Machine Learning

روش‌های چندمتغیره

یادگیری ماشین
فصل پنجم

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
پاییز ۱۳۹۸
احمد محمودی ازناوه

فهرست مطالب

- داده‌های چندمتغیره
- تخمین پارامترها
  - کواریانس و ضریب همبستگی
- دسته‌بندی
- رگرسیون

داده‌های چندمتغیره

- در بسیاری از کاربردها، اندازه‌گیری‌های متفاوتی انجام می‌شود، از این جهت با بردار ورودی (ویژگی) سروکار خواهیم داشت (به عنوان مثال یک بردار d-بعدی).

Data matrix

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

یک نمونه

d inputs/features/attributes

N instances/observations/examples

پارامترهای چندمتغیره

$$\mathcal{M}ean\text{:}E[\boldsymbol{x}] = \mu = [\mu_1, \ldots, \mu_d]^T$$

$$\sigma_i \equiv var(X_i)$$

$$Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i \mu_j$$

$$\Sigma \equiv \mathrm{Cov}(X) = E[(X - \mu)(X - \mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

$$= E[XX^T] - \mu\mu^T$$

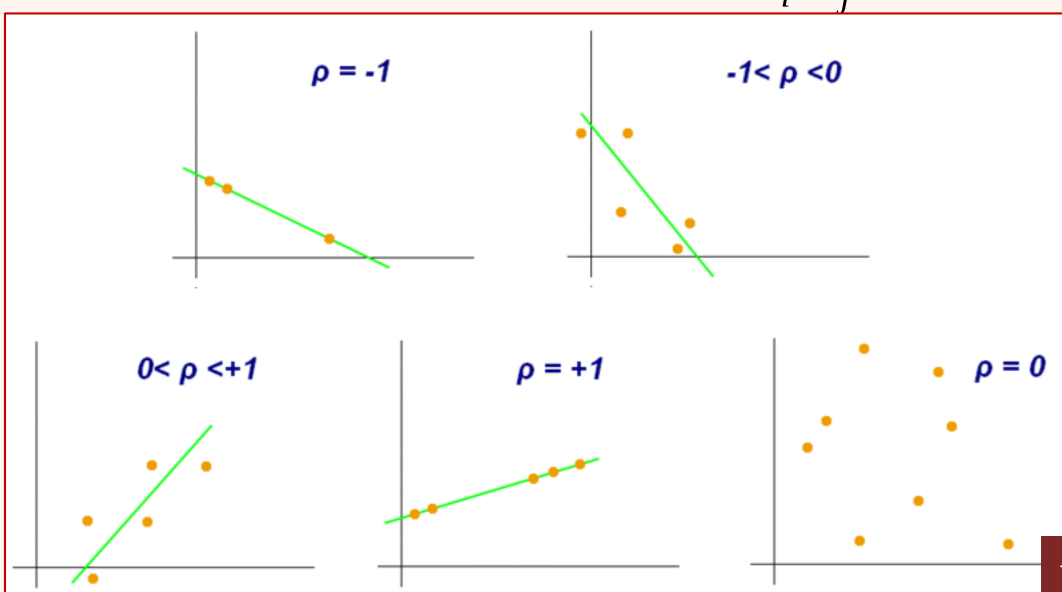
واریانس

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

ضریب همبستگی Pearson

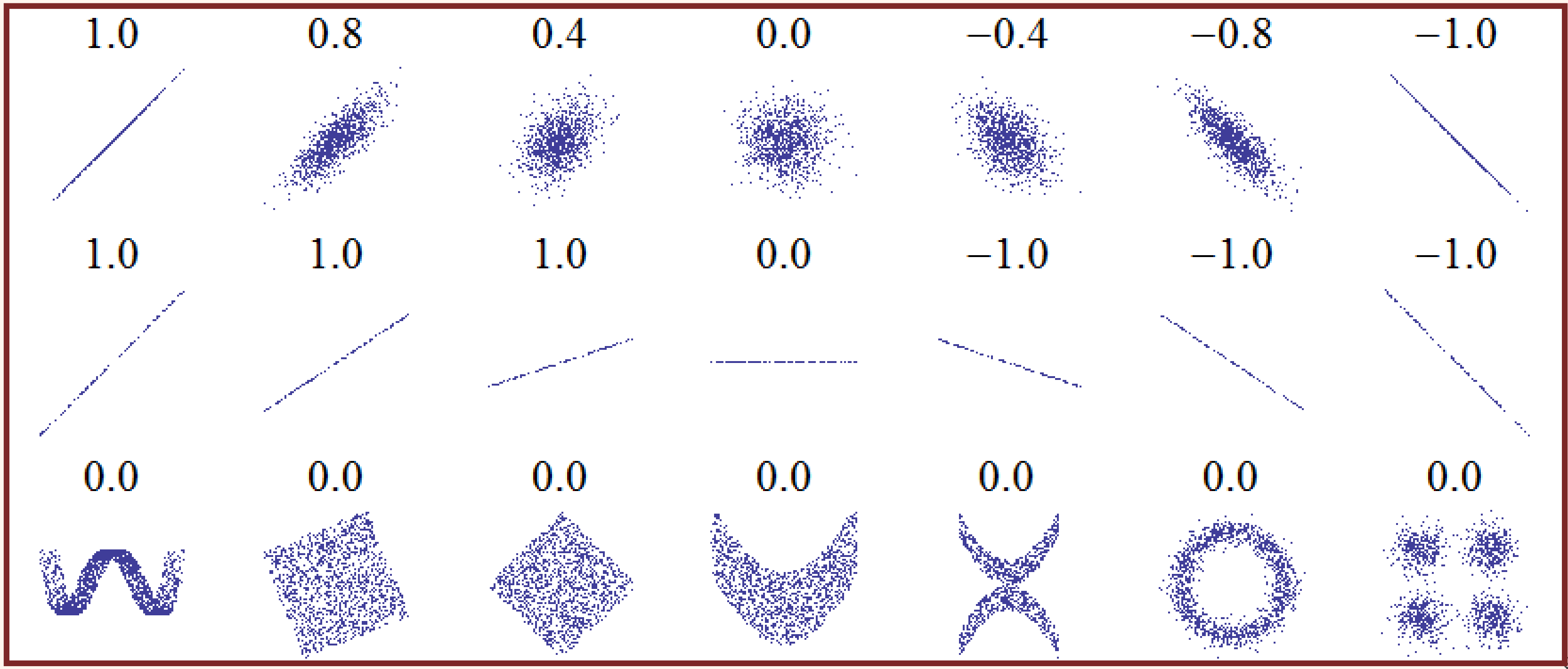
- میزان همبستگی خطی بین دو متغیر تصادفی را می‌سنجد.
  - مقدار این ضریب بین ۱- تا ۱ تغییر می‌کند که «۱» به معنای همبستگی مثبت کامل، «۰» به معنی نبود همبستگی، و «۱-» به معنی همبستگی منفی کامل است.
  - این ضریب که کاربرد فراوانی در آمار دارد، توسط Karl Pearson براساس ایده‌ی اولیه‌ی Francis Galton تدوین شد.

$$Correlation: Corr\left(X_i, X_j\right) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$



wikipedia

مثال



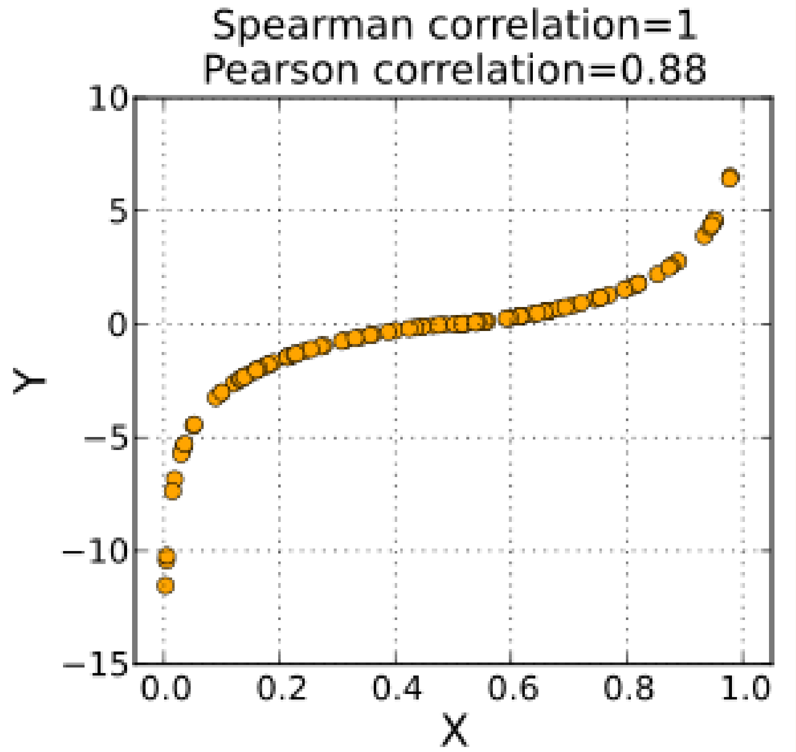
wikipedia

در صورتی که دو متغیر مستقل باشند، کواریانس و در نتیجه ضریب همبستگی آن دو صفر خواهد بود.

اما عکس این مسأله درست نیست.

ضریب همبستگی رتبه‌ای Spearman

Spearman's rank correlation coefficient



$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

تخمین پارامترها

Sample mean $\mathbf{m}: m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1, \ldots, d$

Covariance matrix $\mathbf{S}: s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$

Correlation matrix $\mathbf{R}: r_{ij} = \frac{s_{ij}}{s_i s_j}$

تخمین پارامترهای نامشخص

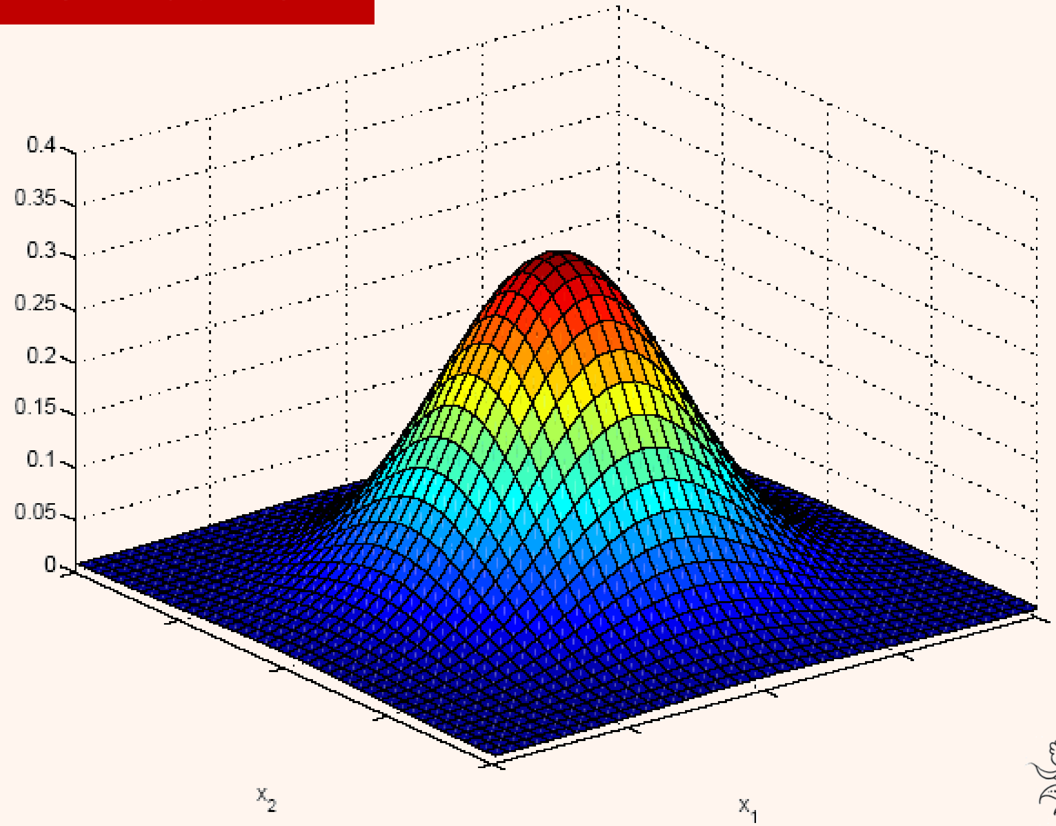
Estimation of Missing Values

- ممکن است در برخی نمونه‌ها، برخی متغیرها در اختیار نباشند.
  - بهترین راه صرفنظر کردن از آن‌هاست، اما این راه در حالتی که داده‌های آموزشی محدود باشد، کارایی ندارد.
  - یک فیلد جدید اضافه کنیم که فقدان مقدار را مشخص می‌کند؛ ممکن است دارای اطلاعات ارزشمندی باشد.
- نسبت دادن مقدار: (imputation)
  - جایگزینی مقدار میانگین (mean imputation)
  - انتساب با رگرسیون (imputation by regression)

توزیع نرمال چند متغیره

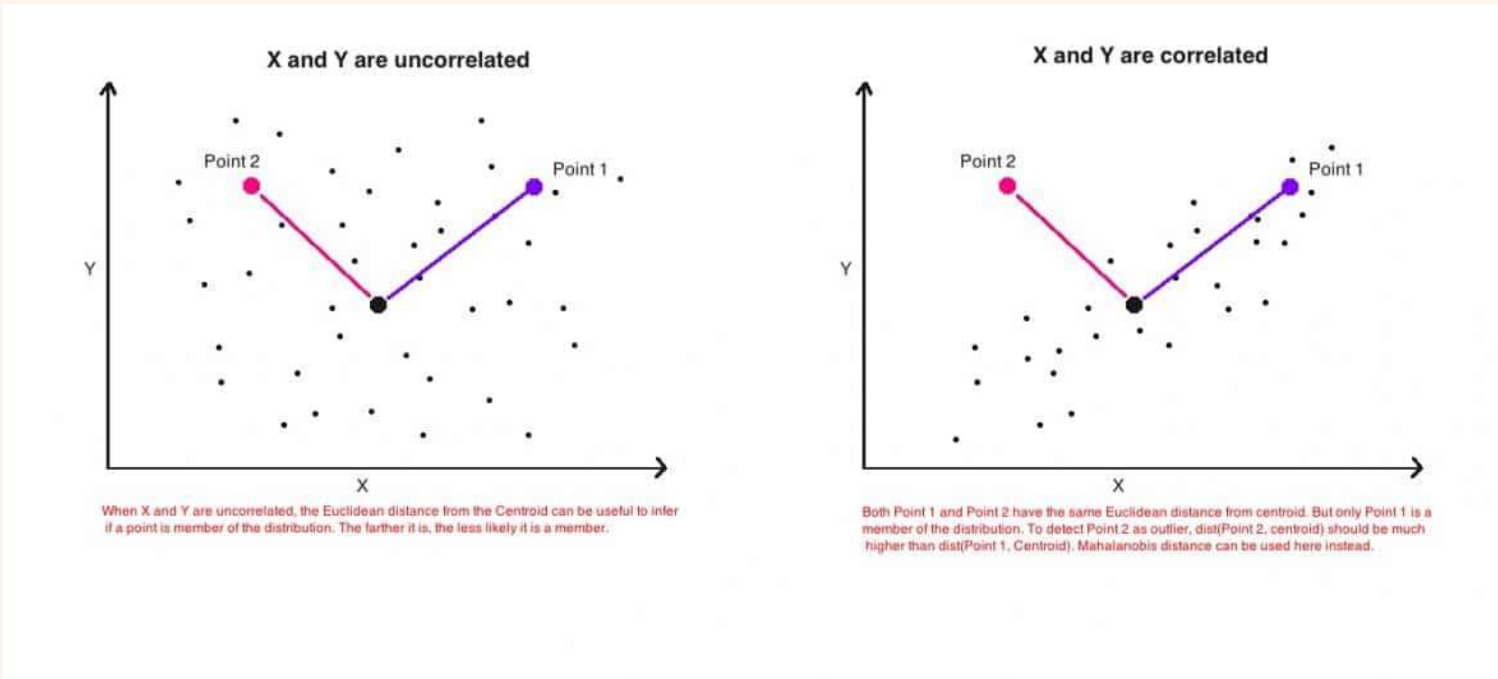
Multivariate Normal Distribution

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



$$\mathbf{x}\sim\mathcal{N}_d(\boldsymbol{\mu},\Sigma)$$

$$p(\mathbf{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$



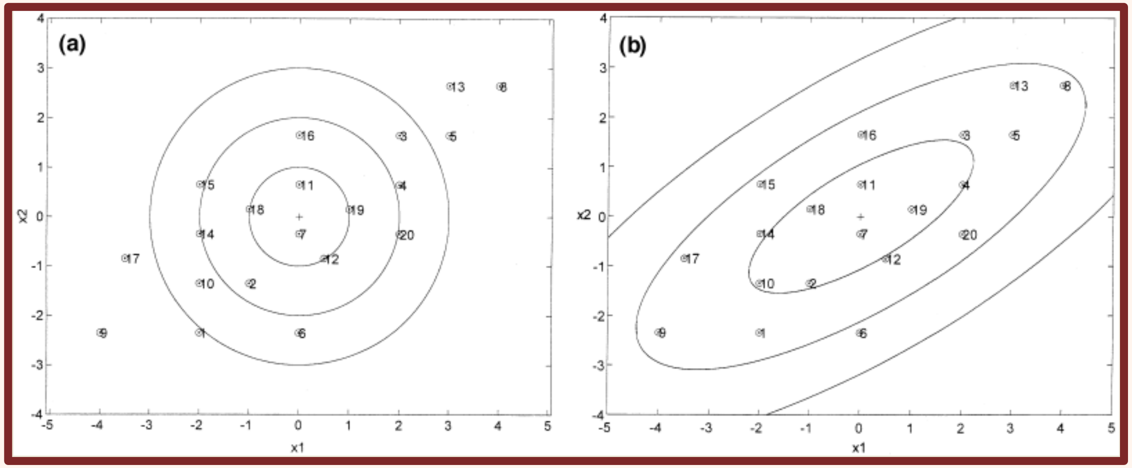
https://www.machinelearningplus.com/statistics/mahalanobis-distance/

توزیع نرمال چند متغیره (ادامه...)

Distance in standard units

- فاصله‌ی Mahalanobis:
  - معیاری برای اندازه‌گیری فاصله‌ی یک نقطه از یک توزیع داده است.

$$(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$$



- $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})=c^2$

ابربیضی

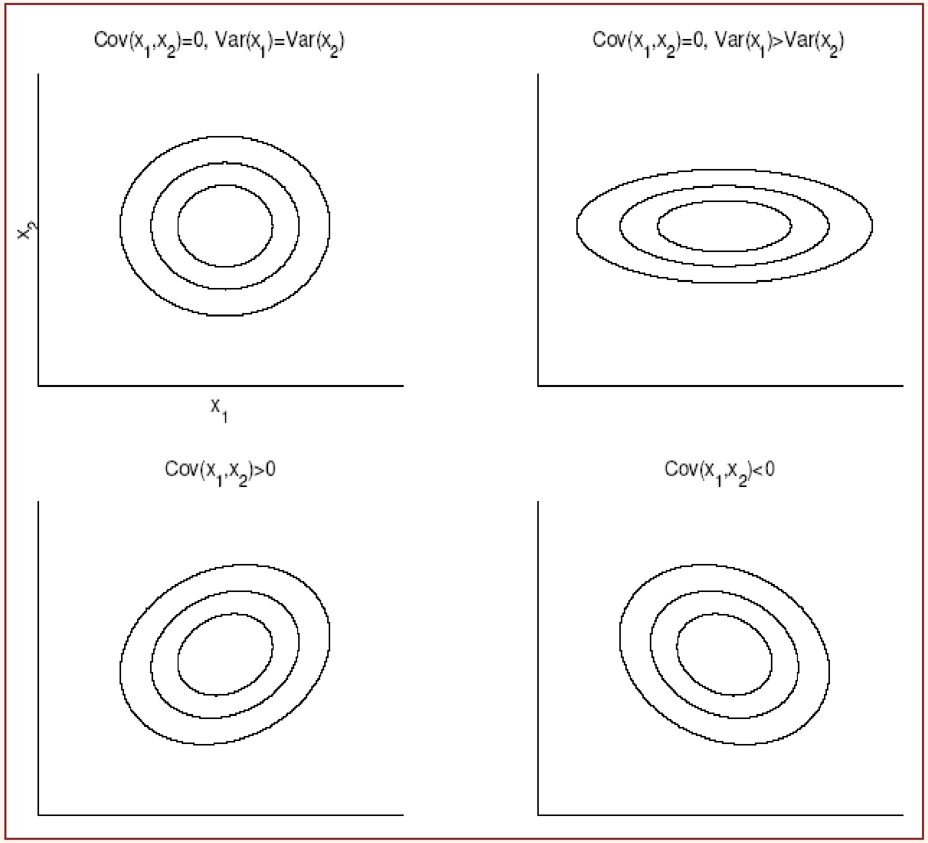
مثال- دو بعدی

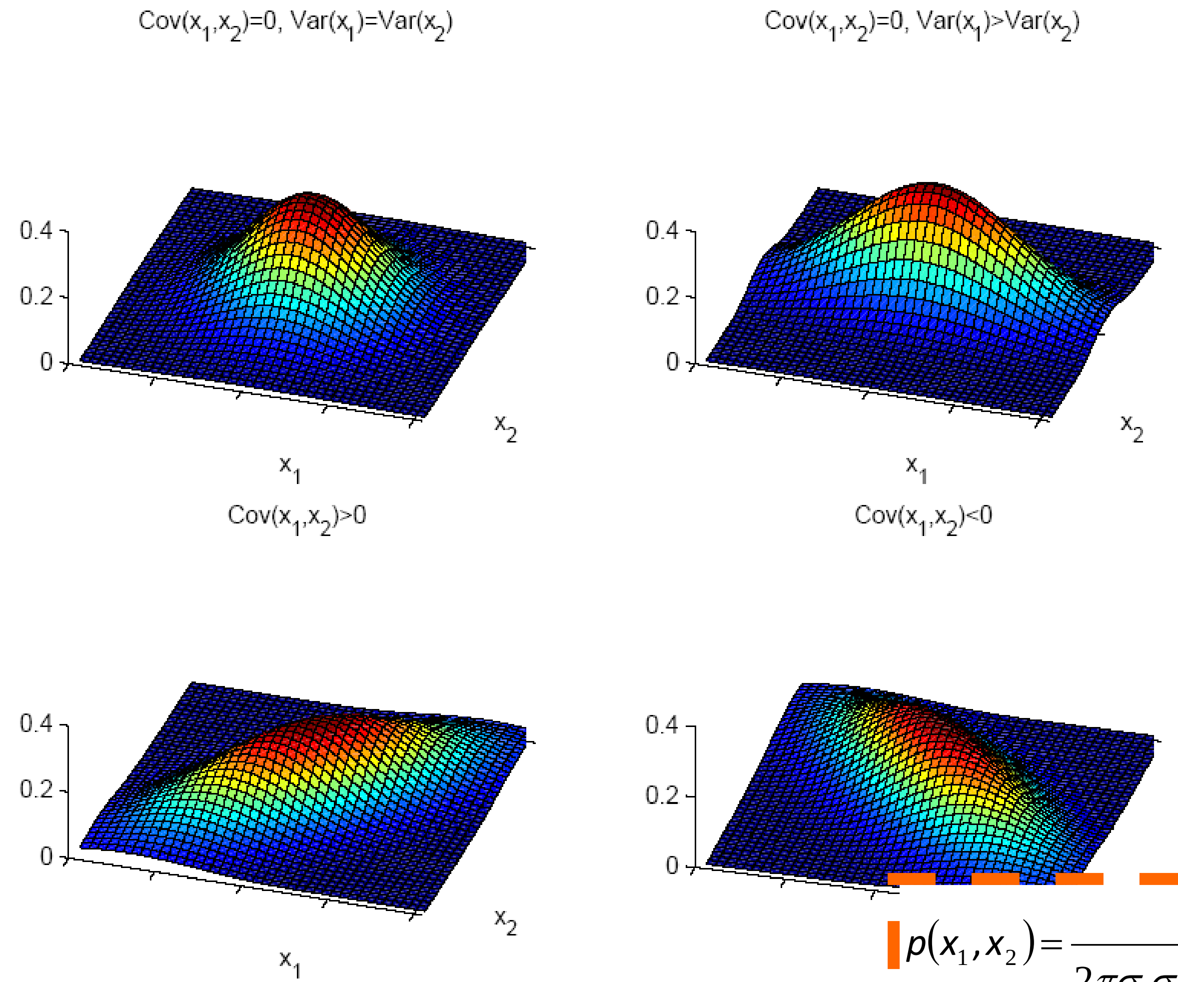
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(z_1^2 - 2\rho z_1 z_2 + z_2^2 \right) \right]$$

$$z_i = (x_i - \mu_i) / \sigma_i$$

z normalization





$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)\right]$$

$$z_i = (x_i - \mu_i)/\sigma_i$$

nonsingular

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right]$$

(var≠0, |ρ|<1)

positive definite

If not ->Dimension reduction

چندنکته

- در صورتی که x دارای توزیع نرمال (چندمتغیره) باشد، متغیر مربوط به هر بعد نیز دارای توزیع نرمال تک‌متغیره است. (عکس این مطلب درست نیست)
  - در واقع، هر زیر مجموعه‌ی kبعدی (k<d) نیز یک توزیع نرمال چندمتغیره است.
- در صورتی که متغیرها مستقل باشند:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{d/2}\prod_{i=1}^{d}\sigma_i}\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

ادامه ...

- در صورتی که

$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma), \quad \mathbf{w} \in \Re^d$$

$$\mathbf{w}^T\mathbf{x} = w_1x_1 + w_2x_2 + \cdots + w_dx_d \quad \sim \mathcal{N}\left(\mathbf{w}^T\boldsymbol{\mu}, \mathbf{w}^T\Sigma\mathbf{w}\right)$$

$$E\left[\mathbf{w}^T\mathbf{x}\right] = \mathbf{w}^T E[\mathbf{x}] = \mathbf{w}^T\mu$$

$$\begin{aligned}\mathrm{Var}\left(\mathbf{w}^T\mathbf{x}\right) &= E\left[\left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mu\right)^2\right] = E\left[\left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mu\right)\left(\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mu\right)\right] \\ &= E\left[\mathbf{w}^T(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\mathbf{w}\right] = \mathbf{w}^T E\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\right]\mathbf{w} \\ &= \mathbf{w}^T\Sigma\mathbf{w}\end{aligned}$$

در صورتی که k بردار در نظر گرفته شود (k<d):

$$\mathbf{w}^T\mathbf{x} \sim \mathcal{N}_k\left(\mathbf{w}^T\boldsymbol{\mu}, \mathbf{w}^T\Sigma\mathbf{w}\right)$$

$$\mathbf{w} \text{ is } d \times k$$

$$p\,(x \mid C_i) \sim N\,(\mu_i\,,\, \Sigma_i)$$

$$p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

Discriminant functions

$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid C_i) + \log P(C_i)$$
$$= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1} (\mathbf{x}-\mu_i) + \log P(C_i)$$

تخمین پارامترها

با تخمین و جایگزینی پارامترها خواهیم داشت:

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

$$\mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t \left(\mathbf{x}^t - \mathbf{m}_i\right)\left(\mathbf{x}^t - \mathbf{m}_i\right)^T}{\sum_t r_i^t}$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\log\left|\mathbf{S}_i\right| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log\hat{P}(C_i)$$

Quadratic discriminant

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}\left(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i\right) + \log\hat{P}(C_i)$$

$$= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

where

$$\mathbf{W}_i = -\frac{1}{2}\mathbf{S}_i^{-1}$$

$$\mathbf{w}_i = \mathbf{S}_i^{-1}\mathbf{m}_i$$

$$w_{i0} = -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{P}(C_i)$$

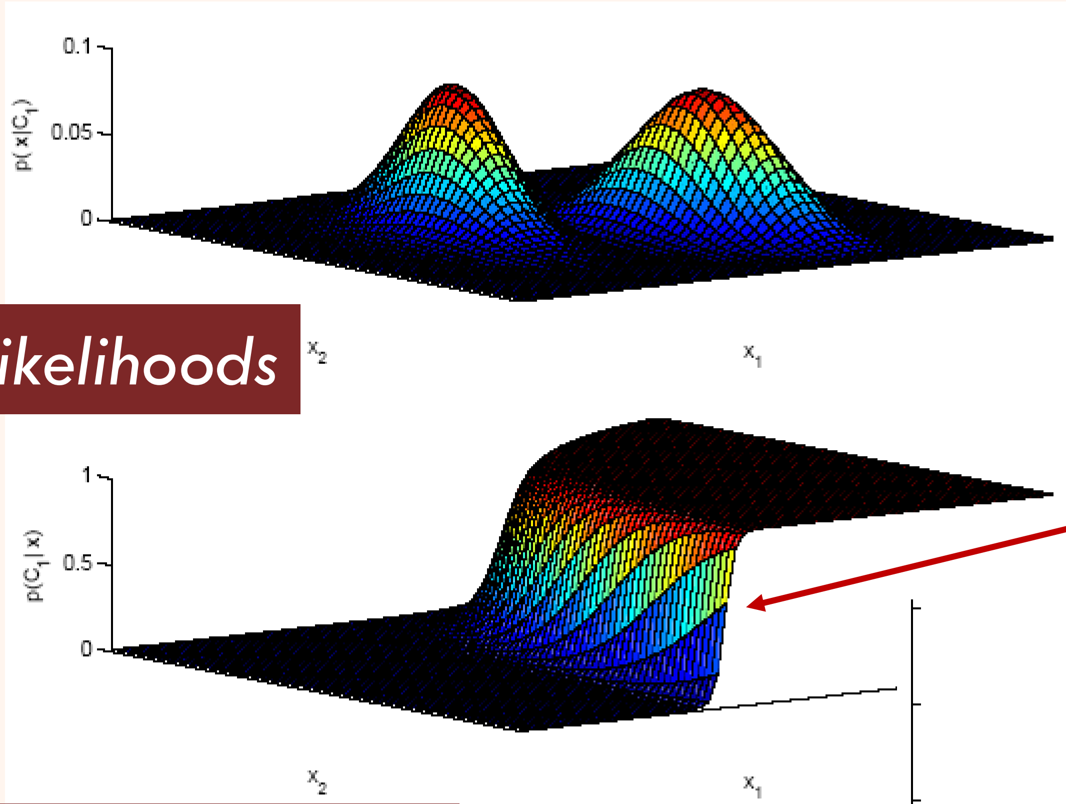
K.d for means

تعداد پارامترها

K.d.(d+1)/2 for covariance

در کم بودن تعداد نمونه‌های آموزشی ممکن است ماتریس singular شود.

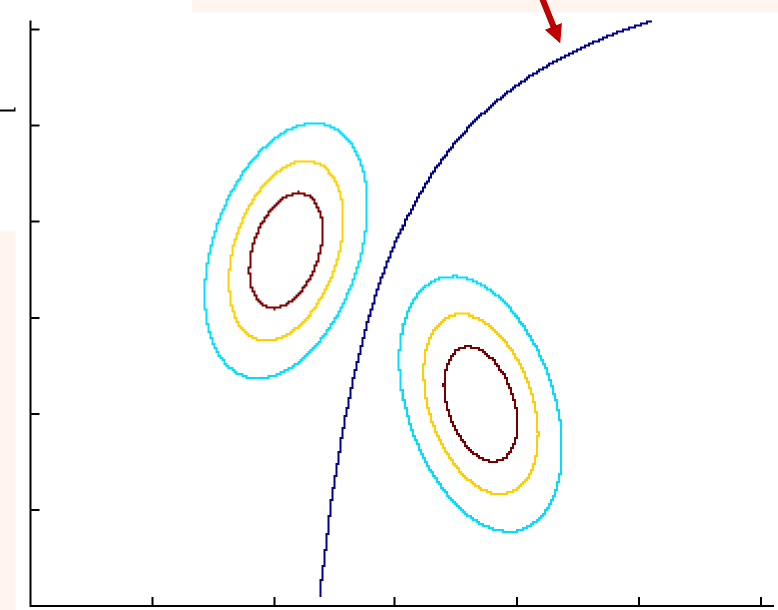
Quadratic discriminant



likelihoods

posterior for C_1

discriminant:
$P\ (C_1\,|\,x\ ) = 0.5$



Common Covariance Matrix S

- در صورت کم بودن تعداد نمونه‌های آموزشی می‌توان ماتریس کوواریانس را برای همه‌ی کلاس‌ها یکسان در نظر گرفت.

$$S = \sum_i \hat{P}(C_i) S_i$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}\left(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i\right) + \log\hat{P}(C_i)$$

- در نتیجه خواهیم داشت:

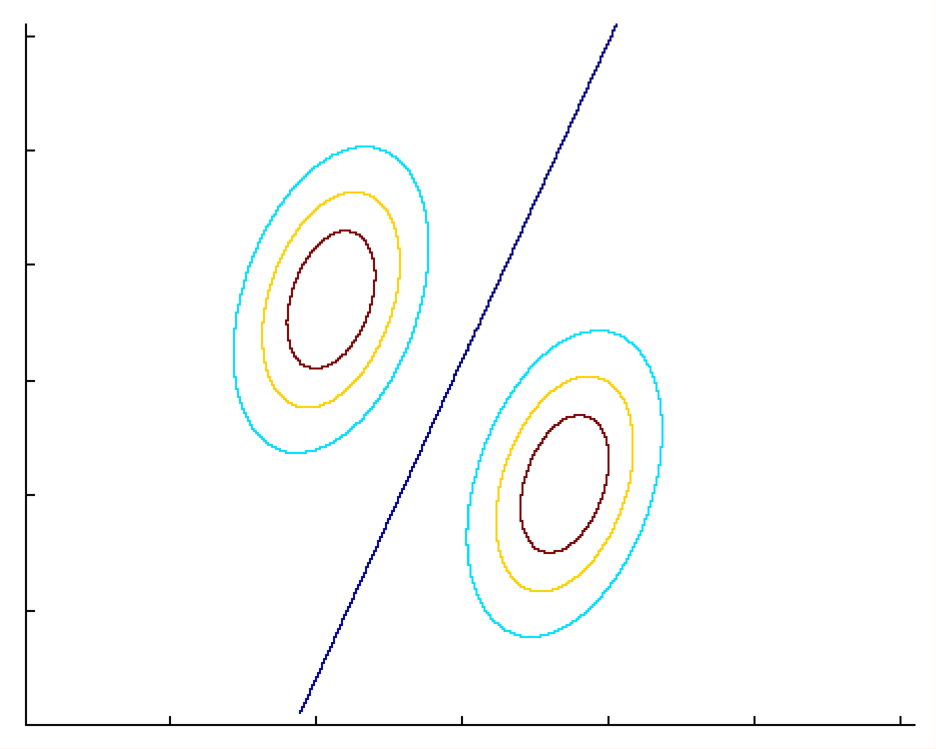
$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^T \mathbf{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) + \log\hat{P}(C_i)$$

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i \quad w_{i0} = -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}^{-1}\mathbf{m}_i + \log\hat{P}(C_i)$$

جداساز خطی



تعداد پارامترها

K.d for means

d.(d+1)/2 for covariance

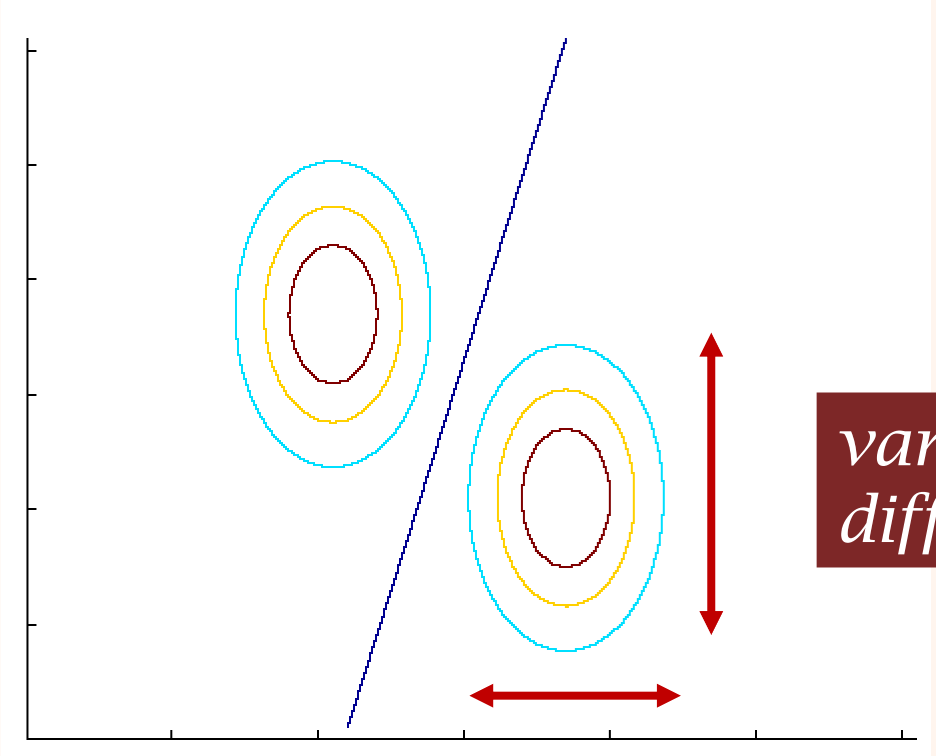
Diagonal S

- در صورتی که متغیرها، مستقل در نظر گرفته شوند، ماتریس کوواریانس قطری خواهد بود:
- $p(\mathbf{x}|C_i) = \prod_j p(x_j | C_i)$

Naïve bayes' classifier

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j^t - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

weighted Euclidean distance



variances may be different

تعداد پارامترها

k.d for means

d for covariance

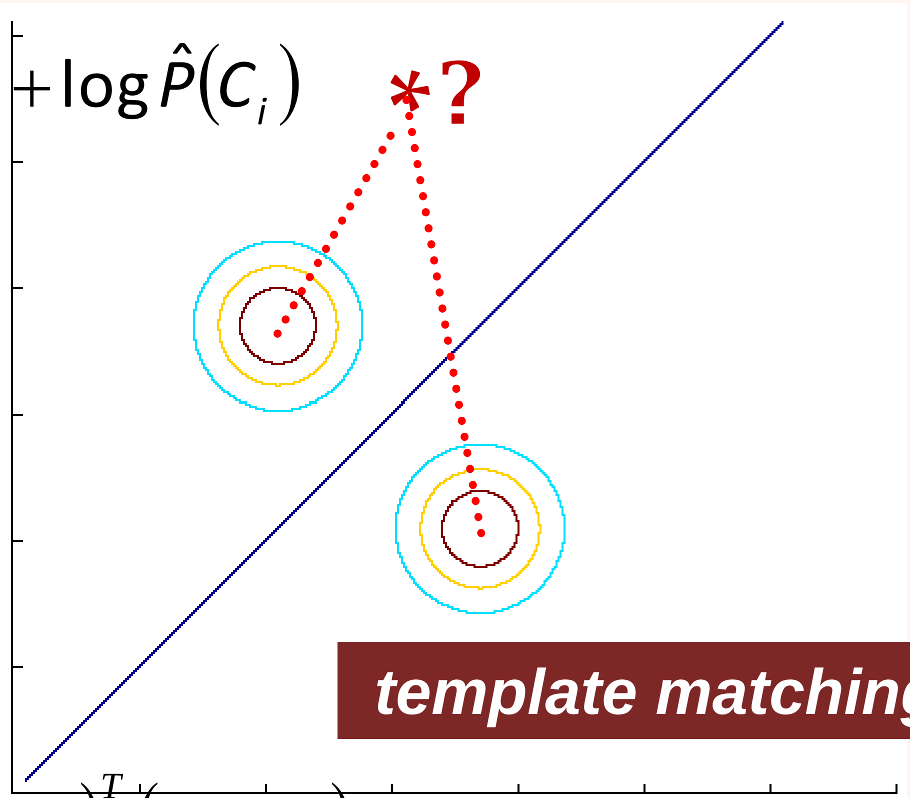
Nearest mean classifier

در صورتی که واریانس متغیرها هم یکسان در نظر گرفته شود:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\mathbf{m}_i\|^2}{2s^2} + \log \hat{P}(C_i)$$

$$= -\frac{1}{2s^2}\sum_{j=1}^{d}\left(x_j^t - m_{ij}\right)^2 + \log \hat{P}(C_i)$$

در صورتی که احتمال تعلق به کلاس‌ها هم‌اندازه باشند، از ضرب داخلی نیز می‌توان استفاده کرد



template matching

$$g_i(\mathbf{x}) = -\|\mathbf{x}-\mathbf{m}_i\|^2 = -(\mathbf{x}-\mathbf{m}_i)^T(\mathbf{x}-\mathbf{m}_i)$$

$$= -\left(\cancel{\mathbf{x}^T\mathbf{x}} - 2\mathbf{m}_i\mathbf{x} + \cancel{\mathbf{m}_i^T\mathbf{m}_i}\right)$$

Tuning Complexity

Assumption	*Covariance matrix*	*No of parameters*
Shared, Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	d
Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$d(d+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ d(d+1)/2$

در نظر گرفتن ماتریس کواریانس مشترک معادل داشتن جداساز خطی است.

فاصله‌ی اقلیدسی زمانی مورد استفاده قرار می‌گیرد، که واریانس همه‌ی متغیرها یکسان در نظر گرفته شود.

رگرسیون خطی

$$g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0$$

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

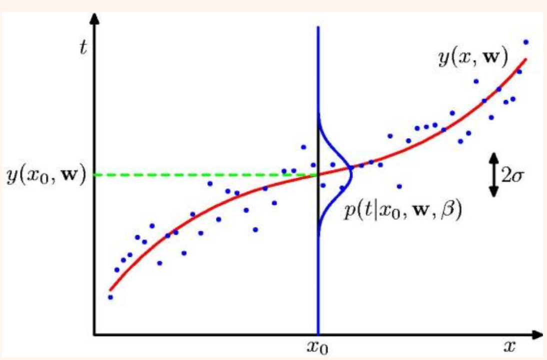
$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

$$\mathbf{w} = \mathbf{A}^{-1} \mathbf{y}$$

رگرسیون چندجمله‌ای

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right)=w_k\left(x^t\right)^k+\cdots+w_2\left(x^t\right)^2+w_1x^t+w_0$$

$$\mathbf{A}=\begin{bmatrix} N & \sum_t x^t & \sum_t\left(x^t\right)^2 & \cdots & \sum_t\left(x^t\right)^k \\ \sum_t x^t & & & \cdots & \sum_t\left(x^t\right)^{k+1} \\ \vdots & & & \cdots & \\ \sum_t\left(x^t\right)^k & \sum_t\left(x^t\right)^{k+1} & \sum_t\left(x^t\right)^{k+2} & \cdots & \sum_t\left(x^t\right)^{2k} \end{bmatrix}$$



$$\mathbf{w}=\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} \qquad \mathbf{y}=\begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t\left(x^t\right)^2 \\ \vdots \\ \sum_t r^t\left(x^t\right)^k \end{bmatrix}$$

رگرسیون چندجمله‌ای

یادآوری

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t & \sum_t \left(x^t\right)^2 & \cdots & \sum_t \left(x^t\right)^k \\ \sum_t x^t & & & \cdots & \sum_t \left(x^t\right)^{k+1} \\ \vdots & & & \cdots & \\ \sum_t \left(x^t\right)^k & \sum_t \left(x^t\right)^{k+1} & \sum_t \left(x^t\right)^{k+2} & \cdots & \sum_t \left(x^t\right)^{2k} \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t \left(x^t\right)^2 \\ \vdots \\ \sum_t r^t \left(x^t\right)^k \end{bmatrix}$$

$$\mathbf{A} = \left(\mathbf{D}^T\mathbf{D}\right) \qquad \mathbf{y} = \mathbf{D}^T\mathbf{r}$$

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & \left(x^1\right)^2 & \cdots & \left(x^1\right)^k \\ 1 & x^2 & \left(x^2\right)^2 & \cdots & \left(x^2\right)^k \\ \vdots & & & & \\ 1 & x^N & \left(x^N\right)^2 & \cdots & \left(x^N\right)^k \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

$$\mathbf{w} = \left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{r}$$

رگرسیون چندمتغیره‌ی خطی

Multivariate linear Regression | **Multiple Regression**

$$r^t = g\left(x^t \mid w_0, w_1, \ldots, w_d\right) + \varepsilon$$

$$= w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_d x_d^t$$

- تابع خطا به صورت زیر به دست می‌آید:

$$E\left(w_0, w_1, \ldots, w_d \mid \mathcal{X}\right) = \frac{1}{2} \sum_t \left[r^t - w_0 - w_1 x_1^t - \cdots - w_d x_d^t\right]^2$$

- مانند آن چه در پیش داشتیم، با مشتق گرفتن، می‌توان ضرایب را به صورت تحلیلی به دست آورد.

رگرسیون چندمتغیره‌ی خطی

$$\mathbf{X}=\begin{bmatrix}1 & x_1^1 & x_2^1 & \cdots & x_d^1\\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2\\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N\end{bmatrix},\ \mathrm{w}=\begin{bmatrix}w_0\\ w_1\\ w_2\\ \vdots\\ w_d\end{bmatrix},\ \mathbf{r}=\begin{bmatrix}r^1\\ r^2\\ \vdots\\ r^N\end{bmatrix}$$

$$\mathbf{A}=\left(\mathbf{D}^T\mathbf{D}\right) \qquad \mathrm{y}=\mathbf{D}^T\mathbf{r}$$

$$\mathbf{w}=\left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{r}$$

$$X^TXW=X^Tr \Rightarrow W=\left(X^TX\right)^{-1}X^Tr$$

این مدل شبیه به مدلی است که برای رگرسیون چند جمله‌ای تک متغیره داشتیم.

$x_1=x,\ x_2=x^2,\ x_3=x^3,\ \ldots.\ x_k=x^k$

برای رگرسیون چندجمله‌ای و چندمتغیره نیز می‌توانیم به صورت مشابه عمل کنیم:

$z_1=x_1,\ z_2=x_2,\ z_3=x_1^2,\ z_4=x_2^2,\ z_5=x_1x_2$