کاربرد آمار در داده کاوی

مقدمه و مقا یسه

آمار شاخه ای از علم ریاضی است که به جمع آوری توضیح و تفسیر داده ها می پردازد.[۳] این مبحث به گونه ای است که روزانه کاربرد زیادی دارد. در مقایسه این عام با data mining قدمت بیشتری دارد و جزء ورشهای کلاسیک داده کاوی محسوب می شود،وجه اشتراک تکنیکهای آماری وdata mining بیشتر درتخمین وپیش بینی است.[۲]البته از آزمونهای آماری در ارزیابی نتایج داده کاوی نیزاستفاده می شود. درکل ا گر تخمین و پیش بینی جزء وظایف data mining در نظر گرفته شوند،تحلیل های آماری،data mining را بیش از یک قرن اجرا کرده است.به عقیده بعضی DM ابتدا ازآمار و تحلیل های آماری تحلیل شروع شد. [۲] می توان تحلیل های آماری از قبیل فاصله اطمینان،رگرسیون و... را مقدمه و پیش زمینه DM را دانست که بتدریج در زمینه های دیگر ومتد های دیگررشد و توسعه پیدا کرد. پس در واقع متدهای آماری جزو روشهای کلاسیک و قدیمی DM محسوب می شوند.در جایی اینگونه بحث می شود که با تعریف دقیق ، آماریا تکنیکهای آماری جزء داده کاوی(data mining) نیستند.این روشها خیلی قبل تر از data mining استفاده می شدند.با این وجود، تکنیکها آماری توسط داده ها بکار برده می شوند و برای کشف موضوعات و ساختن مدلهای پیشگویانه مورد استفاده قرار می گیرند.[۳]

همانگونه که واضح و مشخص است با گذشت زمان علم نیز پیشرفت می کند،هر چه به جلوتر می رویم روشهای جدید تر و بهترمورد استفاده قرار می گیرد،علم امروز نسبت به دیروز جدیدتر است.روشهای جدید علمی در پی کشف محدودیتهای روشهای قدیمی ایجاد می شود،و از آنجایی که روشهای آماری جزء روشهای قدیمیData mining محسوب می شوند،از این قاعده کلی که دارای محدودیت هستند مستثنی نیستند.داشتن فرض اولیه در مورد داده ها، یکی از این موارد است.در اینجا به تشریح بیشتر تفاوتهای بین مباحث و متدهای آماری و دیگر متدهای داده کاوی که در کتابهای مختلف بحث شده است می پردازیم.

تکنیکهای داده کاوی و تکنیکهای آماری در مباحثی چون تعریف مقدار هدف برای پیش گویی،ارزشیابی خوب و داده های دقیق (تمیز)(clean data) خوب عمل می کنند،همچنین این موارد در جاهای یکسان برای انواع یکسا نی از مسایل (پیش گویی ،کلاس بندی و کشف)استفاده می شوند، بنابراین تفاوت این دو چیست؟چرا مآنچنان که علاقه مند بکاربردن روشهای دا ده کاوی هستیم علاقه مند

روشهای آماری نیستیم؟ برای جواب این سوال چندین دلیل وجود دارد اول اینکه روشها ی کلاسیک داده کاوی از قبیل شبکه های عصبی،تکنیک نزدیک ترین همسایه روشهای قوی تری برای داده های واقعی به ما می دهند و همچنین استفاده از آنها برای کاربرانی که تجربه کمتری دارند راحت تر است و بهتر می توانند از آن استفاده کنند.دلیل دیگر اینکه بخاطر اینکه معمولاً داده ها اطلاعات زیادی در اختیار ما نمی گذارند،این روشها با اطلاعات کمتر بهتر می توانند کار کنند وهمچنین اینکه برای داده ها وسیع کابرد دارند.[۳]

در جایی دیگر اینگونه بیان شده که داده ها ی جمع آوری شده نوعاً خیلی از فرضهای قدیمی آماری را در نظر نمی گیرند،از قبیل اینکه مشخصه ها باید مستقل باشند،تعیین توزیع داده ها،داشتن کمترین همپوشانی در فضا و زمان اغلب داده ها هم پوشانی زیاد می دارند،تخلف کردن از هرکدام از فرضها می توان مشکلات بزرگی ایجاد کند،زمانی که یک کاربر(تصمیم گیرنده) سعی می کند که نتیجه ای را بدست آورد. داده های جمع آوری شده بطورکلی تنها مجموعه ای از مشاهدات چندی بعد است بدون توجه به اینکه چگونه جمع آوری شده اند[۵].

در جایی پایه و اساس Data mining به دو مقوله آمار و هوش مصنوعی تقسیم شده است که روشهای مصنوعی به عنوان روشهای یادگیری ماشین در نظر گرفته می شوند.فرق اساسی بین روشهای آماری و روشهای یادگیری ماشین (machine learning) بر اساس فرضها و یا طبیعت داده هایی که پردازش می شوند.بعنوان یک قانون کلی فرضها تکنیکهای آماری بر این اساس است که توزیع داده ها مشخص است که بیشتر موارد فرض بر این است که توزیع نرمال است و در نهایت درستی یا نادرستی نتایج نهایی به درست بودن فرض اولیه وابسته است.در مقابل روشهای یادگیری یادگیری ماشین از هیچ فرض در مورد داده ها استفاده نمی کند و همین مورد باعث تفاوتهایی بین این دو روش می شود.

به هر حال ذکر این نکته ضروری به نظر می رسد که بسیاری از روشهای یادگیری ماشین برای ساخت مدل dataset از حداقل چند استنتاج آماری استفاده می کنندکه این مساله بطور خاص در شبکه عصبی دیده می شود.[۱]

بطور کلی روشهای آماری روش های قدیمی تری هستند که به حالت های احتمالی مربوط می شوند.Data mining جایگاه جدید تری دارد که به هوش مصنوعی یادگیری ماشین سیستمهای اطلاعات مدیریت (MIS) و متدلوژی Database مربوط می شود.

روشهای آماری بیشتر زمانی که تعداد دادهها کمتر است و اطلاعات بیشتری در مورد داده ها می توان بدست آورد استفاده می شوند به عبارت دیگر این روشها با مجموعه داده ها ی کوچک تر سر و کار دارند همچنین به کاربران ابزارهای بیشتری برای امتحان کردن داده ها با دقت بیشتر فهمیدن ارتباطات بین داده ها می دهد. بر خلاف روشهایی از قبیل شبکه عصبی که فرآیند مبهمی دارد. پس به طور کلی این روش در محدوده مشخصی از داده های ورودی بکار می رود.بکار بردن این روشها مجموعه داده های مجموعه داده های زیاد احتمال خطا در این روشها را زیاد می کند.چون در داده ها احتمال noise وخطا بیشتر می شود و نیز روشهای آماری معمولابه حذف noise می پردازند، بنابراین خطای محاسبات در این حالت زیاد می شود.[۷]،[۸]

در بعضی از روشهای آماری نیازداریم که توزیع داده ها را بدانیم. اگر بتوان به آن دسترسی پیدا کرده با بکار بردن روش آماری می توان به نتایج خوبی رسید.

روشهای آماری چون پایه ریاضی دارندنتایج دقیق تری نسبت به دیگر روشها ی Data miningاریه می دهند ولی استفاده از روابط ریاضی نیازمند داشتن اطلا عات بیشتری در مورد داده ها است.

مزیت دیگر روشهای آماری در تعبیر و تفسیر داده ها است.هر چند روشهای آماری به خاطر داشتن ساختار ریاضی تفسیر سخت تری دارند ولی دقت نتیجه گیری و تعبیر خروجی ها در این روش بهتر است بطور کلی روشهای آماری زمانی که تفسیر داده ها توسط روشهای دیگر مشکل است بسیار مفید هستند.

تفاوتهای کلی روشهای آماری و دیگر روشهای Data mining در جدول اریه شده است:

دیگر روشهای Data mining	روشهای آماری
بدون فرض اولیه	داشتن فرض اولیه

در انواع مختلفی از داده ها کاربرد دارند نه فقط داده های عددی	تنها برای داده های عددی کاربرد دارند
در محدوده وسیع تری از داده ها	در محدوده کوچکی از داده ها
Data mining به دادهای درست clean data بستگي دارند	حذفnoise ها ، داده های نامشخص ووفیلتر کردن dirty data
استفاده از شبکه عصبي	روشهای رگرسیون و استفاده از معادلات
استفاده ازData visulization	استفاده از چارتهای دو بعدی و سه بعدی
استفاده از روشهای یادگیری ماشین و هوش مصنوعی	استفاده از روابط ریاضی
در یادگیری غیر نظارتی کاربرد بیشتر دارد	در descriptive statistical و cluster analysis کاربرد دارد.

همچنین می توان گفت که درDM داده ها اغلب بر اساس همپوشانی نمونه هاست،نسبت به اینکه بر اساس احتمال داده ها باشد.همپوشانی نمونه ها برای آشنایی همه انواع پایه ها برای تخمین پا را مترها مشهور است. وهمچنین اغلب استنتاج های آماری نتایج ممکن است مشارکتی باشد تا اینکه سببی باشند. تکنیکهای ماشین را به سادگی می توان تفسیر کرد .مثلاً روش شبکه عصبی بر اساس یک مدل ساده بر اساس مغز انسان استوار است.یعنی همان ساختار مغز انسان را اجرا می کنند ولی خروجی های بسیاری از روشهای آماری ساختار ریاضی دارند،مثلاً یک معادله است که تعبیر و تفسیر آن مشکل تر است.در مورد روش های آماری بایداین مطلب را گفت بدون توجه به اینکه مدل کاربردی،مدل آماری است یا خیر،تستهای آماری می تواند برای تحلیل نتایج مفید باشد.

با ارایه توضیحات داده شده درباره های تفاوتهای روشهای آماری و دیگر روشهای DM در ادامه به کابردهای روش روشهای آماری و بحثهای مشترک آما رو DM می پردازیم .

كاربردهاي روشهاي آماري:

Data mining معمولا وظايف يا به عبارت بهتر استراتژهاي زير را در داده ها بكار مي برد:

- توضيح و تفسير (description)
- تخمين (estimation)
- پيش بيني (prediction)
- كلاس بندي (classification)
- خوشه سازي (clustering)
- وابسته سازي وايجاد رابطه (association)

در جدول زير استراتژي ها و روشهاي هر استراتژي مشخص شده است:

روشها	استراتژيها
تحليل داده ها	توضيح وتفسير
تحليل هاي آماري	تخمين
تحليل هاي آماري	پيش بيني
الگوريتم نزديك ترين همسايه	كلاس بندي
درخت تصميم	كلاس بندي
شبكه هاي عصبي	كلاس بندي
خوشه سازي k-mean	خوشه سازي
شبكه هاي kohonen	خوشه سازي
وابسته سازي و ايجاد رابطه	رابطه سازي

البته بايد گفت كه روشهاي data mining تنها به يك استراتژي خاص محدود نمي شوندو نتايج يك را همپوشاني بين روشها نشان مي دهد. براي مثال درخت تصميم ممكن است كه دركلاس بندي

تخمين وپيش بيني كاربرد داشته باشد. بنابراين اين جدول را نبايد به عنوان تعريف تعريف تقسيم بندي از وظايف در نظرگرفته شود بلكه به عنوان يك خروجي از آنچه كه ما به عنوان وظايف dataminig آشنايي پيدا كرديم در نظر گرفته مي شود.

همانگونه كه ازجدول پيداست روشهاي آماري در مباحث تخمين وپيش بيني كاربرد دارند. در تحليل آماري تخمين وپيش بيني عناصري از استنباطهاي آماري هستند.استنباطهاي آماري شامل روشهايي براي تخمين وتست فرضيات درباره جمعيتي از ويژگيها براساس اطلاعات حاصل از نمونه است .يك جمعيت شامل مجموعه اي از عناصر از قبيل افراد ايتم ياداده ها يي كه دريك مطالعه خاص آمده است. بنابراين در اينجا به توضيح اين دواستراتژي مي پردازيم.

۱- **تخمين**:

در تخمين به دنبال اين هستيم كه مقدار يک مشخصه خروجی مجهول را تعيين كنيم،مشخصه خروجی در مسا يل تخمين بيشتر عددی هستند تا قياسی [۱] .بنابراين مواردی كه بصورت قياسی هستند بايد به حالت عددی تبديل شوند.مثلا موارد بلی،خير به 0 و1 تبديل می شود.

تكنيكهاي نظارتيDM قادرند يكی از دو نوع مسايل كلاس بندی يا تخمين را حل كنند، نه اينكه هر دو را.يعنی اينكه تكنيكی كه كار تخمين را ا نجام می دهد، كلا س بندی نمی كند.

روشهاي آماري مورد استفاده دراين مورد بطوركلي شامل تخمين نقطه و فا صله اطمينان ميباشد. تحليل هاي آماري تخمين وتحليل هاي يك متغيره و...از اين جمله مي باشند.

در توضيح اينكه چرا به سراغ تخمين مي رويم بايد گفت كه مقدار واقعي پارامترها براي ما ناشناخته است.مثلا مقدار واقعي ميانگين يك جامعه مشخص نيست.داده ها ممكن است كه بطور رضايت بخشي جمع آوري نشده باشد يا به عبارتي warehouse نشده باشد. به همين دليل تحليل گران از تخمين استفاده ميكنند.

در خيلي از موارد تعيين ميانگين مجموعه اي از داده ها براي ما مهم است.مثلا ميانگين نمرات درسي يك كلاس،ميانگين تعداد نفراتي كه در يك روز به بانك مراجعه مي كنند،متوسط مقدار پولي كه افراد دريك شعبه خاص از بانك واريز مي كنند وموارد اينچنيني.

زمانی که مقدار یک آماره را برای براورد کردن پارامتر یک جامعه به کار ببریم،آن پارامتر را تخمین زده ایم،وبه مقدار این آماره برآورد نقطه ای پرامتر اطلاق می کنیم.در واقع از کلمه نقطه برای تمایز بین براورد کننده های نقطه ای و فاصله ای استفاده می کنیم.از مهمترین تخمین زننده ها است که به ترتیب برآورد واریانس و میانگین جامعه هستند.خود برآورد کننده ها دارای خاصیت هایی چون ناریبی، کارایی،ناسازگاری،بسندگی و... هستند،که هر یک به بیان ویژگی خاصی از آنها می پردازندو میزان توانایی آنها را در تخمین درست و دقیق یک پارامترتعیین می کنند.

در تخمین نیازمند داشتن اندازه نمونه هستیم،در تعیین اندازه نمونه می توان از رابطه زیر استفاده کرد:

که p احتمال رخدادو e درصد خطای پذیرفته شده است که در اینجا 5% در نظر می گیریم.

پر کاربرد ترین تخمین زننده ،تخمین زننده میانگین جامعه است،ساده ترین رابطه اي که براي میانگین داده ها مي توان نوشت بدین صورت است:

که n تعداد نمونه ها و مقدارهر نمونه است.در اینجا تمام نمونه ها ارزش یکساني دارند ولي گاهي اوقات نیاز است که نمونه ها بر اساس اهمیتي که دارند وزن دهي شوند.

بدین صورت :

که ها در اینجا وؤن هر نمونه i ام هستند. در این حالت برای تعیین مجموع اوزان نمونه ها به جای n، می باشد.

در مواردي نیز تخمین فاصله براي ما اهمیت دارد.فاصله اطمینان شامل فاصله اي است که با درصدي از اطمینان مي توانیم بگوییم که مقدار یک پارامتر درون این این فاصله قرار مي گیرد.به عبارت دیگراگر چه برآورد نقطه ای طریقه متداول توصیف .برآورد هاست اما درباره آن، جا برای پرسشهای زیادی باقی است.مثلا برآورد نقطه ای به ما نمی گوید که برآورد بر چه مقداری از اطلاعات مبتنی است.و چیزی درباره خطا بیان نمی کند. بنابراین می توانیم که برآورد پارامتر را با بعلاوه کردن اندازه کردن اندازه نمونه و مقدار واریانس ،یا اطلاعات دیگری درباره توزیع نمونه گیری کامل کنیم.این کار ما را قادر می سازد که اندازه ممکن خطا را برآورد کنیم.

یک براورد فاصله ای ،فاصله ای به شکل است که درآن و مقادیر متغییرهای تصادفی مناسبی برای هستند،منظور از مناسب آن است که به ازای احتمال مشخصی مانند داریم:

برای مقدار مشخص ، را یک فاصله اطمینان برای می نامیم. همچنین ، درجه اطمینان ،و دو سر فاصله کرانهای اطمینان پایینی و بالایی نامیده می شود.مثلا برای ، درجه اطمینان 95%است ویک فاصله اطمینان 95% بدست می آوریم.فاصله اطمینان از اکثر توزیع ها ،همانند توزیع نرمال ،خی دو،t استودنت وتوزیع F و... استفاده می کند.مثلا اگر مقدار میانگین یک نمونه تصادفی به اندازه n از جامعه نرمال ووایانس معلوم باشد آنگاه

یک فاصله اطمینان برای میانگین جامعه است.[٦]

در خيلي از موارد تعيين نقطه دقيق يک پارامتر ممکن نيست،ولي فاصله اطمينان ،اطمينان ما را از قرار گرفتن مقدار پارامتر در يک بازه تضمين مي کند.فاصله اطمينان را مي توان براي اکثر توزيع ها از جمله توزيع خي دو،توزيع t استودنت وتوزيع F و ...بدست آورد.

۲- پيش بيني(prediction) :

هدف از انجام پیش بینی تعیین ترکیب خروجی با استفاده از رفتار موجود می باشد. یعنی در واقع رسیدن به یک نتیجه بوسیله اطلاعات موجود از داده ها. مشخصه های خروجی در این روش هم می توانند عددی باشند وهم قیاسی.[۱] این استراتژی در بین استراتژی های data mining از اهمییت خاصی بر خوردار است، و مفهوم کلی تری را نسبت به موارد دیگر دارد.خیلی از تکنیکهای نظارتی data mining که براي کلاس بندي و تخمين مناسب هستند در واقع کار پيش بيني انجام مي دهند.

آنچه از کتابهاي آماري وdata minig تحت عنوان پيش بيني برمي آيد رگرسيون و مباحث مر بوط به آن است .در واقع در اکثر اين کتابها هدف اصلي از انجام تحليل هاي آماري براي داده کاوي، رگرسيون داده هاست واين بعنوان وظيفه اصلي متد هاي آماري معرفي مي شود.

اهداف تحليل رگرسيون:

با انجام رگرسيون مي خواهيم اهداف زيررا دنبال کنيم:

۱- بدست آوردن رفتار متغییرy توسط متغییر x ،یعني اینکه متغیرy با تغییرx در نمونه ها چه رفتاري را از خود نشان مي دهد.مثلا در نمونه اي این رفتار خطي است یا اینکه شکل منحني خواهد داشت.

۲- پیش بیني بر اساس دادهها براي نمونه هاي آینده،که هدف اصلي در داده کاوي از طریق متدهاي آماري است.مثلا از روي اطلاعاتي مثل داشتن کارت اعتباري یک فرد جدید،نوع جنسیت او،سن فرد،میزان درآمد سالیانه او بتوان حدس زد که این فرد از بیمه عمر استفاده مي کند یا خیر. ویا اینکه با داشتن اطلاعات در مورد داشتن یا نداشتن کارت اعتباري و بیمه عمر، سن فردبتوان جنسیت فرد را تعیین کرد.

۳- استنباط استنتاجي یا تحلیل حساسیت، تعیین اینکه اگرx به اندازه خاصي تعییر کندy تا چه اندازه تغییر خواهد کرد.هدف از فهمیدن اینکه چگونه تغییرات y تابعي ازx است. باید توجه داشت که نوع تغییرات مدل گرسیوني خاصي را مي دهد.

اهداف مدلسازي براي تشریح ارتباط بین x وy استفاده از نتایج مدل براي پیش بیني کاربردهاي تخمین عبارت است.اما استنباط استنتاجي یک مقوله ظریف تري است.زماني که به استنباط آماري فکر میکنیم در واقع درباره متغییر رفتاري و متغییر هاي کنترل فکر مي کنیم.

متغییرهاي رفتاري مشخصه هایي را ارایه میکنندکه تبحروتجربه خاصي دارندیا اینکه قابلیت آن نبحر را دارند.مثلا مقدار دز دارو که براي بیمار استفاده مي شود در تجربه پزشکي .همچنین متغییرهاي کنترل دیگر ویژگي ها در یک محیط آزمایشي را اندازه میگیرند،از قبیل وزن بیمار که قبل از رفتار اندازه گیري مي شود.

اگر ما براي یکي از متغییر هاي رفتاري، کنترل انجام دهیم، رگرسیون ما احتمالا استنباط هاي استنتاجي را درست حدس میزند.و اگر ما علاقه مند به هر دو مورد پیش بیني انتخاب سهم وتخمین اثرات علتها باشیم تایید هر دومورد را بعنوان متغییرهاي خروجي که همپوشاني دارند در نظر مي گیریم.

روشهاي مختلف رگرسيون براي داده كاوي وجود دارد .رگرسيون خطي بيشترين كاربردرا دارد وهمچنين مشتقات آن حايزاهمييت است.يک نمونه از آن مشتقات آن رگرسيون خطي سلسله مراتبي يا رگرسيون چند سطحي است. اين روش يکي از ابزارهاي تحليل دادههاي پيچيده از قبيل افزايش فر كانس در تحقيقات مقداري را شامل مي شود.مدلهاي رگرسيون چند سطحي براي حالتهايي كه همپوشاني در سطوح مختلف وجود دارد مفيد است. براي مثال اطلاعات آموزشي ممكن است اطلاعاتي از قبيل اطلاعات فردي دانش آموزان (نام، نام خانوادگي و در كل پيش زمينه خانوادگي)،اطلاعات سطح كلاس از قبيل ويژگي هاي معلم وهمچينين اطلاعات درباره مدرسه همانند سياست آموزشي و... باشد. حالت ديگر مد لهاي چند سطحي ،تحليل دادههاي بدست آمده از نمونه هاي خوشه بندي شده است. يک خانواده از مدلهاي رگرسيون، به عنوان متغييرهاي شاخص بري رتبه بندي يا خوشه بندي است علاوه بر اينكه همپوشاني را اندازه مي گيرد. با نمونه خوشه بندي شده مدلسازي چند سطحي براي توسعه نمونه هايي كه داخل خوشه نيستند،لازم است.[٤]

در روش رگرسيون چند سطحي يا سلسله مراتبي محدوديتي براي تعداد سطوح تغيير كه مي تواند انجام شود،وجود نداردروشهاي بيزي در تخمين پارامترهاي مجهول كمک مي كند،هرچند كه محاسبات پيچيده اي دارد.ساده ترين توسعه از رگرسيون همپوشاني مجموعه اي از متغيرهاي شاخص براي كلاس بندي نمونه هاي آموزشي يا رتبه بندي وخوشه بندي درنمونه هاي داده شده است.همچنين به عنوان توسعه رگرسيون خطي در نظر گرفته مي شود،كه در ادامه به توضيح آن مي پردازيم[٤]

١- رگرسيون خطی(Linear regression)

يکی از هدفهای اصلی بسياری از پژوهشهای آماری ا يجاد وابستگی هايی است تا پيش بينی يک يا چند متغير را بر حسب سايرين ممکن می سازد.مثلاً مطالعاتی انجام می شودتا فروشهای بالقوهَ يک محصول جديد را بر حسب قيمت آن،وزن يک بيماررا بر حسب تعداد هفته هايی که پرهيز داشته است،پيش بينی کند.

در عمل مسايل متعددی وجود دارند که در آن ها مجموعه ای از داده ها زوج شده بر آن دلالت می کند که رگرسيون خطی است و در آن توزيع توأم متغيرهای تصادفی تحت بررسی رانمی دانيم اما با اين حال می خواهيم که ضرايب رگرسيون را برآ ورد کنيم.

روش رگرسیون خطی یک تکنیک یادگیری نظارتی است که به وسیله آ ن می خواهیم تغییرات یک متغیر وابسته بوسیلهَ ترکیب خطی از یک یا چند متغیرمستقل مدل کنیم .حالت کلی معادله آن به این صورت است :

$$f(x_1+x_2+\ldots+x_n)=a_1x_1+a_2 \qquad (1)$$

$$x_2+\ldots+a_n\, x_n+b$$

کهxها متغیر مستقل و aهاو b ضرایب ثابت هستند و$f(X_1.X_1\ldots X_n)$ متغیر وابسته می باشند.حالت ساده این معادله بصورت $y=ax+b$ (۲) است که در اینجا y متغیر وابسته است به حالت ساده شده معادله ۱(یعنی معادله۲)shope-intercept from می گویند.

یک روش برای تعیین ضرایب a,b روش حداقل مربعات است.ملاک کمترین مربعات این است که مجموع مربعات ا نحراف ها را مینیمم کنیم؛بنابراین اگر مجموعهای از داده های زوج شده مانند $\{(x_i,y_i),i=1,2,\ldots,n\}$ داده شده باشد، برآ وردهای کم ترین مر بعات ضرایب رگرسیون، مقادیری مانندa,bهستند که به ا زای آنها کمیت

مینیمم است؛

e در شکل مشخص شده است:

e_i

y_i

ax_i+b

بنابر این در حالت ساده اگر یک نمونه n تایی داشته داشته باشیم مقادیر a,bرا از طریق روابط زیر برآورد می کنیم :

مزیت رگرسیون خطی این است که فهمیدن و کار با آن ساده است در حالت کلی برای استراتژی و پیش بینی مناسب است. با بکار بردن این روش از نتایج خروجی می توان دریافت که این روش مناسب بوده یا خیر . بنابر این معیارهایی داریم که با استفاده از آنها می توان دریافت که آیا می توان به نتایج خروجی اطمینان کرد یا خیر.

آنچه در انجام رگرسیون مهم به نظر می رسد،تعیین میزان همبسته بودن داده ها به یکدیگر است.با مشخص کردن میزان همبسته بودن داده های متغیرهای ورودی و خروجی می توان دریافت که رگرسیون خطی برای انجام داده کاوی مناسب است یا خیر، بنابراین ضریب همبستگی و برآوردهای آن در بسیاری از پژوهشهای آماری اهمیت دارند. شرایطی که وقتی چند متغیر پیشگو(x_i) با یکدیگر هم پوشانی دارند،این هم پوشانی منجر نااستواری و تزلزل در فضای جواب می شود،همچنین منجر به نتایج بی ارتباط(بی ربط) می شود.حتی اگر از این تزلزل اجتناب شود هم پوشانی بین متغیرهایی که میزان بین متغیرهایی همبستگی آنها زیاد است ،منجر به تاکید کردن روی بخش خاصی از مدل می شود.[۲]

بنابر این از بین متغیر های ورودی مواردی که با هم بستگی زیادی دارند،نباید با هم در تعیین ارزش متغیر خروجی بکار بره شوند.واز طرفی کاربردرگرسیون خطی منوط به همبستگی متغیر های ورودی و خروجی است.در تحلیل همبستگی نرمال مربوط به داده های زوج شده ،با استفاده از روابط ریاضی می توان به ،که بیانگر ضریب همبستگی نمونه ای ا ست رسید. رابطه چنین است :

که به ترتیب میانگین متغیر های ورودی و خروجی هستند. را معمولاًبا r نمایش می دهند و رابطه ساده شده آن به این صورت است:

که: وشدت بستگی بین x,y را اندازه مي گيرد

در صورتی که r=0 باشد،این دو متغیر (x,y) نسبت به هم نا همبسته اند ،و هر چه صفر فا صله می گیرد، بطرف ۱+و ۱- میزان همبسته بودن آنها زیادتر می شود،و ۱+همبستگی خطی مثبت و ۱- همبستگی خطی منفی را نشان می دهد.در حالت توزیع نرمال دو متغیر، صفر بوده r (r=0) مستقل بودن این دو متغیر را از هم نشان می دهد. رابطه روبرو را در نظر بگیرید:

وقتی که باشد،نتیجه می شودکه و این بدان معنی است که همبستگی خطی کاملی بین x وy موجود است.برای تفسیر مقادیرr بین 0 و1+ یا 0 و1- ، این معادله رانسبت به حل کرده نتیجه را در ۱۰۰ ضرب می کنیم بنابراین داریم:

که در آ ن تغییر کلی yها و تغییر شرطی yها را به ازای مقادیر ثابت xاندازه می گیرند..بنابراین آن قسمت از کل yها که در اثر بستگی به x قابل توضیح است اندازه می گیرد .پس $100r^2$درصد تغییر کلی ا زyها ستکه در اثر بستگی به x قابل توضیح است.

مثلاً وقتی r=%5 در این صورت ۲۵% از تغییر yهاست که در اثر بستگی بهx قابل توضیح است.و وقتی r=%7 در این صورت ٤۹ درصد از تغییر yها در اثر بستگی به x قابل توضیح است.بنابراین می توانیم بگوییم که یک همبستگی r=%7 تقریباً دو برابر قوی تر یک همبستگیr=%5 است.[٦]

همچنین تحلیل رگرسیونی نرمال برای حالت چند گانه رابطه مفید زیر را که بر اساس توزیع t بدست می دهد، ا رایه می کند:

که در این رابطه ضریب متغیر x_i و مقدار عدد ثابت در معادله رگرسیون است.

n تعداد نمونه ها،k تعداد متغیر های ورودی(x_iها)است.

همچنین در نظر بگیرید که یک سری داده متشکل از k متغیر ورودی و یک متغیرخروجی که تعداد هر نمونه از متغیر n تا باشد،با ضرایب و $i=0,1,\ldots,k$ بصورت ماتریس به شکل زیر نمایش داده شود:

همچنین را ترانهاده و را نهادهٔ و)=B در نظر بگیرید.بدین ترتیب و c_{ii} درایه ماتریس معکوس x یعنی است.باید درایه c_{22} در این ماتریس را حساب کنیم.

عبارتt دارای توزیع t با n-k-1 درجه آزادی است.که یک آمارهٔ مناسب برای آزمون میزان تاثیری که ضریب هر یعنی در معادله رگرسیونی دارد.[٦]

۲- Logistic Regression

این روش یکی از تکنیکهای یادگیری نظارتی و در حالتی که نتایج خروجی به صورت binaryهستند،مورد توجه قرار می گیرد.در کل زمانی نتایج خروجی به صورت binary هستند رگرسیون خطی خیلی کارا نیست،در این حالت استفاده از این تکنیک مناسب تر است.نکته دیگر اینکه این روش یک تکنیک رگرسیون غیر خطی است و لزومی ندارد که داده ها حالت خطی داشته باشند.ا گر بخواهیم دلیل استفادهLogistic regression را بیان کنیم باید اینگونه بحث کنیم در رگرسیون خطی علاوه بر اینکه نتایج خروجی باید به صورت عددی باشد،متغیر ها هم باید به صورت عددی باشد بنابراین حالتهایی که به صورت کتگوری (قیاسی) هستند باید به حالت عددی تغییر شکل پیدا کنند.مثلاً جنسیت افراد از حالت زن و مرد بوده به ترتیب به حالتهای 0 و1تغییر پیدا می کند.در این روش اگر نتایج خروجی(متغیر خروجی) بصورت binary باشد می تواند مفید باشد. چون اساس رگرسیون خطی در این حالت ایراد پیدا می کند و ارزش قیدی که بر روی متغیر وابسته قرار می گیرد توسط معادله رگرسیون در نظر گرفته نمی شود.

در وا قع چون رگرسیون خطی،معاد له یک خط را ترسیم می کند،نمی تواند حالت مثبت و منفی یا به عبارتی صفر و یک را در نظر بگیرد.

به همین دلیل برای اینکه بتوان حالتهای binary را هم در نظر گرفت،باید شکل معادله را تغییر داد.با این تغییر شکل معادله رگرسیون احتمال اتفاق افتادن یا اتفاق نیفتادن یک واقعه را بدست می دهد.

با تغییر شکل رگرسیون خطی به حالت Logistic regression این مشکل حل می شود.

معادله خطی را می توان بدین صورت نوشت:

که بیانگر احتمال اینکه متغیروا بسته (y) مقدار 1 را بگیرد به شرط اینکه ترکیبی ا زx را داشته باشیم.بصورت کلی تر و برای حذف محدودیت های مسأله ها حالت احتمالی y=1 را نسبت به y=0 در نظر می گیریم یعنی به صورت

ولگاریتم طبیعی این عبارت را برابر قرار می دهیم که x برداری بصورت است ،ودر نهایت از رابطهln بالا مقدا ر بدست می آید که برابر است با:

این معادله ،معادله Logistic regression را تشکیل می دهد.

اگر بخواهیم منحنی این معادله را نشان دهیم بصورت روبرو می باشد:

که بیانگرغیر خطی بودن این معادله است.در نهایت برای فهم بهتر مسأ له مثا لی ارائه می شود.

داده های زیر را وارد Excel کرده و ضرایب متغیر ها و مقدار ثابت b را توسط تابع LINEST بدست می آ وریم داده ها و نتایج به این صورت می باشد؛

instance	income	Credit card insurance	sex	age	Life insurance promotion	Computed probability
1	40	0	1	45	0	0.007
2	30	0	0	40	1	0.987
3	40	0	1	42	0	0.024
4	30	1	1	43	1	1.000
5	50	0	0	38	1	0.999
6	20	0	0	55	0	0.049
7	30	1	1	35	1	1.000
8	20	0	1	27	0	0.584

9	30	0	1	43	0	0.005
10	30	0	0	41	1	0.981
11	40	0	0	43	1	0.985
12	20	0	1	29	1	0.380
13	50	1	0	39	1	0.999

این مثال ٤ مشخصه ورودی و یک مشخصه خروجی دارد که ضرایب متغیرهای ورودی در زیر محاسبه شده است:

ax+b= 0.0001income+19.827credit card ins-8.314sex+0.415age+17.691

با این معادله می توان نتایج Life Insuranee promotion بدست آ ورد ، که همانطور که در جدول فوق نشان داده شده نتایج ماسبه شده با متغیر وا بسته هم خوا نی زیادی دارد.حال اگر نمونه جدیدی به این صورت داشته باشیم:

In cone=35k credit card Insuranee=1 sex=0 age=39

با محاسبات احتمال بدست آمده برابر 0.999می باشد.که این فرد یک کاندیدا را برای بیمه عمر (Life (Insuranee promotion می باشد حا لت دیگر اینکه نمونه جدید به صورت :

Ineome=35k credit card Insuran =0 sex=1 age=39

باشد در این حالت مقدار احتما لی بدست آمده برابر0.035است که نشان می دهد یک مرد ۳۹ ساله که در آمد سالیانه او ۳۵۰۰۰است و بیمه کارت اعتباری ندارد یک نمونه ضعیف برای داشتن بیمه عمرا ست.

۳- Bayse classsifire

این مقدار یکی از روشهای ساده یادگیری نظارتی است،که در آن فرض می شود که تمام متغیرها ی ورودی به یک اندازه مهم هستند و مستقل از هم می باشند و نیز ا گریکی از شرایط هم برقرار نباشد این روش در شرایطی کاربرد دارد این روش بر اساس تئوری بیز بنا شده است.که این تئوری به صورت زیر است:

که در این جا H متغیر وابسته است و E بوسیله مقدار ویژگی های ورودی تعیین می شود.

Bayse classifier برخلاف اکثر روشهای آماری برای حالتی که مقدار دادهَ یک متغیر ورودی نامعلوم است نیز کاربرد دارد.در ادا مه با ا رائه یک مثال می توان به توضیح این روش پرداخت.

دراینجا نیز ابتدا متغیر خروجی را تعیین می کنیم.فرض کنید که یک سری داده داریم، و با استفاده از آن داده ها می خواهیم برای یک نمونه جدید به شکل زیر،جنسیت فرد را تعیین کنیم:

Magazine promotnio=Yes watch promotion=Yes

Life Insurance Promotion=No credit card Insurance=No

Sex=?

اگر بخواهیم این نمونه را با فرمولBayse classifier بنویسیم داریم:

که برابر است با:

و همچنین باید جنسیت زن نیز محاسبه شود بعنی عبارت زیر

همانند روش فوق باید را محاسبه کرد.

با محاسبه این احتمال داریم:

و چون ۲۸۱%<۵۹۳%بنابر این احتمال اینکه جنسیت فرد در نمونه جدید مرد باشد بیشتر است. پس احتمال اینکه یک نمونه با این مشخصات مذکر باشددو برابر مونث بودن آن است.

مطلب دیکر اینکه زمانی که مقدار یک احتمال صفر باشد چون احتمال ها در هم ضرب می شوند کل احتمال صفر خواهد شد مثلاً وقتی که باشد مقدار احتمال خواهد شد،Bayes classifier برای رفع این مشکل به یک مقدار k به صورت کسر ضرب در یک احتمالp و یک مقدارk به مخرج اضافه می کند.بدین صورت :

می باشد که k مقداری بین صفر و یک دارد که معمولاً مقدار یک می گیرد و نیزp بستگی به تعداد انتخابهای متغیر خروجی دارد مثلاً اگر متغیر خروجی دو حالتی باشد(yes,No)،مقدارp برابر0.5 می باشد. نیز همان مقدار های یا است.مثلاً اگر مقدار

باشد آنگاه برابر خواهد بود.

علاوه برا ین روش (Bayse chassifire) برای حالت Missing data نیز کاربرد دارد.یعنی اگر مقدار یکی از مشخصه های ورودی در یک نمونه جدید را نداشته باشیم،در این روش این مشخصه را کلاً حذف می شود.

حالت دیگری که می توان این روش را بکار برد وجود مشخصه هایی با داده های عددی در بین مشخصه های ورودی است.مثال زیر که یک نمونه جدید است در نظر بگیرید:

Magazine promotion=Yes , wateh promotion=Yes

Life insurance promotion=No , credite card Insurance=No, Age=45

یعنی همان نمونه قبلی با این تغییر که سن نیز به مشخصه های ورودی اضافه شده.در اینجا با استفاده از این روش ابتدا باید باید توزیع مشخصه ورودی تعیین کنیم،که معمولاً فرض می شود که مشخصه از توزیع نرمال پیروی می کند.مثلاً مورد روبرو را می خواهیم حساب کنیم:

که برابر است با:

(عبارت از مثال قبلی که age جز متغیرهای ورودی نبود، بدست آمده)

در اینجا ابتدا باید را حساب کنیم که با بدست آوردن میانگین و واریانس داده های سن داریم:

با قرار دادن در فرمول توزیع نرمال داریم:

که این مقدار برابر با0.03 است و به همین ترتیب برای حالت نیز حساب می کنیم.

نتیجه لازم را از داده های خروجی می گیریم.با حساب کردن احتمال قائده بیز داریم:

که در اینجا نیز احتمال مرد بودن بیشتراست.

ابزار رگرسیون خطی : (توضیح تابع LINEST)

برای اجرای رگرسیون خطی می توان از نرم افزار Excelاستفاده کرد.در قسمت توابع Excle،تابع LINEST برای اجرای یک رگرسیون خطی ایجاد شده است.در اینجا به چگونگی کار با این تابع و استفاده از نتایج بدست آمده اجرای آن می پردازیم.

۱- روش کار با نرم افزار: بعد از باز کردن Excel داده هایی را که می خواهیم بوسیلهَ آنها رگرسیون خطی را اجرا کنیم،وارد می کنیم و نیز محلی را که می خواهیم داده ها خروجی نشان داده شوندتعیین می کنیم سپس از منوی Insert ،furction را انتخاب می کنیم.در قسمت select a cotegory مقولهَ

statistical را انتخاب می کنیم.با انجام این کار تابع های آماری در قسمت پایین همین پنجره نمایش داده می شود.از بین تابع ها تابع LINEST را انتخاب می کنیم و OK می کنیم.

در پنجره LINEST چهار قسمت وجود دارد که باید پر شوند. در قسمت اول باید ستون متغیر وابسته (y) را مشخص می کنیم.مثلاً اگر دادهای شما در ستونEاز ردیف ۲تا۱۲هستند،در این قسمت می نویسیم E2:E12، درقسمت دوم ستون متغیرهای مستقل ها را به همین ترتیب مشخص می کنیم.قسمت سوم مقدار عدد ثابت رگرسیون را به ما می دهد،اگر این مقدار True انتخاب کنیم ،مقدار عدد ثابت را بر میگرداندواگر False باشدمقدار عدد ثابت صفراست. درقسمت چهارم(state) اگر عبارتTure تایپ می شوداطلاعاتی را در مورد نتایج رگرسیون بدست می دهد که مفید است،مثلاً ضریب همبستگی،بین مقدار تخمینی و مقدار واقعی متغیر وابسته،مقدار آماره F و...که با مثال بیشتر توضیح داده می شودواگر این مقدار False باشدExcel این نتایج را ارائه نمی دهد.پس از آنکه هر چهار قسمت پر شد،با نگه داشتن ctrl+shiftو زدنinter(یاokکردن) نتایج اجرای رگرسیون دیده می شود.برای فهم بهتر این ابزار در زیر مثالی آورده می شود.

مثالی را در نظر بگیرید که ٤ متغیر مستقل و یک متغیر وابسته داده ها مربوط به ساختمان اداری می باشد که با داده های ورودی که می گیریم،قصد داریم،ارزش یک ساختمان را به واحد پولی دلار تخمین بزنیم. متغیرهای تا و به شرح زیر می باشند:

توضیح	متغیر
قیمت ساختمان	Y
میزان فضای ساختمان	X_1
تعداد اتاقهای ساختمان	X_2
تعداد ورودی ها	X_3
میزان عمری که ساختمان داشته به سال	X_4

این داده ها بدین شکل در Excel نوشته می شود.

سپس ناحیه ای را که می خواهید داده ها خروجی در آنجا نوشته شود انتخاب کنید

سپس با آدرس زیر پنجرهَ تابع LINEST را باز کنید.

با انتخاب LINEST از مقوله statistical،OK کنید،

تا پنجره LINEST باز شوددرقسمتknown-y sعبارتE2:E12 و در قسمتkown-x s عبارت A2:D12 را تایپ کنید،سپس در قسمت const وstats عبارت True را تایپ کنید.

با نگه داشتن ctrl+shift و زدن ok نتایج در جایی که قبلا" انتخاب کردید، نوشته می شود.

231.8145	2709.2	12618.39	25.5609	56587.02
13.72808	549.07	413.9391	5.617636	12661.69
0.996544	1004.233	#N/A	#N/A	#N/A
432.4997	6	#N/A	#N/A	#N/A
1.74E+09	6050904	#N/A	#N/A	#N/A

۲- تو ضیح نتایج خروجی:

ردیف اول ضرایب متغیر های مستقل و عدد ثابتb را نشان می دهد.در واقع همان ها وb در معادله هستند که بصورت بر عکس از راست به چپ نوشته شده اند.یعنی از سمت راست ترین عدد مقدارb وعدد بعدی ضریبx_1 کهfloor space می باشد نشان داده شده و آخرین عدد سمت چپ ضریبx_4 یعنی Age می باشد.

ردیف دوم نتایج خروجی خطای استاندارد هر ضریب و عدد ثابت b را نشان می دهد.که میزان انحراف نتایج هر ضریب از مقدار میانگین آن نشان می دهد و مثلا"عدد 13.72808 میزان انحراف ضریب x_4 را از مقدار میانگین آن نشان می دهد.

داده اول ردیف سوم میزان ضریب همبستگی بین مقدار تخمینی متغیر وابسته و مقدار واقعی این متغیر را نشان می دهد که بین 1- و1 می باشد و هر چه این ضریب به1-,1 نزدیکتر باشد نشان می دهد که معادله رگرسیون پیشگویی خوبی برای مقدارهای واقعی متغیر وابسته است و هر چه به صفر نزدیک تر باشد نشان می دهد که روش رگرسیون خطی نامعتبر است که یک معیار برای فهمیدن اینکه رگرسیون خطی مناسب است یا خیر می باشد.دادهَ دوم در همین سطر خطای استاندارد متغیر وابسته را از میانگین آن نمایش می دهد.

دادهَ اول سطر چهارم یکی از داده های خروجی مفید است و معیاری خوبیست برای پی بردن به اینکه آیا رگرسیون خطی مناسب است یا خیر. این مقدار آماره F را بدست می دهد.این آماره بعنوان توزیع نمونه گیری دومتغیرتصادفی مستقل که بر درجه آزادیشان تقسیم شده اند،مورد مطالعه قرار می گیرد.برای تفسیر F باید به دو درجه آزادی دسترسی داشته باشیم،این مقادیر اغلب جدول توزیعF را به دو مقدار V_1 وV_2 تفکیک می کند.مقدارV_1 تعداد متغیر های مستقل می باشد که در اینجا برابر٤ میباشد وv_2 حاصل تفریق تعداد کل دادهها از کل متغیرهانیز٥تا می باشد(٤تا مستقل و یک وابسته) پس v_2 برابر است با ٦ =٥-١١ در اینجا از سطح ٥% نیز برای آزمون آماره Fاستفاده می شود.با این توضیح می توان مقدار٦و٥و٥% F را ازجدول توزیع F که در کتابهای آمارموجود می باشد بدست آورد.خلاصه مطلب اینکه اگر مقدار اول داده اول سطر چهارم از مقدار $F_{0/05,V1,V2}$ که از جدول بدست می آید بزرگتر باشد،معادله رگرسیون به دقت توانسته ارزش متغیر وابسته را تعیین کند یا به عبارتی رابطه مشاهده شده بین متغیر های مستقل ومتغیر وابسته اتفاقی نیست واگر این مقدارکوچکترباشداین روش، روش معتبری برای داده ها نیست،یا به صورت آماری می گوییم فرض صفر را مبنی بر اتفاقی بودن رابطهَ متغیر های مستقل و متغیر وابسته را نمی توان رد کرد.

داده دوم در همین سطر درجه آزادی V_2 را نشان می دهد.در سطر آخر نیز نشان دهندهَ مجموع مربعات رگرسیون و مجموع مربعات باقیمانده است که از لحاظ آماری فرمول آن بدین شکل است:

که میانگین است و مقدار هر داده می باشد.