عدد صحیح
sc_w	عرض صفحه نمایش	عدد صحیح
talk_time	بیشترین زمانی که با یک بار شارژ می‌توان از گوشی در تماس استفاده کرد	عدد صحیح
three_g	از 3G پشتیبانی می‌کند یا خیر	عدد صحیح (۰ یا ۱)
touch_screen	صفحه لمسی دارد یا خیر	عدد صحیح (۰ یا ۱)
wifi	از WiFi پشتیبانی می‌کند یا خیر	عدد صحیح (۰ یا ۱)
price_range	رنج قیمت	عدد صحیح (۰،۱،۲ یا ۳)



بردارهای وزن

همانطور که در بخش طبقه‌بندی خطی ذکر شد، امتیازدهی برای تعیین طبقه هر نمونه با استفاده از ضرب داخلی بردار ویژگی هر نمونه با بردار وزن هر یک از دسته‌ها صورت می‌گیرد و دسته‌ی پیش‌بینی شده، دسته‌ای است که بالاترین امتیاز را بین سایر دسته‌ها به خود اختصاص دهد. بردارهای وزن مدل آموزش دیده شده در قالب یک فایل CSV با نام weights.csv در اختیار شما قرار داده شده است که هر سطر از آن مربوط به یک طبقه قیمتی (0,1,2,3) و هر ستون از آن مقدار وزن مربوط به یک ویژگی از میان ۲۰ ویژگی است و ستون آخر مقدار $Bias$ است. هر سطر از این فایل در واقع نشان‌دهنده بردار وزن طبقه قیمتی متناظر با هر طبقه قیمیت در کنار مقدار $Bias$ آن است.

نکات تکمیلی

- در ستون‌های دوربین (fc, pc) عدد 0 نشان‌دهنده این است که آن گوشی دوربین متناظر را ندارد.
- در ستون price_range، اعداد مربوطه به شرح زیر هستند:

ارزان	متوسط	گران	بسیار گران
0	1	2	3

- **تضمین می‌شود** که در داده‌هایی که در اختیار شما قرار گرفته‌است، تنها از کاراکتر ',' برای جداسازی اجزا استفاده شده‌است.

منابع

https://www.kaggle.com/iabhishekofficial/mobile-price-classification

https://www.iconfinder.com/