

یادگیری آماری و کاربردهای آن در آمار رسمی

مجری

زهرا رضائی قهرودی

همکاران

حسن رنجبی

علیرضا رضایی



پژوهشکده‌ی آمار

گروه پژوهشی طرح‌های فنی و روش‌های آمار

بهار ۱۳۹۹

به نام خداوند جان و خرد

پیش‌گفتار

امروزه صحبت از علم داده و ضرورت استفاده از آن در تمامی ابعاد زندگی مطرح است. نوع، حجم و سرعت داده‌های تولیدی علاوه بر اینکه محرک اصلی در توسعه فناوری‌های مه‌داده‌هاست، باعث رونق دانشی تحت عنوان علم داده‌ها شده است. اصل و اساس علم داده، یادگیری بر اساس داده‌های موجود است که از طریق علم یادگیری آماری صورت می‌گیرد.

در سال‌های اخیر شاهد گسترده‌تری، تنوع و سرعت تولید اطلاعات در حوزه‌های مختلف از جمله آمار رسمی هستیم. با توجه به ضرورت جایگزینی پردازش محاسباتی ارزان‌تر، سریع‌تر و دقیق‌تر رایانه‌ها به جای انسان‌ها، موضوع یادگیری ماشین و یادگیری آماری از اهمیت بیشتری برخوردار شده است. پیشرفت‌های اخیر در یادگیری ماشین هم در خصوص توسعه الگوریتم‌ها و نظریه‌های جدید یادگیری بوده است و هم در انقلاب داده‌ها و در دسترس بودن داده‌های آنلاین و آمارهای ثبته‌مبنا با انجام محاسبات کم‌هزینه بوده است. بنا بر این اتخاذ روش‌های یادگیری آماری قدرتمند منجر به تصمیم‌گیری شواهدمبنا در بسیاری از زمینه‌ها از جمله آمار رسمی می‌شود.

یادگیری ماشین و یادگیری آماری، مجموعه وسیعی از ابزارها برای درک داده‌ها هستند که این ابزارها به دو دسته‌ی راهنماییده^۱ و ناراهنماییده^۲ تقسیم می‌شوند. از طرفی یادگیری آماری به عنوان یکی از شاخه‌های پرکاربرد هوش مصنوعی، با استخراج قوانین معنی‌دار از داده‌های خام ذخیره شده، مبنای علمی و فنی مناسبی را برای دانش داده‌کاوی ایجاد نموده است. همچنین یادگیری آماری یک شاخه از آمار کاربردی است که در پاسخ به یادگیری ماشین ظاهر شده است و بر مدل‌های آماری و ارزیابی عدم حتمیت تاکید دارد. یادگیری ماشین نیز الگوریتم‌هایی را ایجاد می‌کند تا بتواند یادگیری از داده‌ها را داشته باشد. روش‌های یادگیری آماری شامل شیوه‌ها و الگوریتم‌هایی است که براساس آنها رایانه‌ها به منظور کشف رفتار داده‌ها، توانایی یادگیری پیدا می‌کنند.

در سال‌های اخیر، پیشرفت‌های زیادی در یادگیری آماری با افزایش دسترسی به نرم‌افزارهای قدرتمند و نسبتاً کاربر پسند، بوجود آمده است. روش‌های اصلی آماری در یادگیری آماری را می‌توان به ۳ دسته‌ی رگرسیون، رده‌بندی و خوشه‌بندی تقسیم کرد که دو روش اول مربوط به روش‌های یادگیری راهنماییده و روش

^۱ supervised learning

^۲ unsupervised learning

آخر مربوط به روش یادگیری ناراهنماییده است که بسته به نوع متغیر مورد بررسی (کیفی در مقابل کمی) فنون یادگیری راهنماییده منجر به رده‌بندی یا رگرسیون و یادگیری ناراهنمایید منجر به خوشه‌بندی می‌شود. روش‌های یادگیری آماری در بسیاری از فرایندهای تولید داده نیز مورد استفاده قرار می‌گیرد.

استفاده از روش یادگیری آماری خوشه‌بندی برای انجام اتصال رکوردها به منظور چارچوب‌سازی و یکپارچه‌سازی داده‌ها، استفاده از ابزار یادگیری آماری رده‌بندی برای کدگذاری رشته فعالیت‌ها، مناطق جغرافیایی، مشاغل و ... استفاده از ابزار یادگیری آماری رده‌بندی و رگرسیون برای جانمایی داده‌های گم‌شده، پیش‌بینی واکنش‌های پاسخگویی، ساخت گروه‌های همگن برای جانمایی، وزن‌دهی مجدد، کالبره یا طبقه‌بندی، استفاده از ابزار یادگیری آماری خوشه‌بندی برای شناسایی نقاط دورافتاده و استفاده از ابزار یادگیری آماری رده‌بندی و رگرسیون برای کنترل افشای داده‌ها از مثال‌های کاربردی یادگیری آماری در آمار رسمی است.

در این طرح مطالعاتی ضمن مرور مفاهیم یادگیری آماری و آشنایی با روش‌های یادگیری آماری، به معرفی روش‌های یادگیری آماری در آمار رسمی و بیان تجربه‌ی کشورهای مختلف در استفاده از روش‌های یادگیری آماری در آمار رسمی (روش‌های کدگذاری، جورسازی داده‌ها، چارچوب‌سازی ...) پرداخته شده است. همچنین سه کاربرد از روش‌های یادگیری آماری در متن‌کاوی شامل کدگذاری خودکار رشته فعالیت‌های اقتصادی، تخصیص کد واجد شرایط بودن یا نبودن به پرسش باز عدم تکمیل پرسشنامه و انتساب کد آماری به آدرس‌های پستی به صورت خودکار و بدون انجام عملیات میدانی، با استفاده از نرم‌افزارهای R و SAS انجام شده است. در مسائل کاربردی اشاره شده، از روش‌های متن‌کاوی برای رده‌بندی پرسش‌های باز استفاده شده است. در مسائل مربوط به تخصیص کد صحیح ISIC یا ISCO یا هر کد دیگر به پرسش‌های باز به صورت خودکار، با تشکیل یک دیکشنری جامع و کامل با استفاده از کدگذاری کتابچه‌های رده‌بندی‌های بین‌المللی مانند رده‌بندی رشته فعالیت‌های اقتصادی و دست‌نوشته‌های مأموران آمارگیری از چند آمارگیری قبلی، امکان کدگذاری خودکار رشته فعالیت‌های اقتصادی به صورت نیمه‌خودکار فراهم می‌شود. کاربرد دیگر متن‌کاوی در پرسش‌های باز، کدگذاری متون نوشته شده در پرسش‌هایی است که یکی از رده‌های آن «سایر با ذکر نام» است. با توجه به اینکه رده‌بندی متون گزینه‌ی «سایر با ذکر علت» نیاز به بررسی و درج کد دارد و انجام این کار به صورت دستی زمان‌بر است، با استفاده از روش‌های یادگیری آماری، امکان اختصاص کد به هر متن نوشته شده در سایر، به صورت نیمه‌خودکار وجود دارد. مثال کاربردی دیگر انتساب آدرس‌های آماری به آدرس‌های پستی به روش خودکار با استفاده از روش‌های یادگیری آماری است که در سرشماری ثبتی مبنا کاربرد دارد. با اتصال آدرس آماری به آدرس‌های پستی، امکان برقراری ارتباط بین سرشماری ثبتی مبنا با سرشماری‌های سنتی قبلی و ارائه‌ی اطلاعات سرشماری ثبتی مبنا به صورت سری‌های زمانی در پایین‌ترین سطوح جغرافیایی نیز فراهم می‌شود. در ایران بیش از ۲۰ درصد کدهای آماری در مرکز آمار منتسب به آدرس‌های پستی نیست. با استفاده از روش‌های یادگیری آماری و آموزش مدل با استفاده از ۸۰ درصد کدهای آماری منتسب به آدرس‌های پستی، امکان انتساب کد آماری به آدرس‌های پستی منطبق نشده فراهم می‌شود.

گروه پژوهشی طرح‌های فنی و روش‌های آمار
پژوهشکده‌ی آمار

فهرست مطالب

آشنایی با مفاهیم یادگیری آماری.....	۱
۱-۱- کلیات.....	۱
۲-۱- تفاوت بین داده‌کاوی، هوش مصنوعی، یادگیری ماشین و یادگیری عمیق.....	۴
آشنایی با روش‌های یادگیری آماری.....	۷
۱-۲- مقدمه.....	۷
۲-۲- آشنایی با روش‌های راهنماییده.....	۸
۱-۲-۲- رگرسیون.....	۹
۲-۲-۲- رده‌بندی.....	۱۳
۳-۲- آشنایی با روش‌های ناراهنماییده.....	۳۲
۱-۳-۲- تحلیل مؤلفه‌های اصلی.....	۳۳
۲-۳-۲- روش‌های خوشه‌بندی.....	۳۷
یادگیری آماری در آمارگیری‌ها.....	۴۷
۱-۳- مقدمه.....	۴۷
۲-۳- مدل عمومی فرایند کسب و کار آماری.....	۴۷
۱-۲-۳- کاربرد روش‌های یادگیری آماری در داده‌های اولیه.....	۵۰
۲-۲-۳- کاربرد روش‌های یادگیری آماری در داده‌های ثانویه.....	۵۳
۳-۳- پردازش زبان‌های طبیعی.....	۵۶
۱-۳-۳- متن‌کاوی.....	۵۸
کاربرد یادگیری آماری در فعالیت‌های مراکز آماری.....	۶۱
۱-۴- مطالعات تطبیقی کدگذاری خودکار به روش یادگیری آماری در مراکز آماری.....	۶۱
۲-۴- کاربرد یادگیری آماری در آمارگیری‌ها و فعالیت‌های مرکز آمار ایران.....	۶۴
۱-۲-۴- روش‌های کدگذاری نیمه‌خودکار رشته فعالیت‌های اقتصادی.....	۶۵
۲-۲-۴- تشخیص خودکار واجد شرایط بودن کارگاه‌ها از پرسش‌های باز.....	۷۴
۳-۲-۴- اتصال کدپستی با آدرس آماری به کمک روش‌های یادگیری آماری.....	۷۹
۴-۲-۴- نتیجه‌گیری.....	۸۴
پیوست‌ها.....	۸۷
پیوست الف: تبدیل متون فارسی در دیکشنری به متغیرهای تک‌نگاشت با نرم‌افزار SAS.....	۸۷
پیوست ب: کد گذاری مشاغل یا رشته فعالیت اقتصادی با استفاده از نرم‌افزار R.....	۹۲
پیوست ج: اتصال کد پستی به آدرس آماری با استفاده از نرم‌افزار SAS.....	۱۰۳
مرجع.....	۱۱۵

فهرست جدول‌ها

- جدول ۱-۲-۱- خلاصه‌ای از ۴ نوع پیوند معمول در خوشه‌بندی سلسله‌مراتبی ۴۴
- جدول ۱-۳-۱- سطوح ۱ و ۲ مدل عمومی فرایند کسب و کار آماری ۴۹
- جدول ۲-۳-۲- کاربرد یادگیری آماری در فعالیت‌های مرتبط با داده‌های اولیه ۵۲
- جدول ۳-۳-۳- ماتریس DTM ۵۹
- جدول ۱-۴-۱- ماتریس DTM داده‌های آموزشی ۶۶
- جدول ۲-۴-۲- ماتریس DTM داده‌های آزمایشی ۶۷
- جدول ۳-۴-۳- تعداد تکرارهای داده‌های آزمایشی با داده‌های آموزشی ۶۷
- جدول ۴-۴-۴- ماتریس DTM داده‌های آزمایشی برای محاسبه‌ی معیار تشابه cosine ۷۰
- جدول ۵-۴-۵- ماتریس DTM داده‌های آموزشی برای محاسبه‌ی معیار تشابه cosine ۷۰
- جدول ۶-۴-۶- محاسبه‌ی $\gamma ci|x$ متغیرهای تک‌نگاشت مقدار ۱ را در صورتی که لغت در متن وجود داشته باشد و صفر در بقیه موارد را اخذ می‌کند ۷۲
- جدول ۴ - ۷- میزان درستی پنج روش برای داده‌های آزمایشی رشته فعالیت اقتصادی کارگاه‌های صنعتی با نرخ آموزشی ۰/۸ به تفکیک احتمال انتساب کد ISIC به رشته فعالیت اقتصادی ۷۳
- جدول ۴-۸- نمونه‌ای از ماتریس درهم‌ریختگی یک ماتریس 2×2 . دو رده‌ی واقعی، P و N در نظر گرفته شده است و خروجی رده‌ی پیش‌بینی شده درست یا غلط در نظر گرفته شده است ۷۵
- جدول ۴-۹- جدول توافقی مقادیر واقعی و مقادیر پیش‌بینی ۷۶
- جدول ۴-۱۰- ماتریس در هم‌ریختگی روش تکرار (نرخ داده‌ی آموزش = ۰/۸) ۷۷
- جدول ۴-۱۱- ماتریس در هم‌ریختگی روش‌های ترکیبی، هیبرید و نزدیک‌ترین همسایه (نرخ داده‌ی آموزش = ۰/۸) ۷۷
- جدول ۴-۱۲- مقایسه‌ی شاخص‌های ارزیابی عملکرد الگوریتم‌های مختلف (نرخ داده‌ی آموزش: ۰/۸) ۷۷
- جدول ۴-۱۳- ماتریس در هم‌ریختگی روش تکرار (نرخ داده‌ی آموزش = ۰/۷) ۷۸
- جدول ۴-۱۴- ماتریس در هم‌ریختگی روش‌های ترکیبی و هیبرید (نرخ داده‌ی آموزش = ۰/۷) ۷۸
- جدول ۴-۱۵- ماتریس در هم‌ریختگی روش نزدیک‌ترین همسایه (نرخ داده‌ی آموزش = ۰/۷) ۷۸
- جدول ۴-۱۶- مقایسه‌ی شاخص‌های ارزیابی عملکرد الگوریتم‌های مختلف (نرخ داده‌ی آموزش: ۰/۷) ۷۸
- جدول ۴-۱۷- تعداد رکوردهای دارای پیوند بین آدرس پستی و آدرس آماری برای خانوارهای شهرهای قم و زنجان ۸۰
- جدول ۴-۱۸- میزان دقت پنج روش برای داده‌های آزمایشی اتصال کدپستی با آدرس آماری شهر قم با نرخ آموزشی ۰/۷ به تفکیک احتمال انتساب (کدپستی به آدرس آماری) ۸۳
- جدول ۴-۱۹- میزان دقت پنج روش برای داده‌های آزمایشی اتصال کدپستی با آدرس آماری شهر زنجان با نرخ آموزشی ۰/۷ به تفکیک احتمال انتساب (کدپستی به آدرس آماری) ۸۳
- جدول ۴-۲۰- میزان دقت روش تکرار داده‌های آزمایشی اتصال کدپستی با آدرس آماری شهر قم به تفکیک احتمال انتساب (کدپستی به آدرس آماری) و نرخ داده‌های آموزشی ۰/۷، ۰/۸، ۰/۸۵ و ۰/۹ ۸۴

فهرست شکل‌ها

- شکل ۱-۱-۱- مقایسه‌ی داده‌های بُعدبالا و بعدپایین ۲
- شکل ۲-۱-۲- مقایسه‌ی یادگیری راهنماییده و یادگیری ناراهنماییده ۳
- شکل ۳-۱-۳- فرایند یادگیری آماری راهنماییده براساس داده‌های آموزشی و آزمایشی ۳
- شکل ۴-۱-۴- ارتباط هوش مصنوعی، یادگیری ماشین و یادگیری عمیق ۵
- شکل ۱-۲-۱- مقایسه‌ی خطای داده‌های آزمایشی، خطای داده‌های آموزشی و R^2 با افزایش تعداد متغیرهای کمکی ۱۰
- شکل ۲-۲-۲- رابطه‌ی تعداد متغیرهای کمکی و مدل، R^2 ، خطای داده‌های آموزشی و خطای داده‌های آزمایشی ۱۱
- شکل ۳-۲-۳- رابطه‌ی پیچیدگی مدل و بیش‌برازشی ۱۱
- شکل ۴-۲-۴- روش‌های اعتبارسنجی (رنگ قرمز: داده‌های اعتبارسنجی و رنگ آبی: داده‌های آموزشی) ۱۲
- شکل ۵-۲-۵- نمایش گرافیکی روش K - نزدیکترین همسایه با $K = ۳$ ۱۴
- شکل ۶-۲-۶- داده‌های دو رده با تفکیک‌کننده‌ی A و B ۱۶
- شکل ۷-۲-۷- ماشین بردار پشتیبان به‌عنوان یک ابرصفحه برای جداسازی خطی نمونه‌ها در فضای داده‌ها. ۱۷
- شکل ۸-۲-۸- انتقال داده‌ها از فضای دوبعدی به سه‌بعدی ۱۸
- شکل ۹-۲-۹- مثالی از ماشین بردار پشتیبان چندرده‌ای ۱۸
- شکل ۱۰-۲-۱۰- خط جداکننده و ایجاد حاشیه با مقدار مثبت و منفی C ۱۹
- شکل ۱۱-۲-۱۱- بردار پشتیبان دو رده ۲۰
- شکل ۱۲-۲-۱۲- نمودار چپ: جداکننده‌ی بردار پشتیبان به یک مجموعه داده‌ی کوچک. ابر صفحه به عنوان خط مشکی و مرزها یا حاشیه‌ها به صورت خط چین تیره. در مورد مشاهدات بنفش، مشاهدات ۳، ۴ و ۵ و ۶ طرف صحیح حاشیه‌ها، مشاهده ۲ روی حاشیه و مشاهده ۱ طرف ناصحیح حاشیه. در مورد مشاهدات آبی، مشاهدات ۷ و ۱۰ طرف صحیح حاشیه‌ها، مشاهده ۹ روی حاشیه و مشاهده ۸ طرف ناصحیح حاشیه. هیچ مشاهده‌ای در طرف نادرست ابرصفحه نیست. نمودار راست: مشابه نمودار چپ با دو نقطه‌ی اضافی ۱۱ و ۱۲. این دو مشاهده در طرف چپ ابرصفحه و طرف اشتباه حاشیه قرار دارند. ۲۴
- شکل ۱۳-۲-۱۳- جداسازی دو رده با استفاده از متغیر ξ ۲۴
- شکل ۱۴-۲-۱۴- شکل چپ: مشاهدات در دو رده با یک مرز غیرخطی بین آنها، قرار دارند. شکل راست: جداکننده‌ی بردار پشتیبان، به دنبال یک مرز خطی است و در نتیجه به نتیجه‌ی ضعیفی دست یافته است. ۲۵
- شکل ۱۵-۲-۱۵- انتقال داده‌ها به فضای متغیرهای با ابعاد بالاتر ۲۶

- شکل ۲-۱۶- شکل چپ: ماشین بردار پشتیبان با هسته از درجه ۳، شکل راست: ماشین بردار پشتیبان با هسته‌ی شعاعی..... ۲۸
- شکل ۲-۱۷- مثالی از درخت تصمیم ۲۹
- شکل ۲-۱۸- درخت تصمیم برای داده‌های پزشکی ۲۹
- شکل ۲-۱۹- اندازه‌ی جمعیت (Pop) و درآمد (Ad) برای ۱۰۰ شهر مختلف به صورت دایره‌های بنفش نمایش داده شده است. خط ممتد نشان‌دهنده‌ی مؤلفه‌ی اصلی اول و خط چین نشان‌دهنده‌ی مؤلفه‌ی اصلی دوم است..... ۳۵
- شکل ۲-۲۰- اندازه‌ی نمودار داده‌های مقیاس‌بندی شدی در مقابل داده‌های مقیاس‌بندی نشده ۳۵
- شکل ۲-۲۱- سمت چپ: نمودار نسبت واریانس بیان‌شده (PVE) با چهار مؤلفه اصلی در داده‌های USArrests سمت راست: نمودار تجمعی نسبت واریانس بیان‌شده با چهار مؤلفه اصلی در مجموعه داده‌های USArrests..... ۳۶
- شکل ۲-۲۲- خوشه‌بندی..... ۳۸
- شکل ۲-۲۳- مجموعه داده‌های شبیه‌سازی شده با ۱۵۰ مشاهده در فضای دو بُعدی. دسته‌ها نتایج حاصل از استفاده از خوشه‌بندی K- میانگین با مقادیر مختلف K را نشان می‌دهد. رنگ هر مشاهده، خوشه‌ای را که با استفاده از الگوریتم خوشه‌بندی K- میانگین به آن اختصاص داده شده است را نشان می‌دهد. ۳۹
- شکل ۲-۲۴- پیشرفت الگوریتم K- میانگین در مثال شکل ۲-۲۳ با $K=3$. بالا سمت چپ: مشاهدات نشان داده شده است. بالا وسط: در مرحله ۱ الگوریتم، هر مشاهده به طور تصادفی به یک خوشه اختصاص داده شده است. بالا سمت راست: در مرحله ۲ (الف)، مرکز خوشه‌ها محاسبه شده است که این مراکز به عنوان دیسک‌های رنگی بزرگ نشان داده شده است. در ابتدا مرکزها تقریباً کاملاً با هم همپوشانی دارند زیرا خوشه‌بندی اولیه به طور تصادفی انتخاب شده است. پایین سمت چپ: در مرحله ۲ (ب)، هر مشاهده به نزدیکترین مرکز خوشه‌ای اختصاص داده شده است. پایین وسط: مرحله ۲ (الف) بار دیگر انجام شده است و منجر به تولید مرکز خوشه‌های جدید شده است. پایین سمت راست: نتایج به دست آمده پس از ده تکرار..... ۴۱
- شکل ۲-۲۵- چهل و پنج مشاهده در فضای دو بُعدی تولید شده است. در واقعیت سه کلاس مجزا وجود دارد که با رنگ‌های جداگانه نشان داده شده‌اند. با این حال، با این که رده‌ها برچسب‌گذاری شده‌اند، برچسب رده‌ها را ناشناخته تلقی کرده و به دنبال طبقه‌بندی مشاهدات خواهیم بود تا رده‌ها از داده‌ها کشف شود ۴۲
- شکل ۲-۲۶- نمایش یک درختواره: سمت چپ: درختواره به دست آمده از خوشه‌بندی سلسله مراتبی داده‌های شکل ۲-۲۵ با پیوند کامل و فاصله اقلیدسی. مرکز: درختواره از شکل سمت چپ، برش در ارتفاع نه (که با خط چین مشخص شده است). این برش منجر به دو خوشه مجزا می‌شود که در رنگ‌های مختلف نشان داده شده است. راست: درختواره از شکل سمت چپ که در ارتفاع پنج برش خورده است. این برش منجر به سه خوشه مجزا می‌شود که در رنگ‌های مختلف نشان داده شده است. ۴۲
- شکل ۲-۲۷- پیوند میانگین، کامل و منفرد برای یک مجموعه داده. پیوند میانگین و کامل تمایل به ایجاد خوشه‌های متعادل‌تر دارد ۴۴
- شکل ۳-۱- مقایسه‌ی مراحل اجرای طرح‌های آمارگیری مرکز آمار ایران با فرایندهای GSBPM ۵۰
- شکل ۴-۱- میزان دقت برای نرخ‌های تولید مشخص به تفکیک پنج روش برای کدگذاری رشته فعالیت اقتصادی با نرخ داده‌ی آموزشی ۰۸ ۷۳

شکل ۴-۲- میزان دقت برای نرخ‌های تولید مشخص به تفکیک چهار روش و دو نرخ مختلف داده‌ی آموزشی الف) نرخ داده‌ی آموزشی ۰/۸ و ب) نرخ داده‌ی آموزشی ۰/۷ ۷۶

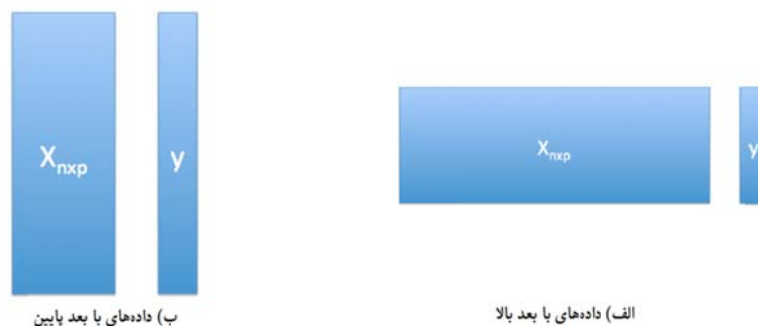
شکل ۴-۳- میزان دقت برای نرخ‌های تولید مشخص به تفکیک پنج روش برای اتصال کدپستی با آدرس آماری با نرخ داده‌ی آموزشی ۰/۷ برای داده‌های آزمایشی قم الف) و زنجان ب) ۸۲

آشنایی با مفاهیم یادگیری آماری

۱-۱- کلیات

امروزه صحبت از علم داده و ضرورت استفاده از آن در تمامی ابعاد زندگی مطرح است. نوع، حجم و سرعت داده‌های تولیدی علاوه بر اینکه محرک اصلی در توسعه فناوری‌های مه‌داده‌هاست، باعث رونق دانشی تحت عنوان علم داده‌ها شده است. علم داده یک علم میان رشته‌ای است که از ابتدای تولید داده، نحوه‌ی گردآوری، ذخیره‌سازی و ... تا کشف دانش از داده‌ها را بر اساس استانداردها و دستورالعمل‌های مناسب معرفی می‌کند. اصل و اساس علم داده، یادگیری بر اساس داده‌های موجود است که از طریق علم یادگیری آماری صورت می‌گیرد. زمانی که با مه‌داده روبرو هستیم، با استفاده از روش‌های یادگیری آماری، دخالت انسان برای انجام تحلیل و محاسبات حذف می‌شود و با طراحی الگوریتم‌ها و استفاده از رایانه، امکان افزایش سرعت محاسبات، کاهش هزینه‌ها و جلوگیری از متوقف شدن اجرای برنامه‌ها برای محاسبات فراهم می‌شود.

از طرفی به دلیل اینکه در حال حاضر نوع، حجم و سرعت داده‌های تولیدی بسیار بالا است و با داده‌های بُعد بالا ($n \ll p$) در مقابل داده‌های بُعد پایین ($n \gg p$) روبرو هستیم (شکل ۱-۱) که در آن‌ها n ، تعداد مشاهدات بسیار کمتر از p ، تعداد متغیرها، پیشگوها، مشخصه‌ها یا ویژگی‌ها است. بنا بر این، استفاده از روش‌های مبتنی بر فناوری اطلاعات و یادگیری آماری جایگزین مناسبی برای روش‌های کلاسیک آماری خواهد بود.



شکل ۱-۱- مقایسه‌ی داده‌های بُعد بالا و بعد پایین

یادگیری آماری، مجموعه‌ی وسیعی از ابزارها برای درک داده‌ها است که این ابزارها به دو دسته‌ی راهنماییده و ناراهنماییده تقسیم می‌شوند. از طرفی یادگیری آماری یک شاخه از آمار کاربردی است که در پاسخ به یادگیری ماشین ظاهر شده است و بر مدل‌های آماری و ارزیابی عدم حتمیت تأکید دارد.

یادگیری راهنماییده شامل ساختن یک مدل آماری برای پیش‌گویی و برآورد خروجی براساس یک یا چند ورودی است. از دیدگاه تئوری یادگیری آماری، یادگیری راهنماییده شامل یادگیری از یک مجموعه‌داده‌ی آموزشی^۳ است که در آن هر نقطه در مجموعه‌داده‌ی آموزشی، یک جفت ورودی/خروجی است، به طوری که در آن ورودی به یک خروجی نگاشت می‌شود. پس از یادگیری یک تابع بر اساس مجموعه‌داده‌های آموزشی، آن تابع بر روی یک مجموعه‌داده‌ی آزمایشی^۴ (که در مجموعه داده‌های آموزشی ظاهر نشده است) اعمال می‌شود. بسته به نوع خروجی، مسائل یادگیری راهنماییده، به رگرسیون یا رده‌بندی تقسیم می‌شوند. اگر خروجی مقادیر پیوسته را اخذ کند، راه حل، رگرسیون است و اگر خروجی مقادیر گسسته باشد، راه حل، روش رده‌بندی است.

یادگیری ناراهنماییده نوعی الگوریتم یادگیری آماری است که برای به دست آوردن استنباط از مجموعه‌داده‌های ورودی بوجود آمده است. به عبارت دیگر در یادگیری ناراهنماییده، تنها داده‌ی ورودی وجود دارد و متغیرهای خروجی متناظر وجود ندارند. در این روش، یادگیری در خصوص ارتباط و ساختار داده‌ها است. معمول‌ترین روش‌های یادگیری ناراهنماییده، کاهش بُعد و خوشه‌بندی است که برای تحلیل کاوشگرانه‌ی داده‌ها و پیدا کردن الگوهای پنهان یا گروه‌بندی داده‌ها مورد استفاده قرار می‌گیرد. خوشه‌بندی و کاهش بُعد برخی روش‌های یادگیری ناراهنماییده هستند. در مسائل آمار رسمی، برای شناسایی داده‌های دورافتاده یا ناهنجار نسبت به سایر داده‌ها می‌توان با استفاده از روش‌های یادگیری آماری خوشه‌بندی به گروه‌بندی داده‌ها پرداخت. هدف این‌گونه مسئله‌ها پیش‌گویی متغیر خروجی نیست و راه حل این مسئله در یادگیری آماری، به خوشه‌بندی معروف است. خوشه‌بندی داده‌های بیان ژن که از مثال‌های کاربردی با داده‌های بُعد بالا است (داده‌هایی که در آن تعداد متغیرها بسیار بیشتر از تعداد مشاهدات است) و در حوزه‌ی پزشکی کاربرد دارد نیز از طریق روش‌های یادگیری ناراهنماییده قابل بحث و بررسی است. شکل ۱-۲ به مقایسه‌ی یادگیری راهنماییده و ناراهنماییده پرداخته است. شکل ۱-۳ نیز به معرفی فرایند یادگیری آماری راهنماییده بر اساس داده‌های آموزشی و آزمایشی می‌پردازد. در شکل ۱-۳ فرایند یادگیری آماری راهنماییده به شش مرحله تقسیم شده است.

^۳training dataset

^۴testing dataset

مرحله ۱. داده‌های ورودی برچسب‌گذاری شده با B به طور تصادفی به مجموعه داده‌ی آموزشی (آبی) و مجموعه داده‌ی آزمایشی (بنفش) تقسیم شده است.

مرحله ۲. مدل یا الگوریتم در نظر گرفته شده، براساس ارتباط بین x و y در مجموعه داده‌ی آموزشی، می‌آموزد.

مرحله ۳. از مدل یا الگوریتم برای پیش‌گویی y ، \hat{y} (نارنجی) از x در مجموعه داده‌ی آزمایشی استفاده می‌شود.

مرحله ۴. مقادیر پیش‌گویی شده \hat{y} با استفاده از مقادیر مشاهده شده y در مجموعه داده‌ی آزمایشی، ارزیابی می‌شود. مراحل ۲ تا ۴ با پارامترهای مختلف تکرار می‌شوند تا خطای پیش‌گویی حداقل شود.

مرحله ۵. مراحل ۱ تا ۴ براساس تجزیه‌های مختلف داده‌ها تکرار می‌شود تا از بیش‌برازشی جلوگیری شود.

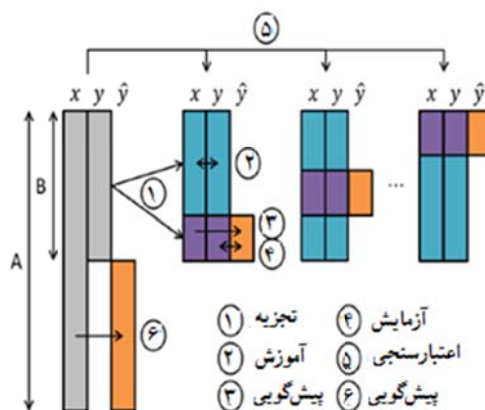
مرحله ۶. مقادیر y مشاهده نشده خارج از نمونه (برچسب‌گذاری شده با A) با استفاده از مدل یا الگوریتم معرفی شده، پیش‌بینی می‌شود.

یادگیری آماری عمدتاً به معرفی مهم‌ترین روش‌های مدل‌سازی و پیش‌بینی براساس روش‌های رگرسیون و رده‌بندی می‌پردازد. عمده روش‌هایی که در یادگیری آماری مورد استفاده قرار می‌گیرند عبارتند از رگرسیون خطی و چندجمله‌ای، رده‌بندی، رگرسیون لوژستیک و تحلیل تشخیصی خطی، مدل‌های غیرخطی، روش‌های مبتنی بر درخت‌های تصادفی و ... (هستی و دیگران، ۲۰۰۹؛ جیمز و دیگران، ۲۰۱۴). با توجه به وجود روش‌های مختلف یادگیری آماری، انتخاب بهترین روش یادگیری آماری با استفاده از اعتبارسنجی متقابل و روش‌های خودگردان‌سازی انجام می‌گیرد.



الف) یادگیری راهنماییده ب) یادگیری ناراهنماییده

شکل ۱-۲- مقایسه‌ی یادگیری راهنماییده و یادگیری ناراهنماییده



شکل ۱-۳- فرایند یادگیری آماری راهنماییده براساس داده‌های آموزشی و آزمایشی