بسم الله الرحمن الرحيم

دانشکده ریاضی و رایانه

گروه آمار

پایان نامه تحصیلی برای دریافت درجه کارشناسی ارشد

رشته آمار ریاضی

ماشین بردار پشتیبان و رگرسیون ماشین بردار پشتیبان

با داده های فازی

استاد راهنما :

دکتر علیرضا عربپور

استاد مشاور:

دکتر محمد علی یعقوبی

مؤلف :

محبوبه درانی

شهریورماه ۱۳۹۰

تقدیم به

پشتیبان و یار همیشگی ام مادر مهربانم و همسر عزیزم که مشوقم بوده .

تشکر و قدردانی

خدای مهربان را شاکرم که لطفش بی اندازه و نعمتش گسترده است.

از استاد ارجمند جناب آقای دکتر عربپور که راهنمایی این پایان نامه را بر عهده داشتند، کمال تشکر را دارم و به خاطر تمام زحمات ایشان در سالهای تحصیلم سپاسگذارم.

از استاد گرامی جناب آقای دکتر یعقوبی که زحمت مشاوره این پایان نامه را تقبل کردند، کمال تشکر را دارم.

از استادان محترم جناب آقای دکتر امیرزاده و جناب آقای دکتر جمالیزاده که داوری این پایان نامه را بر عهده داشتند، تشکر می کنم.

چکیده

یکی از روش های دسته بندی داده ها ماشین بردار پشتیبان است، همچنین ممکن است داده های موجود مبهم باشند که به عنوان مجموعه های فازی توصیف می شوند. در این پایان نامه ابتدا ماشین بردار پشتیبان معرفی شده است سپس مفهوم مجموعه های فازی را وارد ماشین بردار پشتیبان و رگرسیون ماشین بردار پشتیبان کرده ایم. همچنین ضرایب مدل رگرسیون ماشین بردار پشتیبان فازی برآورد شده است.

در فصل اول این پایان نامه مفاهیم وتعاریف مقدماتی مورد نیاز آورده شده است. در فصل دوم روش ماشین بردار پشتیبان و رگرسیون ماشین بردار پشتیبان معرفی شده است. در فصل سوم مروری بر تاریخچه رگرسیون فازی داریم و به بیان روش های برنامه ریزی ریاضی می پردازیم. در فصل چهارم ماشین بردار پشتیبان فازی و مدل رگرسیون ماشین بردار پشتیبان فازی شرح داده شده است و در نهایت در فصل پنجم نتایج عددی بیان شده است.

کلمات کلیدی : ماشین بردار پشتیبان، رگرسیون ماشین بردار پشتیبان، رگرسیون فازی، ماشین بردار پشتیبان فازی.

فهرست مطالب **صفحه**

فصل اول : مفاهیم و تعاریف مقدماتی

۱ – ۱ مجموعه های فازی ۲

۱ – ۱ – ۱ مقدمه ۲

۱ – ۱ – ۲ تعاریف اولیه ۲

۱ – ۱ – ۳ اعداد فازی ۳

۱ – ۱ – ۴ عملگرهای جبری ۴

۱ – ۲ رگرسیون کلاسیک ۵

۱ – ۲ – ۱ مقدمه ۵

۱ – ۲ – ۲ رگرسیون خطی ساده ۵

۱ – ۲ – ۳ رگرسیون خطی چند گانه ۹

۱ – ۳ تحلیل ممیزی ۱۲

۱ – ۳ – ۱ جدا سازی و رده بندی برای دو جامعه : روش فیشر..... ۱۲

۱ – ۴ تعاریف ریاضی ۱۶

فصل دوم : ماشین بردار پشتیبان و رگرسیون ماشین بردار پشتیبان

۲ – ۱ مقدمه ۲۱

۲ – ۲ ماشین بردار پشتیبان (SVM) ۲۲

۲ – ۳ رگرسیون ماشین بردار پشتیبان ۲۸

۲ – ۴ رگرسیون ماشین بردار پشتیبان با انواع توابع زیان ۳۳

۲ – ۴ – ۱ رگرسیون خطی با تابع زیان اپسیلون–غیر حساس ۳۳

۲ – ۴ – ۲ رگرسیون خطی با تابع زیان درجه دو ۳۴

۲ – ۴ – ۳ رگرسیون خطی با تابع زیان هیوبر ۳۵

۲ – ۴ – ۴ رگرسیون غیر خطی با تابع زیان اپسیلون–غیر حساس ۳۵

فصل سوم : رگرسیون فازی

۳ – ۱ مقدمه ۳۹

۳ – ۲ تاریخچه ۴۱

۳ – ۳ روش های برنامه ریزی ریاضی ۴۳

۳ – ۳ – ۱ مدل تاناکا و واتادا ۴۴

۳ – ۳ – ۲ مدل حجتی و همکاران ۴۶

۳ – ۳ – ۳ مدل حسن پور و همکاران ۴۸

فصل چهارم : ماشین بردار پشتیبان فازی و رگرسیون ماشین بردار پشتیبان فازی

۴ – ۱ مقدمه ۵۲

۴ – ۲ ماشین بردار پشتیبان فازی(FSVM) ۵۲

۴ – ۳ رگرسیون خطی SVM فازی ۵۶

۴ – ۴ رگرسیون خطی SVM فازی با ورودی فازی- خروجی فازی ۶۰

۴ – ۵ رگرسیون خطی SVM فازی با ورودی غیر فازی- خروجی فازی ۶۴

۴ – ۶ رگرسیون غیر خطی SVM فازی ۶۸

فصل پنجم : نتایج عددی ۷۲

نتیجه گیری و پیشنهادات ۸۱

ضمیمه : برنامه های R ۸۲

منابع ۱۱۶

فصل اول

تعاریف و مفاهیم مقدماتی

۱- ۱مجموعه های فازی

۱ - ۱- ۱ مقدمه

نظریه ی مجموعه های فازی برای نخستین بار درسال۱۹۶۵ توسط دکترلطفی عسکرزاده [۶۵] دانشمندایرانی تبار واستاد دانشگاه برکلی به دنیای علم عرضه شد. محققین زیادی از آن تاریخ درتوسعه این مفهوم تلاش کرده وآن رادرزمینه های مختلف بکاربسته اند.

در این بخش مفاهیم وتعاریف اولیه مجموعه های فازی ازقبیل:مفهوم مجموعه های فازی، اعداد فازی و... را بیان خواهیم نمود، که در فصل های بعدی به کار گرفته می شود.

۱-۱-۲ تعاریف اولیه

مفهوم مجموعه و نظریه ی مجموعه ها ابزار های قوی در ریاضیات هستند. درنظریه ی مجموعه ـ ها، مجموعه به صورت گردایه ای از اشیاء کاملا مشخص تعریف می شود و عضویت یا عدم عضویت یک شی درمجموعه قطعی است، در بسیاری از مسائل واقعی حدود مجموعه ها کاملا مشخص نیست برای مجموعه هایی که حدودشان کاملا مشخص نیست مفهوم مجموعه فازی و عضویت جزئی مطرح می شود [۲].

تعریف۱-۱-۲-۱: فرض کنیدX یک مجموعه ناتهی باشد. هر زیر مجموعه فازی مانند $\tilde{A}$ از X توسط یک تابع$\mu_{\tilde{A}}: X \to [۰,۱]$ به عنوان تابع عضویت مشخص می شود که در آن برای هرx درX، مقدار $\mu_{\tilde{A}}(x)$ میزان عضویت x در آن زیر مجموعه فازی را نشان می دهد[۶].

به عنوان مثال $\mu_{\tilde{A}}(x) = ۰$ یعنی x قطعاً متعلق به A نیست و $\mu_{\tilde{A}}(x) = ۱$ یعنی x بدون شک متعلق به A است.

اگر برای هر $x\epsilon X$ ، $\mu_{\tilde{A}}(x) = ۰$ یا $\mu_{\tilde{A}}(x) = ۱$ در اینصورت $\tilde{A}$ یک مجموعه غیرفازی است و تابع عضویت به تابع نشانگر مجموعه A تبدیل می گردد.

تعریف۱- ۱ - ۲-۲ : تکیه گاه[۱] مجموعه ی فازی $\tilde{A}$ از X که با نماد $S_{\tilde{A}}$ نشان می دهیم به صورت زیر تعریف می شود:

$$S_{\tilde{A}} = \{x\epsilon X | \mu_{\tilde{A}}(x) > ۰\}.$$

تعریف۱- ۱- ۲-۳ : برای هر[۱و۰] $\alpha \epsilon$ ، α- برش[۲] یک مجموعه ی فازی $\tilde{A}$ از X که با نماد $\tilde{A}_\alpha$ نمایش داده می شود به صورت زیر تعریف می گردد:

$$\tilde{A}_\alpha = \begin{cases} \{x\epsilon X | \mu_{\tilde{A}}(x) \geq \alpha\}, & \forall \alpha > ۰, \\ \{x\epsilon X | \mu_{\tilde{A}}(x) > ۰\}, & \alpha = ۰. \end{cases}$$

۱- ۱-۳ اعداد فازی

تعریف۱- ۱-۳- ۱ : کمیت فازی $\tilde{A}$ را عدد فازی L - R گوییم هر گاه تابع عضویت $\tilde{A}$ به صورت زیر باشد :

$$\mu_{\tilde{A}}(x) = \begin{cases} L\left(\frac{m-x}{\alpha}\right), & x \leq m, (\alpha > ۰), \\ R\left(\frac{x-m}{\beta}\right), & x \geq m, (\beta > ۰). \end{cases}$$

---

[۱] support
[۲] α-level

که Lو$R : [۰, \infty) \mapsto [۰, ۱]$ توابعی پیوسته، نزولی و روی بازه ی [۰, ۱] معکوس پذیر هستند. $R(۰)=L(۰)=۱$ و $R(۱)=L(۱)=۰$، به توابع L و R توابع مرجع می گوییم. همچنین به m، α و β به ترتیب مرکز، پهنای چپ و پهنای راست عدد فازی $\tilde{A}$ می گوییم. برای سادگی عدد فازی $\tilde{A}$ را با نماد $\tilde{A} = (m, \alpha, \beta)_{LR}$ نشان می دهیم. اگر $L=R$ و $\alpha=\beta$ ، $\tilde{A}$ را عدد فازی متقارن می گوییم و آن را با نماد $\tilde{A} = (m, \alpha)_{L}$ نشان می دهیم.

تعریف۱- ۱- ۳- ۲ : به عدد فازی $\tilde{A}$ ، $L - R$ با توابع مرجع $L(x) = R(x) = \max(۰, ۱-x)$ عدد فازی مثلثی می گوییم و آن را با $\tilde{A} = (m,\alpha,\beta)$ نشان می دهیم [۶].

۱- ۱- ۴ عملگر های جبری

اگر $\tilde{A}_۱ = (m_۱, \alpha_۱, \beta_۱)$ و $\tilde{A}_۲ = (m_۲, \alpha_۲, \beta_۲)$ دو عدد فازی مثلثی باشند آنگاه :

$$۱)\quad k\tilde{A}_۱ = \begin{cases} (km_۱, k\alpha_۱, k\beta_۱), & k > ۰, \\ (km_۱, |k|\beta_۱, |k|\alpha_۱), & k < ۰. \end{cases}$$

$$۲)\quad \tilde{A}_۱ + \tilde{A}_۲ = (m_۱ + m_۲, \alpha_۱ + \alpha_۲, \beta_۱ + \beta_۲),$$

$$۳)\quad \tilde{A}_۱ - \tilde{A}_۲ = \tilde{A}_۱ + (-\tilde{A}_۲) = (m_۱ - m_۲, \alpha_۱ + \beta_۲, \alpha_۲ + \beta_۱).$$

در ادامه به مفاهیم مقدماتی رگرسیون کلاسیک می پردازیم که در فصل های بعدی مورد استفاده قرار می گیرد.

۱-۲ رگرسیون کلاسیک

۱- ۲- ۱مقدمه

بسیاری از پدیده ها و پیشامدها یی که در جهان اتفاق می افتند تابع برخی از پیشامدهای دیگر می باشند. یکی از هدفهای مهم در تحقیقات علمی،کشف وجود یا عدم وجود رابطه بین پدیده ـ هاست. یکی از روشهایی که به طور گسترده برای این منظور به کار می رود روش تحلیل رگرسیون می باشد [۸]. در واقع رگرسیون روشی است برای بررسی رابطه ی بین دو یا چند متغیر به طوری که یک متغیر از روی دیگری یا بقیه می تواند پیش بینی شود.

۱-۲-۲ رگرسیون خطی ساده

در ساده ترین حالت در یک مدل رگرسیون فرض می کنیم y متغیر وابسته و x متغیر مستقل باشد و این رابطه به صورت زیر باشد:

$$y = \beta_0 + \beta_1 x . \qquad (۱-۱)$$

با توجه به این که نقاط دقیقاً روی خط راست در یک امتداد قرار نمی گیرند رابطه ی (۱-۱) باید تعدیل شود.

بنابراین مدل (۱-۱) به این صورت اصلاح می گردد.

$$y = \beta_0 + \beta_1 x + \varepsilon. \qquad (۱-۲)$$

که ε اختلاف بین مقدار y (متغیر وابسته) وخط $\beta_0 + \beta_1 x$ را بیان می کند و مقدار خطای آماری نامیده می شود. مدل اصلاح شده ی (۱-۲) مدل رگرسیون خطی ساده نامیده می شود.

اعتبار یک مدل رگرسیونی به برقرار بودن شرایط زیر بستگی دارد:

۱- میانگین خطاها برابرصفرو واریانس خطاها مقدار ثابتی باشد.

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2, \quad i = 1,2, \dots, n.$$

۲- خطا ها نسبت به هم ناهمبسته باشند.

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \qquad i,j = 1,2, \dots, n, \qquad i \neq j.$$

۳- خطاها دارای توزیع نرمال با میانگین صفرو واریانس ثابت باشند.

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1,2, \dots, n.$$

با استفاده از روش حداقل مربعات خطا می توان پارامترهای مدل یعنی β_1 و β_0 را برآوردکرد[۸].

فرض می کنیم داده ها به صورت زوجی (x_1, y_1) و (x_2, y_2) و ... و (x_n, y_n) وجود دارند، خطای هرمشاهده به این صورت است :

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1,2, \dots, n.$$

در این روش پارامترها را به گونه ای برآورد می کنیم که مجموع مربعات خطا مینیمم شود.

به عبارتی تابع زیر را مینیمم می کنیم :

$$S\left(\beta_{۰} و \beta_{۱}\right) = \sum_{i=۱}^{n} (y_i - \beta_{۰} - \beta_{۱} x_i)^{۲}.$$

در صورتی که $\hat{\beta}_{۰}$ و $\hat{\beta}_{۱}$ برآورد پارامترهای $\beta_{۰}$ و $\beta_{۱}$ به روش کمترین مربعات خطا باشند آنگاه باید داشته باشیم :

$$\frac{\partial S(\beta_{۰} و \beta_{۱})}{\partial \beta_{۰}} |_{\hat{\beta}_{۰} و \hat{\beta}_{۱}} = -۲ \sum_{i=۱}^{n} (y_i - \hat{\beta}_{۰} - \hat{\beta}_{۱} x_i) = ۰,$$

$$\frac{\partial S(\beta_{۰} و \beta_{۱})}{\partial \beta_{۱}} |_{\hat{\beta}_{۰} و \hat{\beta}_{۱}} = -۲ \sum_{i=۱}^{n} x_i (y_i - \hat{\beta}_{۰} - \hat{\beta}_{۱} x_i) = ۰.$$

با ساده کردن روابط بالا داریم:

$$n\hat{\beta}_{۰} + \hat{\beta}_{۱} \sum_{i=۱}^{n} x_i = \sum_{i=۱}^{n} y_i ,$$

$$\hat{\beta}_{۰} \sum_{i=۱}^{n} x_i + \hat{\beta}_{۱} \sum_{i=۱}^{n} x_i^{۲} = \sum_{i=۱}^{n} x_i y_i .$$

با حل معادلات فوق که معادلات نرمال نامیده می شوند برآورد پارامترها به این صورت به دست می آیند:

$$\hat{\beta}_{۰} = \bar{y} - \hat{\beta}_{۱} \bar{x}, \qquad \hat{\beta}_{۱} = \frac{S_{xy}}{S_{xx}}.$$

که

$$S_{xy} = \sum_{i=۱}^{n} x_i y_i - \frac{۱}{n}\left(\sum_{i=۱}^{n} x_i\right)\left(\sum_{i=۱}^{n} y_i\right),$$

$$S_{xx} = \sum_{i=۱}^{n} x_i^{۲} - \frac{۱}{n}\left(\sum_{i=۱}^{n} x_i\right)^{۲},$$

$$\bar{y} = \frac{۱}{n}\left(\sum_{i=۱}^{n} y_i\right),$$

$$\bar{x} = \frac{۱}{n}\left(\sum_{i=۱}^{n} x_i\right).$$

پس از برآورد پارامترها، مدل رگرسیونی به صورت زیر در می آید :

$$\hat{y} = \hat{\beta}_{.} + \hat{\beta}_{۱} x.$$

۱-۲-۳ رگرسیون خطی چندگانه

در رگرسیون خطی چندگانه می خواهیم متغیر وابسته y را براساس متغیرهای مستقل $x_1, x_2, \dots, x_k$ پیش بینی کنیم، برای این منظور مدل خطی زیر پیشنهاد می شود:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon .$$

که همانند رگرسیون خطی ساده، ε خطای آماری نامیده می شود و اختلاف بین y و رابطه ی خطی $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ می باشد.

حال بر اساس یک نمونه ی n تایی $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$ $i=1,2,\dots,n$ پارامترهای مدل یعنی $\beta_0, \beta_1, \dots, \beta_k$ را به روش کمترین مربعات خطا برآورد می کنیم. برای این منظور باید تابع زیررا مینیمم کنیم.

$$S\left(\beta_0 و \beta_1 و \dots و \beta_k\right) = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 .$$

درصورتی که $\beta_0, \beta_1, \dots, \beta_k$ برآورد کمترین مربعات خطا ی پارامترها باشند آن گاه باید داشته باشیم:

$$\frac{\partial S(\beta_0 و \dots و \beta_k)}{\partial \beta_0} \Big|_{\hat{\beta}_0 و \hat{\beta}_1 و \dots و \hat{\beta}_k} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0,$$

$$\frac{\partial S(\beta_0 و \dots و \beta_k)}{\partial \beta_1} \Big|_{\hat{\beta}_0 و \hat{\beta}_1 و \dots و \hat{\beta}_k} = -2 \sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0,$$

⋮

$$\frac{\partial S(\beta_0 \text{ و } \ldots \text{ و } \beta_k)}{\partial \beta_k}\Big|_{\hat{\beta}_0 \text{ و } \hat{\beta}_1 \text{ و } \ldots \text{ و } \hat{\beta}_k} = -2\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0.$$

باساده کردن روابط فوق داریم

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i ,$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^{2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i ,$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{ik} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^{2} = \sum_{i=1}^{n} x_{ik}y_i .$$

با حل $p = k+1$ معادله ی فوق که معادلات نرمال نامیده می شوند $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ به دست می آیند.

برای سادگی کار وانجام محاسبات مدل رگرسیون چندگانه را به شکل دیگری نیز می توان نوشت.

تعریف می کنیم

$$\mathbf{y}_{n\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \quad p = k+1,$$

$$\mathbf{X}_{n\times p} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}.$$

با توجه به تعاریف بیان شده شکل ماتریسی مدل رگرسیون چندگانه را به صورت زیر می توان نوشت:

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \qquad \boldsymbol{\varepsilon}_{n\times 1}\sim N_n(\mathbf{0}_{n\times 1},\ \sigma^2 I).$$

$$\mathbf{y}_{n\times 1}\sim N_n\left(\mathbf{X\beta},\ \sigma^2\mathbf{I}\right) \qquad \text{بنابراین}$$

و در نتیجه برآورد کمترین مربعات خطا به صورت زیر به دست می آید.

$$S\left(\beta_0 \text{ و } \beta_1 \text{ و } \dots \text{ و } \beta_k\right) = \boldsymbol{\varepsilon}'\,\boldsymbol{\varepsilon} = (\mathbf{y}-\mathbf{X\beta})'\,(\mathbf{y}-\mathbf{X\beta})$$

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} = 0 \ \mapsto\ \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\,\hat{\boldsymbol{\beta}}$$

منظور از $\mathbf{X}'$ ترانهاده ی ماتریس $\mathbf{X}$ می باشد.

اگر $(\mathbf{X}'\mathbf{X})$ معکوس پذیر باشد برآورد کمترین مربعات $\boldsymbol{\beta}$ عبارت است از:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbf{y}.$$

بنابراین برآورد مدل رگرسیون به صورت $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ می باشد که

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{\beta}_0 + x_{11}\hat{\beta}_1 + \cdots + x_{1k}\hat{\beta}_k \\ \hat{\beta}_0 + x_{21}\hat{\beta}_1 + \cdots + x_{2k}\hat{\beta}_k \\ \vdots \\ \hat{\beta}_0 + x_{n1}\hat{\beta}_1 + \cdots + x_{nk}\hat{\beta}_k \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}.$$

در ادامه به تعریف روش تحلیل ممیزی یکی از روش های دسته بندی و رده بندی می پردازیم که در فصل های بعد مورد استفاده قرار خواهد گرفت.

۱ – ۳ تحلیل ممیزی

تحلیل ممیزی و رده بندی تکنیک های چند متغیری هستند که با جدا کردن مجموعه های متمایز مشاهدات و تخصیص دادن مشاهدات جدید به دسته های تعریف شده ی قبلی سروکار دارد. در تحلیل ممیزی تاکید روی به دست آوردن قاعده ای است که از آن بتوان برای تخصیص بهینه ی یک شیئی جدید به رده های مشخص استفاده کرد. اصطلاح ممیزی توسط فیشر[۳] [۳۰] در اولین کار پیشرفته ی مسائل جداسازی معرفی گردیده است . در زیر روش ممیزی پیشنهاد شده توسط فیشر بیان می شود.

۱ – ۳ – ۱ جداسازی و رده بندی برای دو جامعه (روش فیشر)

در این بخش دو رده (جامعه) را با π_1 و π_2 مشخص می کنیم. مشاهدات به طور معمول بر مبنای اندازه هایی مانند p متغیر تصادفی $\mathbf{X}' = (X_1, \cdots, X_p)$ ، جداسازی یا رده بندی می شوند. به عنوان مثال جامعه های π_1 و π_2 می تواند دو گونه علف هرز **و** متغیر$\mathbf{X}$ درازای کاسبرگ و گلبرگ، عمق شکاف گلبرگ و ... باشد.

فیشر در ابتدا ترکیبات خطی از $\mathbf{x}$ را برای ایجاد y ها پیشنهاد می کند، زیرا آن ها توابع ساده ای از $\mathbf{x}$ هستند و کار کردن با آن ها از نظر ریاضی آسان تر است. اگرμ_{1y} میانگین y های به دست آمده از $\mathbf{x}$ های متعلق به π_1 و μ_{2y} میانگین y های به دست آمده از $\mathbf{x}$ های متعلق به π_2 باشد، آن

[۳] Fisher