یادگیری آماری و کاربردهای آن در آمار رسمی

مجری

زهرا رضائی قهرودی

همکاران

حسن رنجی

علیرضا رضایی

پژوهشکده‌ی آمار

گروه پژوهشی طرح‌های فنی و روش‌های آماری

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 گستردگی، 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 ثبتی‌مبنا 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 جانهی داده‌های گم‌شده، پیش‌بینی واکنش‌های پاسخگویی، ساخت گروه‌های همگن برای جانهی، وزن‌دهی مجدد، کالیبره 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 ثبتی‌مبنا کاربرد دارد. با اتصال آدرس آماری به آدرس‌های پستی، امکان برقراری ارتباط بین سرشماری ثبتی‌مبنا با سرشماری‌های سنتی قبلی و ارائه‌ی اطلاعات سرشماری ثبتی‌مبنا 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 آماری
پژوهشکده‌ی آمار

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

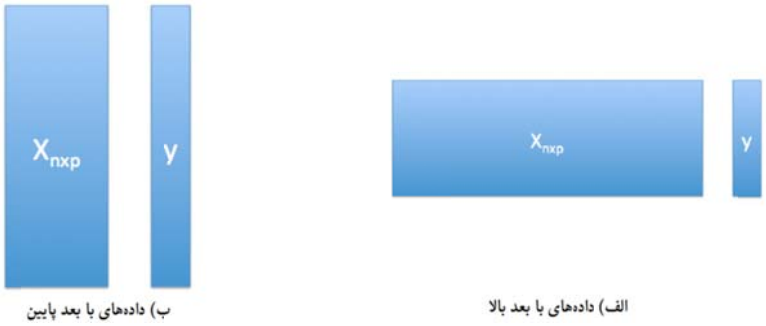
۱

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 مِه‌داده‌هاست، 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 مِه‌داده 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 بُعدبالا ($n \ll p$) در مقابل داده‌های بُعدپایین ($n \gg p$) 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 بُعدبالا و بعدپایین

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 عدم‌حتمیت تأکید دارد.

یادگیری راهنماییده شامل ساختن یک مدل آماری برای پیش‌گویی و براورد 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 شناسائی 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 بُعدبالا 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.

---

[۳] training dataset

[۴] testing dataset

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

مرحله‌ی ۲. مدل یا الگوریتم درنظر گرفته شده، براساس ارتباط بین x و y در مجموعه‌داده‌ی آموزشی، می‌آموزد.

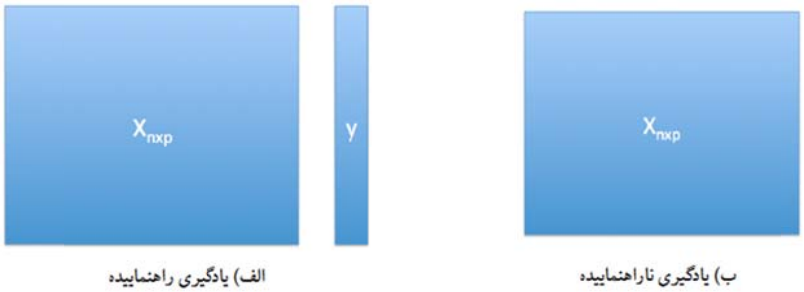
مرحله‌ی ۳. از مدل یا الگوریتم برای پیش‌گویی y ، $\hat{y}$ (نارنجی) از x در مجموعه‌داده‌ی آزمایشی استفاده می‌شود.

مرحله‌ی ۴. مقادیر پیش‌گویی‌شده $\hat{y}$ با استفاده از مقادیر مشاهده‌شده‌ی y در مجموعه‌داده‌ی آزمایشی، ارزیابی می‌شود. مراحل ۲ تا ۴ با پارامترهای مختلف تکرار می‌شوند تا خطای پیش‌گویی حداقل شود.

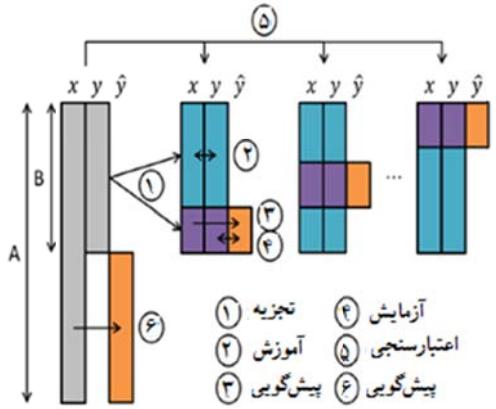
مرحله‌ی ۵. مراحل ۱ تا ۴ براساس تجزی‌های مختلف داده‌ها تکرار می‌شود تا از بیش‌برازشی جلوگیری شود.

مرحله‌ی 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.



شکل ۱-۲- مقایسه‌ی یادگیری راهنماییده و یادگیری ناراهنماییده



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