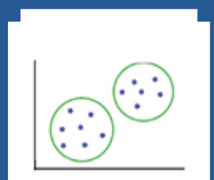
شناسایی سریع

تصویر ۱



تصویر ۲

الگوریتم‌های یادگیری ماشین به روش‌های «نظارتی، غیرنظارتی و نیمه‌نظارتی» تقسیم می‌شوند. در روش‌های نظارتی، هر نمونه از داده‌ها یک ویژگی به نام برچسب دارند. ویژگی برچسب، دسته یا کلاس نمونه‌ها را مشخص می‌کند. بیماران یک بیمارستان را در نظر بگیرید. هر بیمار ویژگی‌هایی نظیر فشارخون، قند خون، سن، جنسیت، کلسترول خون، قد و وزن دارد. ویژگی برچسب است که نوع بیماری فرد را مشخص می‌کند. هدف از یادگیری نظارتی، یادگیری ارتباط بین ویژگی‌هایی با ویژگی برچسب است، به‌گونه‌ای که اگر یک نمونهٔ جدید و بدون برچسب را به رایانه بدهیم، دستگاه بتواند برچسب آن نمونه را پیش‌بینی کند. برای مثال فرض کنید علائم یک بیمار را به دستگاه می‌دهیم و از رایانه می‌خواهیم بر اساس این علائم، نوع بیماری را پیش‌بینی کند. از آنجا که برچسب‌گذاری داده‌ها ممکن است کاری پرهزینه و زمان‌بر باشد، این نوع داده‌ها در اکثر محیط‌ها در دسترس نیستند. برای مثال، در برخی مواقع ممکن است پزشک متخصصی که بتواند نوع بیماری را شناسایی و درمان مناسب را تجویز کند، در دسترس نباشد. برای حل این مشکل، از نوع دیگری از فن‌ها با عنوان «یادگیری غیرنظارتی» استفاده قرار می‌گیرند. در ادامه روش‌های غیرنظارتی را بررسی کرده و کاربردهای آن را بررسی می‌کنیم

در روش‌های غیرنظارتی، داده‌ها بر اساس ویژگی‌های ذاتی‌شان به گروه‌های (خوشه‌های) متفاوت تقسیم می‌شوند. داده‌هایی که در یک گروه قرار می‌گیرند، بیشترین شباهت را به هم دارند و داده‌هایی که در گروه‌های متفاوتی قرار دارند، کمترین شباهت را به هم خواهند داشت. شکل یک و دو، تفاوت روش‌های نظارتی و غیرنظارتی را نشان می‌دهد:

در تصویر شماره ۱، داده‌ها دو ویژگی «yZ» دارند. بنابراین، هر نمونه را می‌توان در فضایی دوبعدی نمایش داد. علاوه بر این، داده‌ها یک ویژگی برچسب دارند که با رنگ نشان داده شده است. هدف از یادگیری نظارتی، ساخت یک خط، منحنی یا صفحه است که بتواند داده‌ها با رنگ‌های (برچسب‌های) متفاوت را از هم تفکیک کند. این منحنی به‌عنوان نمونهٔ نهایی شناخته می‌شود و می‌تواند برای پیش‌بینی برچسب داده‌های جدید مورد استفاده قرار گیرد. سامانه با دریافت یک نمونهٔ جدید، موقعیت آن را نسبت به نمونهٔ نهایی می‌سنجد. اگر آن نمونه در بالای خط قرار بگیرد، برچسب قرمز و در غیر این صورت برچسب آبی را به آن تخصیص می‌دهد. مدل ساخته‌شده ممکن است خطای جزئی داشته باشد و همهٔ داده‌ها را به‌درستی دسته‌بندی نکند. همان‌طور که در تصویر شمارهٔ ۱ مشخص است، یک نمونهٔ قرمز به‌اشتباه، در پایین منحنی و یک نمونهٔ آبی به‌اشتباه در بالای منحنی قرار دارد. این خطا دقت مدل را نشان می‌دهد. هر چه خطا کمتر باشد، پیش‌بینی برچسب داده‌های جدید قابل‌اعتمادتر است. لازم به ذکر است ، هزینهٔ خطای شناسایی آبی به‌عنوان قرمز با هزینهٔ خطای شناسایی قرمز به‌عنوان آبی برابر نیست. برای مثال فرض کنید می‌خواهیم بیماری کرونا را تشخیص دهیم. اگر فرد بیماری را به‌اشتباه سالم تشخیص دهیم و او را قرنطینه نکنیم، فرد مبتلابه بیماری کرونا سایرین را نیز مبتلا خواهد کرد. اما اگر فرد سالمی را به‌عنوان بیمار تشخیص دهیم، آن فرد چند روز در قرنطینه خواهد بود و نسبت به حالت اول هزینهٔ کمتری دارد. در تصویر شمارهٔ ۲، داده‌ها برچسب ندارند و تنها دو ویژگی «X,Y» دارند. هدف از این نوع یادگیری، گروه‌بندی داده‌ها بر اساس شباهت آن‌هاست. مفهوم شباهت با معیارهای متعددی مانند فاصلهٔ اقلیدسی، فاصلهٔ همینگ، شباهت کسینوسی و شباهت پیرسون ارزیابی می‌شود. در اینجا فاصلهٔ اقلیدسی بین داده‌ها به‌عنوان معیاری برای سنجش شباهت در نظر گرفته شده است. به این معنا که هر چه دو نمونه به هم نزدیک‌تر باشند، احتمالاً شباهت بیشتری به هم دارند و در یک خوشه قرار می‌گیرند. یادگیری غیرنظارتی، کاربردهای متعددی از جمله خوشه‌بندی، شناسایی ناهنجاری، تشخیص تقلب و نمونه‌برداری داده‌ها دارد که در ادامه به چند مورد از آن‌ها خواهیم پرداخت.

یکی از راه‌های تشخیص بیماری کرونا، انجام آزمایش (تست) کروناست. فرض کنید امکان گرفتن آزمایش کرونا برای همهٔ افراد، به دلیل محدودیت بسته (کیت) آزمایشگاهی، وجود نداشته باشد. از طرف دیگر می‌دانیم که علائم بیماری کرونا مشابه سرماخوردگی و آنفولانزا هستند. می‌خواهیم افرادی را که بیماری کرونا دارند شناسایی و قرنطینه کنیم. برای این کار از یادگیری غیرنظارتی استفاده می‌کنیم. با دریافت علائم بیماران، آن‌ها را به سه گروه تقسیم می‌کنیم؛ به‌طوری‌که بیماران با علائم مشابه در گروه یکسانی قرار بگیرند. سپس از هر گروه چند نفر به‌عنوان نمونه انتخاب می‌شوند و آزمایش کرونا روی آن‌ها انجام می‌شود. بر اساس نتایج آزمایش، همهٔ اعضای گروهی که نتیجهٔ آزمایش کرونا برای نمونه‌های آن‌ها مثبت باشد، به‌عنوان بیمار شناسایی و قرنطینه می‌شوند. این روش دقت کمتری دارد، اما در زمان کمتر و با هزینهٔ کمتری قابل انجام است. این یک مثال از خوشه‌بندی است که از قبل تعداد خوشه‌ها مشخص و به تعداد بیماری‌هاست (کرونا، سرماخوردگی و آنفولانزا).

در بسیاری از مسائل از قبل تعداد خوشه‌ها را نمی‌دانیم. به‌طور مثال فرض کنید قرار است اخبار یک سال گذشتهٔ یک سایت خبری را با توجه به محتوای خبر گروه‌بندی کنیم. یک راه ساده این است که از یک فرد خبره (متخصص) بخواهیم همهٔ هزاران خبر را بخواند و آن‌ها را از نظر مشابهت در گروه‌هایی قرار دهد. این روش مستلزم صرف زمان و هزینهٔ زیادی است. روش دیگر این است که اخبار را به یک ماشین بدهیم و با استفاده از یادگیری غیرنظارتی آن‌ها را به گروه‌های متعدد تقسیم کنیم. هر خبر بر اساس کلمات تشکیل‌دهنده‌اش در یک یا چند خوشه قرار می‌گیرد.در مثال تشخیص بیماری، هر نمونه (بیمار) به یک خوشه تعلق می‌گیرد که به آن خوشه‌بندی انحصاری می‌گویند. اما در مثال گروه‌بندی اخبار، ممکن است یک خبر به گروه‌های متفاوت متعلق باشد.

برای مثال، یک خبر می‌تواند هم سیاسی و هم اقتصادی باشد. در این مواقع خوشه‌بندی از نوع هم‌پوشان است.یکی دیگر از کاربردهای روش‌های غیرنظارتی، تشخیص ناهنجاری و شناسایی تقلب است. در تشخیص ناهنجاری، موارد (آیتم‌ها) یا رویدادهای غیرمنتظره در مجموعهٔ داده‌ها که با رفتار کلی سامانه متفاوت هستند، شناسایی می‌شوند. تشخیص ناهنجاری دو فرض اساسی دارد: ۱. ناهنجاری‌ها به‌ندرت در داده‌ها رخ می‌دهند؛ ۲. ویژگی‌های آن‌ها به‌طور قابل‌توجهی با نمونه‌های معمولی متفاوت هستند. فرض کنید، بانک الگویی عادی از واریز و برداشت از کارت شما دارد. اگر در یک روز مبالغ هنگفتی یکی پس از دیگری خرج شود و این رفتار معمول شما نباشد، بانک می‌تواند این موضوع را به‌عنوان ناهنجاری شناسایی و کارت شما را مسدود کند. زیرا حدس می‌زند احتمالاً سارقان اطلاعات کارت بانکی شما را به دست آورده و از آن استفاده می‌کنند.کاربرد دیگری را که می‌توان برای یادگیری غیرنظارتی در نظر گرفت، نمونه‌برداری داده‌هاست. فرض کنید میلیون‌ها داده داریم. از آنجا که بررسی همهٔ آن‌ها کار بسیار دشواری است، بنابراین، از روش‌های نمونه‌برداری استفاده می‌کنیم و مجموعهٔ کوچک‌تری از داده‌ها می‌سازیم که رفتاری مشابه رفتار کل داده‌ها داشته باشد و همهٔ تنوع داده‌های کل را حفظ کنند. برای این کار می‌توان داده‌ها را خوشه‌بندی و از هر خوشه نمایندگانی را به‌عنوان نمایندهٔ سایر اعضا انتخاب کرد.