تمرین کامپیوتری سوم

دانشکده مهندسی برق و کامپیوتر

سیستم‌های عامل - بهار 1400

استاد:

دکتر مهدی کارگهی

پیش‌زمینه

مسئولان تمرین:

مبینا شاه‌بنده، محمدصابر ابراهیم نژاد

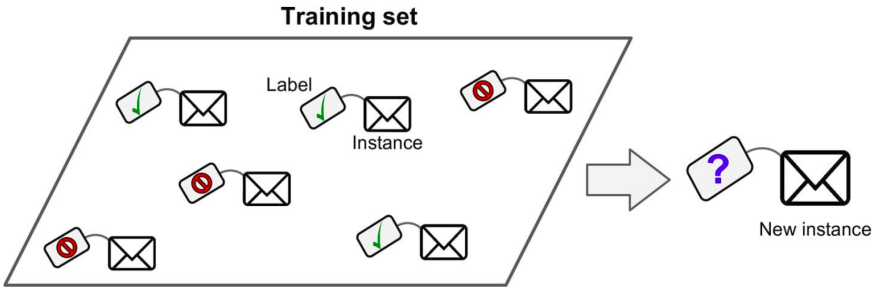
مقدمه

در این تمرین شما به تحلیل داده‌هایی که از مشخصات و قیمت فروش خانه ها 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.

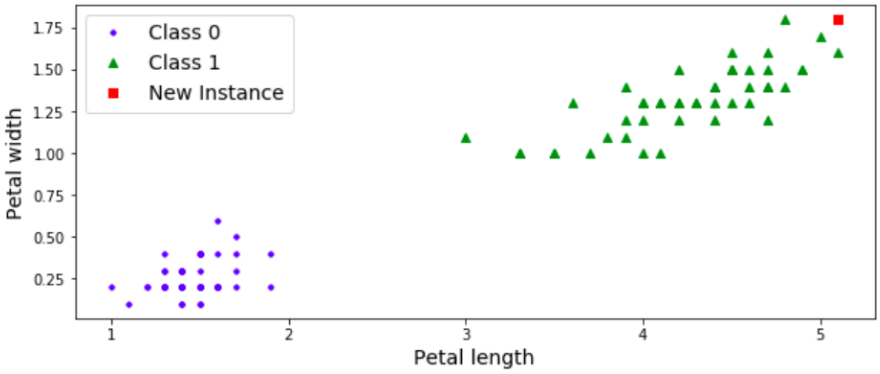
[1] Machine Learning

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 بر اساس میانگین و انحراف معیار[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. یکی از روش های طبقه بندی داده ها، استفاده از میانگین و انحراف معیار است که در آن ابتدا میانگین و انحراف معیار یک ویژگی از داده ها بدست آمده و سپس بر اساس آن، بازه تخمینی ساخته می شود. این بازه تخمین به صورت زیر بدست می آید:

$$(mean - std, mean + std)$$

برای مثال تصور کنید که طبقه‌بند توانایی تشخیص دو نوع گل از یکدیگر را بر اساس طول گلبرگ آنها دارد. حال فرض کنید گل هایی که در کلاس $\propto$ هستند، دارای میانگین طول گلبرگ 0.7 و انحراف معیار طول گلبرگ 0.1 هستند و این مقادیر برای گل های کلاس β به ترتیب برابر با 0.3 و 0.2 باشد. حال این طبقه‌بند با متغیرهای آماری مذکور، قصد تشخیص نمونه‌ای که دارای بردار ویژگی زیر است را دارد:

$Length$	$Width$
0.78	0.15

ستون‌های $Length$ و $Width$ همانطور که از نام آن‌ها برمی‌آید معرف طول و عرض گلبرگ مربوط به گل‌ها است. با توجه به اینکه طبقه بندی بر اساس طول گلبرگ است، مقدار این ویژگی را برای این نمونه بررسی می کنیم. بازه تخمینی برای کلاس $\propto$، برابر است با: $(0.7 - 0.1, 0.7 + 0.1)$. با توجه به اینکه طول گلبرگ این نمونه برابر با 0.78 است، پس در بازه تخمین کلاس $\propto$ حضور دارد؛ بنابراین کلاس این نمونه کلاس $\propto$ تعیین می شود. در صورتی که مقدار ویژگی مورد نظر نمونه ای

[4] Mean-Standard Deviation Classification

[5] Feature Vector

در هیچ یک از دو بازه تخمین دو کلاس (یا چند کلاس) حضور نداشته باشد، یکی از کلاس ها به صورت پیش فرض در نظر گرفته می شود.

مجموعه داده[6]

مجموعه داده‌ای که در این تمرین به شما داده شده‌است در قالب CSV[7] 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 خانه ها

اطلاعات مربوط به خانه ها در پرونده dataset.csv در اختیار شما قرار داده شده‌است. در ادامه درباره‌ی هر ویژگی و نوع داده[8]ی مربوط به آن، توضیح مختصری آمده‌است.

نام ویژگی	توضیح	نوع داده
LotArea	مساحت خانه در واحد فیت مربع	عدد صحیح

[6] Dataset

[7] Comma-Separated Values

[8] Data Type

OverallQual	کیفیت کلی متریال های به کار رفته در ساخت خانه	عدد صحیح (1 تا 10)
OverallCond	رتبه بندی وضعیت کلی خانه	عدد صحیح (1 تا 10)
YearBuilt	سال ساخت خانه	عدد صحیح
TotalBsmtSF	مساحت فضای زیرزمین در واحد فیت مربع	عدد صحیح
GrLivArea	مساحت فضای روزمین در واحد فیت مربع	عدد صحیح
GarageCars	ظرفیت پارکینگ خانه در واحد تعداد خودرو	عدد صحیح
GarageArea	مساحت پارکینگ در واحد فیت مربع	عدد صحیح
SalePrice	قیمت فروش خانه در واحد دلار	عدد صحیح

منابع

https://www.kaggle.com/lespin/house-prices-dataset

https://www.vectorstock.com/