((به نام خدا))

متدهای آماری در داده کاوی

Statistical Methods in Data Mining

تهیه کنندگان :

مهندس ایمان اشکاوندراد

مهندس وحید ابریشمی

مقدمه :

فن آمارگری علم جمع آوری و سازمان دهی اطلاعات و ترسیم نتیجه از مجموعه داده ها است. سازمان دهی و توصیف مشخصه های عمومی مجموعه داده ها موضوع مورد مطالعه در حیطه آمار تشریحی است .این فصل بر روی اصول اولیه استنتاج آماری تاکید می کند .

تحلیل داده های آماری دارای بیشترین مجموعه متدولوژی ها برای داده کاوی است. به صورت تاریخی اولین کامپیوتر بر پایه برنامه های تحلیل داده با پشتیبانی از آمارگرها توسعه پیدا کرد. آمار روشهای متنوعی شامل رگرسیون و تحلیلی تفکیکی برای داده کاوی عرضه می کند .

۵-۱- استنتاج آماری

مجموع مشاهداتی که برروی تحلیل آماری مورد نظری صورت می گیرد صرف نظر از متناهی یا نامتناهی بودن آن بستگی به واژه ای به نام جمعیت دارد. در زمینه استنتاج آماری، علاقه مندیم وقتی که برسی تمام مشاهدات موجود غیر ممکن یا غیر عملی باشد به یک نتیجه مطلوب برسیم. به طور مثال تست کردن تمام لامپ های روشنایی یک محصول خاص برای رسیدن به میانگین طول عمر لامپ به صورت عملی غیر ممکن است. بنابر این باید زیرمجموعه ای از جمعیت برای تحلیل آماری بسنده کنیم که به آن نمونه یا مجموعه داده ها هم میگویند.

از مجموعه داده های داده شده یک مدل آماری از جمعیت می سازیم که به ما کمک می کند استنتاج مورد نظر را از جمعیت تولید کنیم. اگر نتایج استنتاج معتبر بود باید به نمونه هایی برسیم که معرف جمعیت باشد. اغلب سعی می کنیم مجموعه داده های نمونه را از عناصری انتخاب کنیم که در دسترس هستند، ولی روش مذکور ممکن است باعث ایجاد خطا در استنتاج شود . بنابر این سعی می کنیم داده های انتخابی به صورت کاملاً تصادفی انتخاب شوند.

تئوری استنتاج آماری شامل آن دسته از روشهایی است که تولید استنتاج از جمعیت می کنند ، این روشها به دو دسته روش تخمینی و روش تست فرضیه تقسیم می شوند.

در روش تخمینی هدف رسیدن به یک یا مجموعه ای از مقادیر پذیرفتنی برای پارامترهای ناشناخته سیستم است.

هدف بدست آوردن اطلاعات از مجموعه داده های T به ترتیبی که بتوان تخمینی از پارامترهای W وابسته به مدل دنیای حقیقی سیستم $F(X,W)$ زد.

$T=\{(x11,\ldots,x1n),(x21,\ldots,x2n),\ldots\}$
$X=\{x1,\ldots,xn\}$

زمانی که پارامترهای مدل تخمین زده شدند ، میتوانیم از آنهای برای پیشگویی متغیر تصادفی Y متعلق به مجموعه اولیه X بر پایه $x^* = x-y$ استفاده کنیم .

اگر Y مقداری عددی بود با رگرسیون و اگر مقداری گسسته بود با طبقه بندی روبرو هستیم .

در مورد روش تست فرضیه آماری ، شخص میخواهد تصمیم بگیرد که از لحاظ تحلیل مجموعه داده ها آیا فرضیه مورد نظر باید پذیرفته یا رد شود. یک فرضیه آماری ادعا یا تخمینی درباره یک یا چند مجموعه است. درستی یا غلطی فرضیه آماری به صورت مطلق هیچ وقت معلوم نمی شود مگر با آزمایش تمام جمعیت . بنابراین در اکثر موقعیت ها غیر کاربردی است.
تست فرضیه را برروی مجموعه ای تصادفی از داده ها انجام می دهیم. شواهد نشان می دهد مجموعه داده هایی که با حالت فرضیه ناسازگارند منجر به رد آن و در غیر این صورت قبول آن و یا به صورت دقیقتر عدم وجود شواهد قوی برای رد آن می شوند .
H0 در این تست وضعیتی است که در صورتی رد میشود که شواهد قوی مبنی بر درست بودن آن وجود داشته باشد رد H0 منجر به پذیرش فرضیه مقابل در مورد جمعیت می شود.

۵-۲- ارزیابی تفاوتها در مجموعه داده ها

برای وظیفه های داده کاوی زیادی آگاهی یافتن از خصوصیات عمومی مجموعه داده ها سودمند است.
دوگروه از این خصوصیات شامل :

۱. احتمال مطابقت داده با مقادیر مورد انتظار
۲. پراکندگی داده ها

یکی از شایع ترین راه های اندازه گیری عددی مرکز داده ها میانگین است

$$mean = 1/n \sum_{i=1}^{n} x_i$$

در برخی موارد هر xi دارای وزنی است که منعکس کننده تناوب رخ داد آن داده است که به این صورت محاسبه می شود

$$mean = \sum_{i=1}^{n} w_i \, x_i / \sum_{i=1}^{n} w_i$$

برای داده های اریب معیار اندازه گیری بهتری به نام میانه وجود دارد

$$median = \begin{cases} x_{(n+1)/2} & \text{if n is odd} \\ (x_{n/2} + x_{(n/2)+1})/2 & \text{if n is even} \end{cases}$$

معیار اندازه گیری دیگرMode است که ارزشی است که بیشترین تناوب رخ داد را در مجموعه داشته باشد.

$$mean - mode = 3 \times (mean - median)$$

درجه ای که داده های عددی گرایش به پخش شدن نسبت به مرکز داده را دارند را پراکندگی و معیار اندازه گیری آنرا انحراف معیار استاندارد و واریانس گویند .

$$\sigma^2 = (1/(n-1)) \sum_{i=1}^{n} (x_i - mean)^2$$

خصوصیات اولیه انحراف معیار

- حول میانگین است و زمانی که از میانگین استفاده شود قابل استفاده است.
- اگر صفر باشد یعنی پراکندگی نداریم در واقع یعنی همه داده ها با هم برابر هستند.

۵-۳- استنباط بیزین(Bayesian Inference)

خیلی سخت نیست حالتی را فرض کنیم که در آن داده تنها منبع اطلاعات قابل دسترس در مورد جمعیت نیست. متد بیزین فراهم کننده یک راه اصولی برای یکپارچه کردن این اطلاعات خارجی به فرایند تحلیل داده است.

این فرایند با یک احتمال توزیع از پیش داده شده برای تحلیل مجموعه داده ها شروع می شود. به این توزیع اولیه پیش توزیع و به توزیع بروز رسانی شده فرایند پس توزیع گویند که اساس آن قضیه Bayes است.

$$P(H/X) = [P(X/H) \cdot P(H)]/P(X)$$

بخاطر اینکه محاسبه $P(x/ci)$ خیلی پیچیده است ، مخصوصا برای مجموعه داده های بزرگ ، فرض ساده ای از استقلال شرطی بین خواص تولید می شود. با استفاده از این فرض می توانیم $P(x/ci)$ را به این صورت محاسبه کنیم :

$$P(X/C_i) = \prod_{t=1}^{n} P(x_t/C_i)$$

در تئوری Bayesian Classifier دارای کمترین نرخ خطا در مقایسه با دیگر Classifier های توسعه یافته در داده کاوی است. در عمل همیشه این موضوع به علت نادرستی و کم دقتی در فرضهای صفات و شروط استقلال کلاسها برقرار نمی باشد.

۴.۵ رگرسیون پیشگویانه :

پیشگویی مقادیر پیوسته می تواند توسط یک مدل آماری به نام رگرسیون مدل شود . هدف رگرسیون پیدا کردن بهترین مدل برای مربوط کردن یک متغیر خروجی به مقادیر مختلف ورودی است . به طور رسمی تر تحلیل رگرسیون فرآیند محاسبه این است که چگونه یک متغیر مانند Y وابسته به متغیرهای دیگری مانند X1 و X2 و ... و Xn است . معمولا Y را جواب خروجی یا متغیر وابسته می نامند و Xi را ورودی ها ، رگرسورها ، متغیرهای توضیحی یا متغیرهای مستقل می نامند .

دلایل رایج برای انجام تحلیل رگرسیون عبارتند از :

۱ . اندازه گیری خروجی ها پرهزینه است اما اندازه گیری ورودی ها کم هزینه .

۲ . مقادیر ورودی ها قبل از خروجی ها معلوم است بنابراین نیاز به پیش بینی است .

۳ . کنترل مقادیر ورودی ، ما می توانیم رفتار خروجی های خروجی های وابسته را پیش بینی نماییم .

۴ . ممکن است یک رابطه سببی بین ورودی ها و خروجی باشد و ما می خواهیم این رابطه را کشف نماییم .

مدل های خطی عمومی اکنون رایج ترین تکنیک آماری پذیرفته شده است . آن ها برای توضیح ارتباط بین تمایل یک متغیر و مقادیری که توسط متغیرهای دیگر اخذ شده است به کار می روند . مدل کردن این نوع از ارتباط ، رگرسیون خطی نامیده می شود . متناسب سازی یک مدل تنها وظیفه در مدل سازی آماری نیست . ما معمولا می خواهیم تنها یک مدل را از میان مدل های ممکن که از بقیه مناسب تر است را انتخاب کنیم .یک روش برای انتخاب از میان چندین مدل آنالیز واریانس نامیده می شود که در قسمت ۵.۵ توضیح داده خواهد شد .

رابطه ای که یک مجموعه از داده ها را تطبیق می دهد ، به وسیله یک مدل پیش بینی که معادله رگرسیون نامیده می شود ، مشخص می شود . رایج ترین فرم استفاده مدل رگرسیون ، مدل خطی عمومی است که به صورت زیر نوشته می شود :

$$Y = \alpha + \beta_1 . x + \beta_2 . x_2 + \beta_3 . x_3 + ... + \beta_n . x_n$$

به وسیله اعمال این معادله بر روی هر یک از این نمونه ها ، مجموعه معادلات جدیدی به دست می آوریم :

$$y_j = \alpha + \beta_1 . x_{1j} + \beta_2 . x_{2j} + \beta_3 . x_{3j} + ... + \beta_n . x_{nj} + \varepsilon_j$$

جایی که ε_j ها خطای رگرسیون برای هریک از m نمونه داده شده است . مدل خطی از آنجا خطی نامیده می شود که که مقدار مورد انتظار y_j یک تابع خطی است : جمع وزنی مقادیر ورودی .

رگرسیون خطی با استفاده از یک متغیر ساده ترین فرم رگرسیون است که یک متغیر تصادفی Y (متغیر پاسخ نامیده می شود) به وسیله یک تابع خطی از یک متغیر تصادفی دیگر X (متغیر پیش بینی نامیده می شود) را مدل می کند .

رگرسیون خطی با داشتن n نمونه به فرم (x_1, y_1) , (x_2, y_2), …, (x_n, y_n) ، به صورت زیر بیان می شود:

$$Y = \alpha + \beta.X$$

در جایی که α و β ضرایب رگرسیون هستند . با فرض اینکه واریانس Y ثابت است ، این ضرایب می توانند با استفاده از کمترین مربعات که خطای بین نقاط داده واقعی و خط تخمین زده شده را مینیمم می کند ، به دست آیند . مجموع باقیمانده مربعات معمولا در مورد خط رگرسیون ، مجموع مربعات خطاها نامیده شده و به صورت SSE نمایش داده می شود .

$$SSE = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - y_i')^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

به طوری که y_i مقدار واقعی خروجی برای مجموعه داده ، داده شده است و y_i' مقدار پاسخ است که از مدل به دست آمده است . با مشتق گیری از رابطه بالا نسبت به α و β داریم :

$$\partial(SSE)/\partial\alpha = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)$$

$$\partial(SSE)/\partial\beta = -2\sum_{i=1}^{n}((y_i - \alpha - \beta x_i).x_i)$$

معادله های بالا را برابر با صفر قرار داده و نتایج زیر حاصل می شود :

$$n\alpha + \beta\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} x_i y_i$$

که می توانند به صورت همزمان برای محاسبه یک فرمول برای محاسبه α و β به کار روند . به کار بردن روابط استاندارد برای مقادیر میانگین ، ضرایب رگرسیون برای این مورد ساده بهینه سازی عبارت است از :

$$\beta = \left[\sum_{i=1}^{n}(x_i - mean_x).(y_i - mean_y)\right] / \left[\sum_{i=1}^{n}(x_i - mean_x)^2\right]$$

$$\alpha = mean_y - \beta.mean_x$$

که $mean_x$ و $mean_y$ مقادیر میانگین برای متغیرهای تصادفی X و Y در مجموعه داده آموزشی است . باید این موضوع مهم را به خاطر داشته باشیم که مقادیر ما از α و β که بر مبنای مجموعه داده ، داده شده است ، تنها تخمینی از تمام جمعیت است . معادله $Y = \alpha + \beta.X$ ممکن است برای محاسبه مقدار میانگین پاسخ y_0 برای ورودی x_0 استفاده شود ، که لزوما از مجموعه داده اولیه نیست .

به عنوان مثال اگر مجموعه داده نمونه به صورت جدول باشد (جدول ۲.۵) و ما خواسته باشیم رگرسیون خطی بین دو متغیر (متغیر پیش بینی A و متغیر پاسخ B) را تحلیل نماییم ، معادله رگرسیون به شکل زیر در می آید :

$$B = \alpha + \beta.A$$

به طوری که α و β ضرایب رگرسیون می توانند براساس فرمول های پیشین به دست آیند .

$$\alpha = 0.8$$
$$\beta = 1.04$$

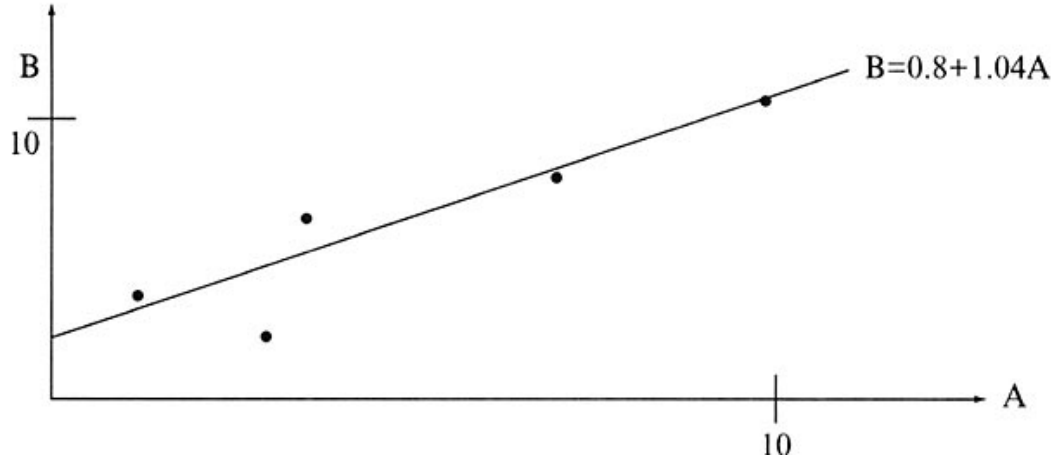
خط رگرسیون بهینه به شکل زیر به دست می آید :

$$B = 0.8 + 1.04.A$$

جدول ۲.۵ : یک پایگاه داده برای کاربرد مدل رگرسیون

A	B
1	3
8	9
11	11
4	5
3	2

مجموعه داده اولیه و خط رگرسیون در شکل ۲.۵ نمایش داده شده است :



شکل ۲.۵ : رگرسیون خطی برای مجموعه داده در جدول ۲.۵

رگرسیون چندتایی یک گسترش از رگرسیون خطی با یک ورودی است و شامل بیشتر از یک متغیر پیشگویی . متغیر پاسخ Y به صورت یک تابع خطی از چندین متغیر پیش بینی می تواند مدل شود . برای مثال اگر متغیرهای پیشگویی به صورت x_1و x_2و x_3 باشد آنگاه معادله رگرسیون خطی چندتایی به شکل زیر در می آید :

$$Y = \alpha + \beta_1.x + \beta_2.x_2 + \beta_3.x_3 + ... + \beta_n.x_n$$

به طوری که α و β_1 و β_2 و β_3 ضریب هایی هستند که با استفاده از روش کمترین مربعات به دست آمده اند . برای رگرسیون خطی با بیشتر از یک متغیر ورودی ، بهتر است که فرآیند محاسبه پارامترهای β را به وسیله محاسبات ماتریسی تحلیل نماییم :

$$Y = \alpha.\beta.X$$

به طوری که $\beta = \{\beta_0, \beta_1, ..., \beta_n\}, \beta_0 = \alpha$ و X و Y ماتریس های ورودی و خروجی برای مجموعه داده آموزشی هستند . مجموع مربعات خطاها نیز به صورت ماتریسی و به شکل زیر خواهد بود :

$$SSE = (Y - \beta.X)'.(Y - \beta.X)$$

و بعد از بهینه سازی :

$$\partial(SSE)/\partial(\beta) = 0 \Rightarrow (X'.X).\beta = X'.Y$$

بردار نهایی β معادله ماتریس را ارضا می نماید :

$$\beta = (X'.X)^{-1}(X'.Y)$$

به طوری که β بردار تخمین زده شده برای ضرایب ، در رگرسیون خطی است . ماتریس های X و Y ابعاد یکسانی دارند و برابر با مجموعه داده است . بنابراین یافتن یک راه حل بهینه برای بردار β در مسائلی با چند صد نمونه کار ساده ای است . برای مسائل داده کاوی در دنیای واقعی ممکن است تعداد نمونه ها به چندین میلیون افزایش یابد . در این مواقع به خاطر وجود ابعاد زیاد در ماتریس و افزایش نمایی پیچیدگی الگوریتم ، باید از تخمین ها را در مسئله پیدا کنیم ، یا از یک روش متفاوت رگرسیون استفاده نماییم .

یک کلاس بسیار بزرگ از مسائل رگرسیون وجود دارند که غیر خطی هستند . برای مثال به صورت یک چند جمله ای به صورت زیر :

$$Y = \alpha + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_1X_3 + \beta_4.X_2X_3$$

که می تواند با قرار دادن $X_4 = X_1 \cdot X_3$ و $X_5 = X_2 \cdot X_3$ به فرم خطی تبدیل شود . همچنین رگرسیون های چند جمله ای می توانند به وسیله اضافه کردن یک عبارت چند جمله ای به یک مدل خطی پایه ، مدل شوند . به عنوان مثال فرم یک منحنی چندجمله ای درجه ۳ به شکل زیر است :

$$Y = \alpha + \beta_1.X + \beta_2.X^2 + \beta_3.X^3$$

با اعمال انتقال بر متغیرهای پیش بینی $(X_1 = X, X_2 = X^2, X_3 = X^3)$، امکان تبدیل آن به فرم خطی و انتقال آن به یک مسئله رگرسیون چندتایی را فراهم ساخته که حل به وسیله کمترین

مربعات ممکن می شود . بایستی به این نکته توجه شود که عبارت خطی در عبارت مدل سازی عمومی خطی ، در مورد متغیرهای وابسته که یک تابع خطی از پارامترهای ناشناخته هستند ، صدق می نماید . بنابراین یک مدل خطی عمومی ممکن است شامل متغیرهای غیروابسته از درجات با لاتر مانند $e^{\beta x}, X_1 \cdot X_2, 1/X, \text{or } X^3_2$ باشد . اصل این است که بتوانیم انتقال مناسب از متغیرهای ورودی یا ترکیب آن ها به دست آوریم .
بعضی از انتقالات مفید برای خطی سازی مدل رگرسیون در جدول ۳.۵ داده شده است .

جدول ۳.۵ : بعضی انتقالات برای خطی سازی رگرسیون

Table 5.3: Some useful transformations to linearize regression

Function	Proper transformation	Form of simple Linear regression
Exponential: $Y = \alpha e^{\beta x}$	$Y^* = \ln Y$	Regress Y^* against x
Power: $Y = \alpha x^{\beta}$	$Y^* = \log Y$; $x^* = \log x$	Regress Y^* against x^*
Reciprocal: $Y = \alpha + \beta(1/x)$	$x^* = 1/x$	Regress Y against x^*
Hyperbolic: $Y = x/(\alpha + \beta x)$	$Y^* = 1/Y$; $x^* = 1/x$	Regress Y^* against x^*

بیشترین تلاش در سمت کاربر ، در به کار بردن تکنیک های رگرسیون چندتایی، در شناخت متغیرهای غیروابسته مربوط از مجموعه اولیه و در انتخاب مدل رگرسیون که فقط متغیرهای مربوط را به کار می برد نهفته است . دو روش رایج برای انجام این وظیفه وجود دارد :
۱ = جستجوی ترتیبی : ساخت یک مدل رگرسیون ابتدایی و سپس اضافه یا حذف کردن متغیرها تا زمانی که معیارهایی ارضا شود .
۲ = روش ترکیبی : تمام ترکیب های متغیرهای غیر وابسته را ترکیب کرده و یک مدل رگرسیون مناسب را انتخاب می کنیم .
مراحل اضافی پس پردازش می توانند کیفیت مدل رگرسیون خطی را تخمین بزنند . آنالیز وابستگی تلاش می نماید که قدرت رابطه بین دو متغیر را اندازه گیری نماید . یک پارامتر که نمایش دهنده قدرت بین دو متغیر با استفاده از یک عدد است ، ضریب وابستگی نامیده می شود . محاسبه این ضریب نیازمند بعضی نتایج میانی در تحلیل رگرسیون است .

$$r = \beta . \sqrt{(S_{xx} / S_{yy})} = S_{xy} . \sqrt{(S_{xx} / S_{yy})}$$

که :

$$S_{xx} = \sum_{i=1}^{n} (x_i - mean_x)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - mean_y)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - mean_x)(y_i - mean_y)$$

مقدار r بین ۱ و -۱ است . مقادیر منفی برای r مربوط به خط های رگرسیون با دامنه منفی است و مقادیر مثبت نشان دهنده شیب مثبت است . بایستی در تحلیل مقدار r ، بسیار دقت نماییم . برای مثال اگر r=0.3 و r=0.6 به این معنی است که ما تنها دو وابستگی مثبت داریم که دومی از اولی قوی تر است . این نتیجه گیری که وابستگی که r=0.6 نشان می دهد ، دو برابر قوی تر از r=0.3 است ، اشتباه است .

برای مثال ساده رگرسیون خطی ما که در ابتدای این بخش مطرح شد ، مدل به دست آمده به صورت B = 0.8 + 1.04A بود . ممکن است ما کیفیت این مدل را به وسیله ضریب همبستگی به دست آوریم :

$$S_{AA} = 62$$

$$S_{BB} = 60$$

$$S_{AB} = 52$$

و در نهایت داریم :

$$r = 52 / \sqrt{62.60} = 0.85$$

یک ضریب وابستگی به شکل r=0.85 نشان دهنده یک ارتباط خوب بین دو متغیر است . یک تفسیر اضافی دیگر نیز ممکن است . چون $r^2 = 0.72$ ما می توانیم بگوییم که ۷۲٪ از تغییرات در مقدار B توسط ارتباط خطی با A ایجاد می شود .

۵.۵ تحلیل واریانس :

بعضی از اوقات تحلیل واریانس به عنوان روشی برای تعیین کیفیت رگرسیون به کار می رود . در این روال تمام تغییرات متغیر وابسته تقسیم به یک سری اجزای معنی دار می شود که سپس با آنها به شکلی سیستماتیک برخورد می شود .

تحلیل واریانس یا ANOVA یک روش برای شناسایی این است که کدامیک از β ها در مدل رگرسیون خطی ، غیر صفر هستند . فرض کنید پارامترهای β قبلا توسط الگوریتم کمترین مربعات خطاها تعیین شده اند . باقیمانده ها عبارتند از تفاوت بین خروجی مشاهده شده و مقادیر تناسب داده شده :

$$R_i = y_i - f(x_i)$$

سایز باقیمانده ها برای همه m نمونه در مجموعه داده وابسته است به سایز واریانس که می تواند توسط فرمول زیر تخمین زده شود :

$$S^2 = \left[\sum_{i=1}^{m}(y_i - f(x_i))^2\right]/(m-(n-1))$$

نکته کلیدی درباره S^2 آن است که به ما اجازه می دهد که مدل های خطی مختلف را مقایسه نماییم . اگر مدل تخمین زده شده مناسب باشد ، آن گاه S^2 تخمین مناسبی از σ^2است . اگر مدل تخمین زده شده شامل عبارات باقیمانده باشد (بعضی از β واقعا صفر هستند) آن گاه S^2 هنوز خوب و نزدیک به σ^2 است . تنها اگر مدل تخمین زده شده شامل یا چند متغیر ورودی نشود در حالی که باید شامل می شد ، آن گاه مقدار S^2 بسیار بزرگتر از مقدار σ^2 خواهد بود . این پارامترها قدم های ابتدایی تصمیم گیری در الگوریتم ANOVA هستند که در آن ما تاثیر یک متغیر را در یک مدل نهایی بررسی می نماییم . ابتدا ما از تمام ورودی ها آغاز کرده و مقدار S^2 برای این مدل محاسبه می نماییم . سپس ما متغیرهای ورودی را یک به یک حذف می نماییم . اگر یک متغیر ورودی مفید را حذف نماییم ، مقدار S^2 به صورت چشمگیری افزایش خواهد یافت . اما اگر یک متغیر ورودی اضافی را حذف نماییم ، مقدار S^2 تغییر چندانی نمی کند . توجه کنید حذف یک متغیر از مدل به معنای وادار کردن β آن به صفر است . در اصل ، در هر تکرار ما دو مقدار S^2 را با یکدیگر مقایسه می نماییم و تفاوت آنها را تحلیل می نماییم . برای این منظور ما تست F-ratio یا F-static را به فرم :

$$F = S^2_{new} / S^2_{old}$$

معرفی می نماییم . اگر مدل جدید (بعد از حذف یک ورودی یا بیشتر) کافی باشد ، آنگاه F به یک نزدیک خواهد بود و در غیر این صورت این مقدار بسیار بیشتر از یک خواهد بود . با به کاربردن این روش تکراری ANOVA ما می توانیم بفهمیم که خروجی به کدام ورودی ها مربوط و به کدام ها مربوط نمی باشد . روال ANOVA تنها هنگامی به درستی عمل می نماید که مدل هایی که با هم مقایسه می شوند به صورت لانه ای باشند ، به عبارت دیگر یک مدل یک نوع خاص از مدل دیگر است .

فرض کنید که مجموعه داده دارای سه متغیر ورودی به شکل x_1و x_2و x_3و یک متغیر خروجی به شکل Y باشد . فرض کنید ما روش ANOVA را بر روی آن اعمال نموده ایم . نتایج را در جدول ۴.۵ مشاهده می نمایید .

جدول ۴.۵ : تحلیل ANOVA برای مجموعه داده با سه ورودی

Table 5.4: ANOVA analysis for a data set with three inputs x_1, x_2, and x_3

Case	Set of inputs	S^2_i	F
1	x_1, x_2, x_3	3.56	
2	x_1, x_2	3.98	$F_{21} = 1.12$
3	x_1, x_3	6.22	$F_{31} = 1.75$
4	x_2, x_3	8.34	$F_{41} = 2.34$
5	x_1	9.02	$F_{52} = 2.27$
6	x_2	9.89	$F_{62} = 2.48$

نتیجه تحلیل ANOVA نشان می دهد که ویژگی x3 تاثیری در خروجی ندارد زیرا مقدار F-ratio نزدیک به ۱ است :

$$F_{21} = S_2 / S_1 = 3.98/3.56 = 1.12$$

در بقیه موارد ، زیر مجموعه های ورودی به طور قابل توجهی F-ratio را افزایش می دهند .
بنابراین دیگر امکانی برای حذف متغیرهای دیگر ورودی وجود نداشته و بنابراین مدل به شکل زیر در می آید :

$$Y = \alpha + \beta_1.X_1 + \beta_2.X_2$$

تحلیل چند تغییری واریانس یک تعمیم از تحلیل ANOVA است که توجه به تحلیل داده ها در شرایطی دارد که خروجی به جای یک مقدار یک بردار است . یک راه حل برای تحلیل این نوع از داده مدل کردن هر یک از خروجی ها به صورت جداگانه است که ممکن است باعث نادیده گرفتن ارتباط بین خروجی ها شود . به عبارت دیگر ، تحلیل بر مبنای این است که خروجی ها به هم مرتبط نیستند . تحلیل چند مقداری واریانس یک فرم از تحلیل بوده که اجازه ارتباط بین خروجی ها را می دهد . با استفاده از مجموعه داده شده از ورودی ها و خروجی ها ، ما می توانیم مجموعه داده موجود را به وسیله مدل خطی چند تغییری تحلیل نماییم :

$$y_j = \alpha + \beta_1.x_{1j} + \beta_2.x_{2j} + \beta_3.x_{3j} + ... + \beta_n.x_{nj} + \varepsilon_j$$

جایی که n تعدا ابعاد ورودی ، m تعداد نمونه ها و y_j یک بردار با ابعاد c*1 است و c تعداد خروجی ها است . این مدل می تواند توسط تخمین حداقل مربعات مانند یک مدل خطی حل شود . یک راه برای انجام این تناسب سازی ، نسبت دادن یک مدل خطی به هریک از خروجی ها ، یکی بعد از دیگری است . باقیمانده وابسته برای هر بعد به صورت $y_i - y'$ خواهد بود که که y_i

مقدار واقعی خروجی برای مجموعه داده ، داده شده است و y_i' مقدار پاسخ است که از مدل به دست آمده است .

باقیمانده مجموع مربعات برای مدل خطی تک تغییری یک ماتریس از باقیمانده مجموع مربعات برای مدل خطی چند تغییری است . ماتریس R به صورت زیر تعریف می شود :

$$R = \sum_{j=1}^{m} (y_j - y_j')(y_j - y_j')^T$$

ماتریس R شامل باقیمانده مجموع مربعات برای هر یک از c بعد ، بر روی قطر اصلی است . عناصری که بر روی قطر اصلی نیستند ، جمع باقیمانده ها جفت ها به صورت عرضی هستند . اگر بخواهیم دو مدل خطی لانه ای را به منظور محاسبه اینکه بعضی از β ها برابر صفر هستند ، با هم مقایسه کنیم ، می توانیم یک ماتریس اضافی مجموع مربعات ایجاد کرده و یک روش مانند ANOVA به صورت چند تغییری اعمال نماییم . در حالی که ما در ANOVA ، F-static را داریم ، MANOVA بر اساس ماتریس R و ۴ تست رایج آماری به نام های Roy's greatest root, the Lawley-Hotteling trace, the Pillai trace, and Wilks' lambda است .

۵-۶- رگرسیون Logistic

رگرسیون خطی برای مدل کردن توابع با مقادیر پیوسته به کار می رود. به طور کلی مدل های رگرسیون تعمیم یافته اساس تئوری رگرسیون خطی است که می تواند برای مدل هایی که متغیر پاسخ آنها گسسته (دسته ای) می باشد ، استفاده شود. یک نوع معمول مدل خطی رگرسیون لجستیک است. مدل رگرسیون لجستیک احتمال اتفاق افتادن تعدادی رویداد به عنوان توابع خطی از یک مجموعه متغیرهای پیشگویی شده را نشان می دهد.

مدل رگرسیون لجستیک بجای پیش بینی مقادیر متغیرهای وابسته ، سعی در برآورد احتمال P در متغیرهای وابسته که دارای مقدار معینی هستند، دارد. برای مثال به جای پیش بینی این که آیا یک مشتری نرخ اعتبار خوب یا بدی دارد، رگرسیون لجستیک تلاش می کند که احتمال خوب بودن نرخ اعتبار مشتری را تخمین بزند. حالت واقعی متغیرهای وابسته ، با احتمالات تخمین زده شده تعیین می‌شود. اگر احتمال برآورده شده بزرگتر از ۰/۵ باشد آنگاه تخمین به Yes (درجه اعتبار خوب) نزدیکتر است در غیر این صورت خروجی به No (درجه اعتبار بد) نزدیک است. بنابراین در رگرسیون لجستیک احتمال P ، احتمال موفقیت نامیده می شود.

هنگامی از رگرسیون لجستیک استفاده می کنیم که متغیرهای خروجی در دسته های دوتایی تعریف شوند. از طرف دیگر، هیچ دلیلی برای این که داده ها کمی نباشند وجود ندارد و بنابراین رگرسیون لجستیک از یک مجموعه داده های ورودی جامع پشتیبانی می کند.

فرض کنید که خروجی Y دارای دو مقدار قطعی ممکن ۰ و ۱ باشد. بر پایه داده های موجود ما میتوانیم احتمال قابل دسترس بودن برای هر دو مقدار از ورودیهای نمونه محاسبه کنیم.

$$P(y_j = 1) = P_j \quad , \quad p(y_j = 0) = 1 - P_j$$

مدلی که برای احتمال مورد نظر مناسب می باشد ، رگرسیون خطی است که به صورت زیر محاسبه می‌گردد:

$$\log(p_j/(1-p_j)) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \ldots + \beta_n \cdot X_{nj}$$

این معادله مدل لجستیک خطی نامیده می شود. تابع $\log(P_j /(1- P_j))$ اغلب به صورت Logit(p) نوشته می شود. دلیل اصلی استفاده از شکل خروجی Logit این است که با توجه به اینکه دامنه P_j ،$[0,1]$ است از ایجاد دامنه دیگری برای پیشگویی احتمال P_j جلوگیری می شود. در اینجا ، مدل تضمینی بر اسا مجموع داده ترتیبی و به کار گیری تابع رگرسیون خطی در نظر بگیرید. رگرسیون خطی با معادله زیر ارائه شده است:

$$Logit(p) = 1.5 - 0.6 \,.\, X_1 + 0.4 \,.\, X_2 - 0.3 \,.\, X_3$$

و همچنین فرض کنید که نمونه جدید برای دسته بندی مقادیر ورودی را به صورت زیر داریم :

$$\{X_1, X_2, X_3\} = \{1,0,1\}$$

استفاده از مدل لجستیک خطی این را امکان پذیر می سازد که احتمال خروجی با مقدار 1 و (P(Y=1)) را برای این نمونه تخمین بزنیم. اول: Logit(p) را محاسبه می کنیم:

$$\text{logit}(p) = 1.5 - 0.6 \cdot 1 + 0.4 \cdot 0 - 0.3 \cdot 1 = 0.6$$

و سپس احتمال مقدار خروجی یک را برای ورودی معین می کنیم :

$$\text{Log}\,(\,p\,/\,(1\text{-}p)) = 0.6$$

$$P = e^{-0.6}/\,(1+ e^{-0.6}) = 0.35$$

بر اساس مقدار نهایی برای احتمال P ، ممکن است به این نتیجه برسیم که احتمال y=۱ کمتر از دیگر مقادیر رده بندی با مقدار y=۰ است. این مثال نشان می دهد که رگرسیون لجستیک یک ابزار دسته بندی خیلی ساده ولی قوی در کاربردهای داده کاوی است.

مجموعه ی داده (مجموعه آموزشی) این امکان را فراهم می کند که یک مدل رگرسیون لجستیک را ایجاد کنیم و با مجموعه داده دیگر (مجموعه تست) ما ممکن است کیفیت مدل را در پیش بینی مقادیر قطعی تحلیل کنیم. در نتیجه رگرسیون لجستیک ممکن است با دیگر روشهای داده کاوی برای دسته بندی فعالیتهایی از قبیل قوانین تصمیم گیری ، شبکه های عصبی و طبقه بندی Bayesian مقایسه شود.

۵ -۷- مدل Log خطی

مدل Log خطی یک روش برای تحلیل رابطه بین متغیرهای قطعی(یا کمی) است. مدل Log خطی توزیع های احتمال چند بعدی را به صورت مجزا و گسسته تخمین می زند. در مدل خطی تولید شده خروجی Y_i یک توزیع پوآسون با مقدار μ_j فرض شده است.

لگاریتم طبیعی μ_j با تابع خطی از متغیر های وابسته (ورودیهای) زیر می با شد :

$$\log(\mu_j) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \ldots + \beta_n \cdot X_{nj}$$

از آنجایی که همه متغیرها جز متغیرهای رده بندی قرار مـی گیرنـد، از یـک جـدول فراوانـی بـرای نمایش توزیع عمومی داده ها استفاده می شود.

هدف در مدل Log خطی تشخیص ارتباط بین متغیرهای رده بندی است .در این مدل ارتباط بـین دو متغیر متناظر با اثر متقابل بین آنهاست: همچنین مسئله دیگر پیدا کردن همه β هایی است کـه در مدل ، صفر هستند.

مشابه مسئله فوق در تحلیل ANOVA نیز برقرار است. اگر یک تعامل بین متغیرها در مدل Log خطی وجود داشته باشد نشان دهنده این است که متغیرها وابسته نیستند ولی با هم نسبت دارند و β برابر با صفر نیست. در این تحلیل لازم است متغیرهای رده بندی مثل خروجـی هـا (پاسـخ هـا) مطرح شوند. اگر پاسخ ها معلوم شود، آنگاه علاوه بر مدل Log خطـی ، مـی تـوانیم از رگرسـیون لجستیک نیز برای تحلیل استفاده کنیم. سپس ما تحلیل Log خطـی را بـرای یـک مجموعـه داده بدون متغیرهای پاسخ را توضیح می دهیم . همه متغیرها از نوع داده های رده بنـدی هسـتند و مـا ارتباط ممکن بین آنها را تحلیل خواهیم کرد مسلماً این کار وظیفه تحلیل متناظر[1] می باشد.

تحلیل متناظر، مجموعه داده های رده بندی برای تحلیل ماتریس ها را نشـان مـی دهـد. خروجـی تحلیل ماتریس سئوال زیر را پاسخ می دهد که : " آیا رابطه ای بین صفات تحلیل شده وجـود دارد یا نه "

یک مثال از ماتریس ۲×۲ را در جدول ۵.۵ را مشاهده می نمایید.

		Support		
		Yes	No	Total
	Femal	309	191	500
Sex	Male	319	281	600
	Total	628	472	1100

جدول ۵-۵ یک ماتریس ۲x۲ برای ۱۱۰۰ نمونه از نظرسنجی در خصوص سقط جنین

جدول مربوطه نتیجه مطالعه یک آزمون می باشد که نظریه رابطه جنسیت مذکر و مونث را در مورد سقط جنین بررسی نموده است. جمع مجموعه نمونه ها ۱۱۰۰ است و هر نمونه شـامل خـواص رده بندی با مقادیر مشابه می باشد. برای صفت جنسیت ، مقادیر ممکن مذکر و مونث می باشند و برای صفت پشتیبان مقادیر بله و نه هستند . نتیجه کلی برای همه نمونه ها در ۴ عضـو جـدول احتمـال نشان داده شده است .

" آیا تفاوتی بین اندازه نظریه جمعیت مذکر و مونث وجود دارد ؟" این سؤال ممکن اسـت بـه ایـن صورت پرسیده شود که : " چه سطحی از وابستگی بین دو صفت جنسیت و نظریه فوق وجـود دارد ؟ " اگر یک ارتباط وجود داشت آنگاه اختلاف مهمی در نظریه بین جمعیـت مـذکر و مونـث وجـود دارد در غیر این صورت هر دو جمعیت نظریه مشابه ای دارند.

---

[1] Correspondence

با توجه به این که در مدل Log خطی ، بین متغیرهای رده بندی شده ارتباط وجود دارد ، تلاش می‌کنیم در این مدل با استفاده از جدول توافقی یک مقدار (اندازه) پیدا کنیم . اما ما نمی توانیم این را انجام دهیم. در عوض الگوریتمی را بر اساس دو جدول توافقی برای ارتباط آنها پیدا می کنیم.

۱- نخستین مرحله : انتقال جدول توافقی به یک جدول با داده های مورد انتظار. این مقادیر با فرض این که متغیرها مستقل هستند محاسبه می شوند.

۲- در مرحله دوم ، ما دو ماتریس را با استفاده از اندازه گیری توان دوم فاصله و تست Chi-Square به عنوان معیاری برای ارتباط بین دو متغیر رده بندی مقایسه می کنیم .

پردازش قابل محاسبه این دو مرحله برای جدول توافقی ۲ × ۲ خیلی ساده است. این فرآیند همچنین برای جدول توافقی با ابعاد بیشتر قابل تعمیم است. (تحلیل متغیرهای رده بندی با بیش از دو مقدار مثل ۴ × ۳ یا ۹ × ۶ قابل اجرا است). اجازه بدهید که ما یک سری نکات را معرفی کنیم .

جدول توافقی را به نام $X_{m \times n}$ در نظر بگیرید. جمع ردیف ها برای هر جدول عبارتند از :

$$X_{j+} = \sum_{i=1}^{n} X_{ji}$$

این رابطه برای هر ردیف ($j = ۱,...,m$) معتبر می باشد به طور مشابه ما می توانیم جمع ستون ها را به صورت زیر تعریف کنیم :

$$X_{+i} = \sum_{j=1}^{m} X_{ji}$$

جمع کلی به صورت جمع مجموع ردیف ها تعریف می شود:

$$X_{++} = \sum_{j=1}^{m} X_{j+}$$

یا به صورت جمع مجموع ستون ها به دست می آید :

$$X_{++} = \sum_{i=1}^{n} X_{+i}$$

با استفاده از این فرمول ارائه شده می توانیم جدول توافقی برای مقادیر مورد انتظار با فرض این که هیچ ارتباطی بین متغیرهای ردیف و ستون وجود ندارد، را محاسبه کنیم.

مقادیر مورد انتظار برای $j = 1,\ldots,m$, $i= 1,\ldots,n$ عبارتند از :

$E_{ji} = (X_{j+} . X_{+j}) / X_{++}$

که برای هر حالت از جدول توافقی محاسبه می شوند. نتیجه نهایی از نخستین مرحله رویهمرفته یک جدول جدید است که فقط شامل مقادیر مورد انتظار خواهد بود و دو جدول ابعاد مشابه ای دارند.

برای مثال ما در همه مجموع ها (ستون ها ، سطرها و جمع کلی) قبلاً دو جدول توافقی در شکل ۴.۷ نشان داده شده است. بر پایه این مقادیر ما می توانیم جدول توافقی که شامل مقادیر مورد انتظار است را ایجاد کنیم. مقدار مورد انتظار در محل برخورد سطر اول و ستون اول به صورت زیر است :

$$E11= (X_{1+} \ . X_{+1}) / X_{++} = 500.628 / 1100 = 285.5$$

به طور مشابه ما می توانیم مقادیر مورد انتظار دیگری را محاسبه نماییم و در نهایت جدول توافقی به صورت جدول ۵.۶ در خواهد آمد .

		Support		
		Yes	No	Total
	Femal	285.5	214.5	500
Sex	Male	342.5	257.5	600
	Total	628	472	1100

جدول ۶-۵ ماتریس ۲x۲ از مقادیر مورد انتظار برای داده های ارائه شده در جدول ۵-۵

مرحله بعد در تحلیل ارتباط خواص رده بندی تست Chi-Square می باشد . فرضیه اولیه H_0 بر اساس دو خاصیت غیر مرتبط می باشد و آن به وسیله فرمول Pearson's Chi-Square تست شده است که به صورت زیر می باشد:

$$\chi^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} ((X_{ji} - E_{ji})^2 / E_{ji})$$

مقادیر بزرگتر از χ^2 ، نشان دهنده گواهی قوی تری بر علیه فرض H_0 می باشد. برای مثال ما جدول های ۵.۵ و ۵.۶ را مقایسه کردیم نتیجه تست عبارتند از :

$$\chi^2 = 8.2816$$

درجه آزادی برای یک جدول با ابعاد $m \times n$ به صورت زیر محاسبه می شود :

$$d.f = (m-1)(n-1) = (2-1)(2-1) = 1$$

به طور کلی فرضیه H_0 با سطح معنی دار α رد می شود اگر :

$$\chi^2 \geq T(\alpha)$$

باشد در حالی که $T(\alpha)$ ، مقدار آستانه ای در جدول توزیع χ^2 می باشد. برای مثال ما α =۰.۰۵ انتخاب کردیم و مقدار آستانه را به این شکل به دست آوردیم :

$$T(0.05) = \chi^2(1 - \alpha, d.f.) = \chi^2(0.95, 1) = 3.84.$$

و بنابراین ما می توانیم نتیجه بگیریم که فرض H_0 رد خواهد شد. خواص در یک برآورد با ارتباط بیشتر تحلیل می شوند. به عبارت دیگر نظریه در مورد سقط جنین ، نشان دهنده اختلاف بین جنسیت مذکر و مونث می باشد.

روال های مشابه ممکن است تعمیم پیدا کند و بر روی جدول توافقی به کار بسته شود در جایی که خواص رده بندی بیش از دو مقدار داشته باشد.مثال بعدی نشان می دهد که چگونه روال توضیح داده شده قبلی می تواند بدون تغییر برای جدول توافقی ۳x۳ به کار بسته شود. مقدار اولیه جدول را در جدول الف - ۵.۷ با جدول ب- ۵.۷ که شامل مقادیر تخمین زده شده است را مقایسه می کنیم و آزمون تطابق به صورت X^2= ۳.۲۲ محاسبه می شود در این حالت درجه آزادی به صورت زیر محاسبه می شود :

$$d.f = (n-1)(m-1) = (3-1)(3-1) = 4$$

		Attribute 1			
		Low	**Med.**	**High**	**Totals**
	Excell.	21	11	4	36
Attribute2	Good	3	2	2	7
	Poor	7	1	1	9
	Total	31	14	7	52
a) A 3 × 3 contingency table of observed values					
		Attribute 1			
		Low	Med.	High	Totals
	Excell.	21.5	9.7	4.8	36
Attribute2	Good	4.2	1.9	0.9	7
	Poor	5.4	2.4	1.2	9
	Total	31	14	7	52
b) A 3 × 3 Contingency table of expected values under H_0					

جدول ۵-۷ جدول توافقی برای خواص رده بندی با ۳ مقدار

ما بایستی در توصیف استنتاج های اضافی و آنالیزهای بیشتر مجموعه داده خیلی دقیق باشیم. کاملاً آشکار است که اندازه نمونه ها بزرگ نبوده و تعداد مشاهدات در بعضی خانه های جدول کوچک است. در اینجا با توجه به اندازه نمونه ، یک مشکل جدی وجود دارد و بنابراین برای تحلیل آماری اضافی نیاز است که بررسی شود که آیا نمونه مورد نظر یک نماینده خوبی از کل جمعیت هست یا نه؟ ما در اینجا تحلیل فوق را بررسی نخواهیم کرد زیرا در بیشتر مسائل دنیای واقعی داده کاوی مجموعه داده ها آنقدر بزرگ است که احتمال رویداد این کمبودها را حذف کنیم.

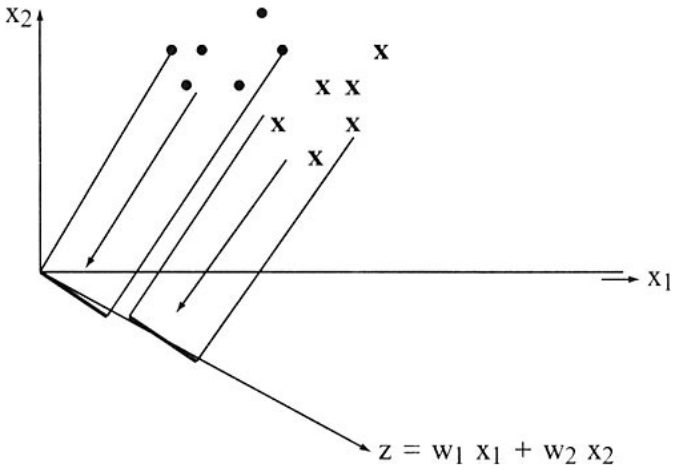
تا اینجا مراحل تولید جداول توافقی با داده های رده بندی را ارائه کردیم. روش تحلیل جداول توافقی سه یا چند بعدی در کتابهای پیشرفته آماری توضیح داده می شود. که روال ارتباط بین چندین خاصیت را که به طور همزمان تحلیل شده اند، ارائه گردیده است.

۵-۸- تحلیل ممیز خطی (LDA)

تحلیل ممیز خطی[۲] با مسائل دسته بندی متغیرهای وابسته رده بندی مرتبط است، به نحوی که متغیرهای وابسته به دسته های اسمی و ترتیبی و متغیرهای غیر وابسته به دسته های عددی تقسیم بندی می شوند. هدف LDA ایجاد یک تابع ممیزی می باشد که برای دسته های پاسخ (خروجی) متفاوت ، امتیازات متفاوتی را نشان می دهد. یک تابع ممیز خطی به شکل زیر است :

$$Z = X_1W_1 + X_2W_2 + \ldots + X_k W_k$$

به نحوی که X_1 , X_2 , ... و X_k متغیرهای مستقل باشند. کمیت Z امتیاز ممیزی نامیده می شود. W_1و W_2 , ... , W_k وزن نامیده می شوند. امتیاز ممیزی روی خط با مجموعه پارامترهای وزن مشخص شده را در شکل ۵.۳ مشاهده می نمایید.



شکل ۳-۵ تفسیر هندسی از امتیازات ممیزی

اساس تابع ممیز کننده Z این است که نسبت واریانس بین درون کلاس ها و بین کلاس ها را در یک مجموعه داده ای که از قبل دسته بندی شده اند را افزایش دهد. تابع ممیزی Z برای پیش بینی یک کلاس برای نمونه جدید غیر دسته بندی شده مورد استفاده می گیرد. برش امتیازات به عنوان یک معیار در مقابل امتیازات ممیزی تقسیم شده به کار می رود. انتخاب برش های امتیازات بستگی به توزیع نمونه ها در کلاس ها دارد. اجازه دهیدZ_a و Z_b را میانگین امتیازات ممیزی نمونه های پیش دسته بندی شده از کلاس A , B در نظر بگیریم. انتخاب بهینه برای برش $Z_{cut\text{-}ab}$ به شکل زیر است :

$$Z_{cut\text{-}ab} = (Za + Zb) / 2$$

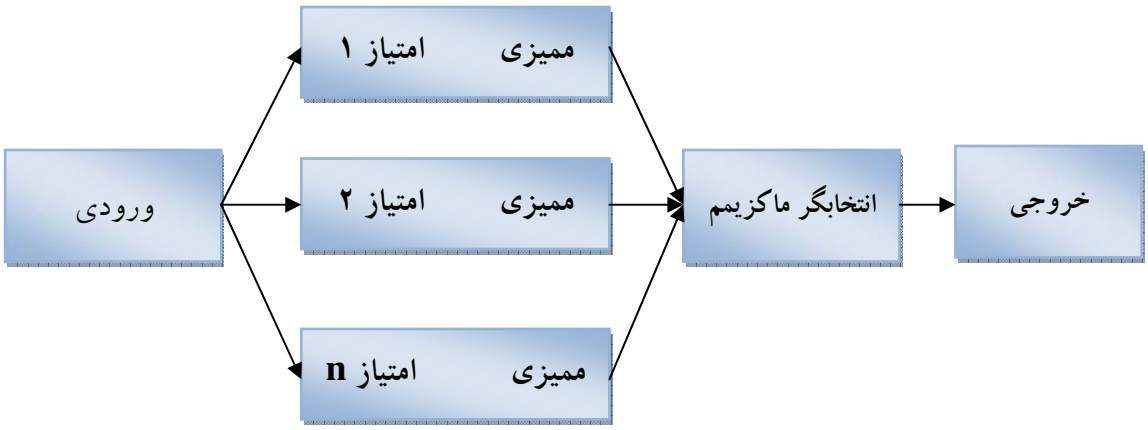
هنگامی که دو دسته یا کلاس نمونه دارای اندازه مساوی و با واریانس یک شکل توزیع شده اند. یک نمونه جدید طوری دسته بندی خواهد شد که امتیازات یا به صورت $Z > Z_{cut\text{-}ab}$ و یا $< Z_{cut\text{-}ab}$

[2] Linear Discriminant Analysis

Z خواهد بود. زمانی میانگین وزن برای امتیازات ممیزی در یک برش امتیاز بهینه استفاده می شود که سایر دسته ها هم اندازه نباشند:

$$Z_{cut-ab} = (n_a.Z_a + n_a.Z_b) / (n_a + n_b)$$

کمیت های n_a و n_b نشان دهنده تعداد نمونه ها در هر کلاس هستند. اگرچه یک تابع ممیزی منفرد با چند برش ممیزی می تواند نمونه ها را به چند دسته چندتایی تفکیک کند، اما تحلیل ممیزی چندگانه برای مسائل پیچیده تر استفاده می شود. تحلیل ممیزی چندگانه در هر حالت هنگامی قابل استفاده است که منابع ممیزی برای هر دسته به صورت مجزا ایجاد شده باشد. " به بالاترین امتیاز ممیزی توجه شود" این قانون از شکل ارائه شده استنباط می شود که این قانون در شکل ۴-۵ شرح داده شده است.



شکل ۴- ۵ فرآیند رده بندی در تحلیل ممیزی چندگانه