机器学习的数学基础

第5次课

正态分布的概率密度函数

张朝晖

河北师范大学软件学院

2017.10.19

问题的引入：

- 统计决策规则中，涉及“连续随机变量或向量的**概率密度函数**”。
- “正态分布”的概率密度函数，特点：

物理上的合理性。如果在特征空间中的某一类样本，较多地分布在这一类均值附近，远离均值的样本比较少，此时用正态分布作为这一类的概率模型是合理的。

数学上，比较简便。正态分布概率模型有许多好的性质，有利于作数学分析。

→ 简单、符合一些实际情况

主要内容

1. 单个随机变量的正态分布

2. 随机向量的正态分布

连续随机变量X概率密度函数定义

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

记 $p_X(x)\sim N(\mu,\sigma^2)$

标准正态分布：$p_X(x)\sim N(0,1)$

$$\text{其中}\begin{cases} p_X(x)-\text{随机变量(标量)}\\ \mu-\text{随机变量}X\text{的期望，或均值}\\ \qquad \mu\equiv E\{X\}=\int_{-\infty}^{\infty}xp_X(x)dx\\ \sigma-x\text{的标准差，描述}x\text{的分散程度}\\ \sigma^2-x\text{的方差：}\ \sigma^2\equiv E\{(X-\mu)^2\}=\int_{-\infty}^{\infty}(x-\mu)^2p_X(x)dx \end{cases}$$

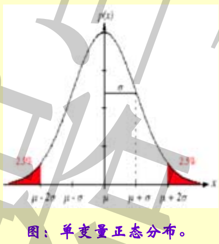
性质：$p_X(x)\ge 0,-\infty<x<+\infty$

$$\int_{-\infty}^{\infty}p_X(x)dx=1$$

$$\int_{\mu-\sigma}^{\mu+\sigma}p_X(x)dx=0.683$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma}p_X(x)dx=0.9544$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma}p_X(x)dx=0.9974$$



图：单变量正态分布。

其它：熵 $H(p_X(x))=-\int p_X(x)\ln p_X(x)dx>0$

Mahalanobis距离 $r=\frac{|x-\mu|}{\sigma}$

$$p(X=\mu)=\frac{1}{\sqrt{2\pi}\sigma}$$

$$E\{f(X)\}=\int_{-\infty}^{\infty}f(x)p_X(x)dx$$

主要内容

1. 单个随机变量的正态分布

2. 随机向量的正态分布

[1]多元正态分布的概率密度函数 -- $p_X(x)$

定义：$p_X(x)=\dfrac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$

记为 $p_X(x)\sim N(\boldsymbol{\mu},\Sigma)$

其中：
- $\boldsymbol{X}$ $-d$维列向量，$\boldsymbol{X}=[X_1,X_2,...,X_d]^T$
  X_i为随机变量，$i=1,2,...,\boldsymbol{d}$；
- $\boldsymbol{\mu}$ $-d$维均值向量，$\boldsymbol{\mu}=[\mu_1,\mu_2,...,\mu_d]^T$
  μ_i为第i维随机变量X_i的均值，$i=1,2,...,\boldsymbol{d}$；
- $\Sigma-d\times d$维协方差矩阵(*covariance matrix*)
- Σ^{-1} – 矩阵Σ的逆矩阵，精度矩阵
- $|\Sigma|$ – 矩阵Σ的行列式

[1]多元正态分布概率密度函数定义(续) -- $\boldsymbol{\mu}$

$\boldsymbol{\mu}$ $-d$维均值向量，$\boldsymbol{\mu}=[\mu_1,\mu_2,...,\mu_d]^T$

$$\boldsymbol{\mu}\equiv\boldsymbol{E}\{\boldsymbol{X}\}=\int_{\Re^d}\boldsymbol{x}p_X(x)d\boldsymbol{x}$$

$$=\left[\boldsymbol{E}\{x_1\},\boldsymbol{E}\{x_2\},...,\boldsymbol{E}\{x_d\}\right]^T=[\mu_1,\mu_2,...,\mu_d]^T$$

其中：

$$\begin{cases}\mu_i=\boldsymbol{E}\{X_i\}=\displaystyle\int_{\Re^d}x_i p_X(x)d\boldsymbol{x}\\ \quad=\displaystyle\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}x_i p_X(x)dx_1dx_2...dx_d=\int_{-\infty}^{+\infty}x_i p(x_i)dx_i\\ \text{边缘密度函数}\quad p_{X_i}(x_i)=\displaystyle\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p_X(x)dx_1...dx_{i-1}dx_{i+1}...dx_d\end{cases}$$

[1]多元正态分布概率密度函数定义(续) -- Σ

$\Sigma-d\times d$维协方差矩阵(*covariance matrix*)：

$$\Sigma\equiv\boldsymbol{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\}=\int_{\Re^d}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T p_X(x)dx$$

$$\Sigma=\boldsymbol{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\}=\boldsymbol{E}\left\{\begin{bmatrix}X_1-\mu_1\\X_2-\mu_2\\\vdots\\X_d-\mu_d\end{bmatrix}\begin{bmatrix}X_1-\mu_1 & X_2-\mu_2 & \cdots & X_d-\mu_d\end{bmatrix}\right\}$$

$$=\boldsymbol{E}\begin{bmatrix}(X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_d-\mu_d)\\(X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_d-\mu_d)\\\vdots & \vdots & & \vdots\\(X_d-\mu_d)(X_1-\mu_1) & (X_d-\mu_d)(X_2-\mu_2) & \cdots & (X_d-\mu_d)^2\end{bmatrix}$$

=(下页待续)

[1]多元正态分布概率密度函数定义(续) -- Σ

$\Sigma-d\times d$维协方差矩阵(*covariance matrix*)：

$$\Sigma\equiv\boldsymbol{E}\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\right]$$

$$=\begin{bmatrix}\boldsymbol{E}\{(X_1-\mu_1)^2\} & \boldsymbol{E}\{(X_1-\mu_1)(X_2-\mu_2)\} & \cdots & \boldsymbol{E}\{(X_1-\mu_1)(X_d-\mu_d)\}\\\boldsymbol{E}\{(X_2-\mu_2)(X_1-\mu_1)\} & \boldsymbol{E}\{(X_2-\mu_2)^2\} & \cdots & \boldsymbol{E}\{(X_2-\mu_2)(X_d-\mu_d)\}\\\vdots & \vdots & \ddots & \vdots\\\boldsymbol{E}\{(X_d-\mu_d)(X_1-\mu_1)\} & \boldsymbol{E}\{(X_d-\mu_d)(X_2-\mu_2)\} & \cdots & \boldsymbol{E}\{(X_d-\mu_d)^2\}\end{bmatrix}$$

$$=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1d}\\\sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d}\\\cdots & \cdots & \cdots\cdots & \cdots\\\sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd}\end{bmatrix}\quad\Sigma\text{对称非负定}\begin{cases}\text{对角线元素}\sigma_{ii}=\sigma_i^2,\quad X_i\text{方差}\\\text{非对角元素}\sigma_{ij},\ X_i、X_j\text{协方差}\end{cases}$$

这里：只考虑**Σ对称正定**

[1]多元正态分布概率密度函数定义(续) -- Σ

$\Sigma-d\times d$维协方差矩阵(*covariance matrix*)：

$$\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1d}\\\sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d}\\\cdots & \cdots & \cdots\cdots & \cdots\\\sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd}\end{bmatrix}$$

$$\sigma_{ij}=\boldsymbol{E}\{(X_i-\mu_i)(X_j-\mu_j)\}=\int_{\Re^d}(x_i-\mu_i)(x_j-\mu_j)p_X(\boldsymbol{x})d\boldsymbol{x}$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(x_i-\mu_i)(x_j-\mu_j)p_X(\boldsymbol{x})dx_1dx_2...dx_d$$

$$=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_i-\mu_i)(x_j-\mu_j)p_{X_iX_j}(x_i,x_j)dx_idx_j$$

其中，边缘密度：

$$p_{X_iX_j}(x_i,x_j)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p_X(\boldsymbol{x})dx_1...dx_{i-1}dx_{i+1}...dx_{j-1}dx_{j+1}...dx_d$$

[2]多元正态分布概率密度函数性质

性质1　多元正态分布完全由μ、Σ决定

参数共计$\boldsymbol{d}+\dfrac{\boldsymbol{d}(\boldsymbol{d}+1)}{2}$个

$$\boldsymbol{\mu}=[\mu_1,\mu_2,...,\mu_d]^T$$

$$\Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots\cdots & \sigma_{1d}\\\sigma_{21} & \sigma_{22} & \cdots\cdots & \sigma_{2d}\\\cdots & \cdots & \cdots\cdots & \cdots\\\sigma_{d1} & \sigma_{d2} & \cdots\cdots & \sigma_{dd}\end{bmatrix}\text{为对称矩阵}$$

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

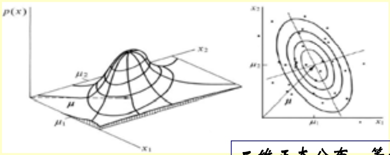
$$p_X(\boldsymbol{x})\sim N(\boldsymbol{\mu},\Sigma)$$

性质2 等概率密度点的轨迹为一超椭球面

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

$\mathbf{x}$到$\boldsymbol{\mu}$的 **Mahalanobis距离**平方 $\quad \mathbf{r}^2=(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$

等概率密度点: $\mathbf{r}^2=$常数



二维正态分布，等概率密度点的轨迹为椭圆周。

超椭球面中心$\boldsymbol{\mu}$

超椭球面主轴$\begin{cases}\text{方向取决于}\Sigma\text{本征向量}\\ \text{长度与}\Sigma\text{本征值的平方根成正比。}\end{cases}$

性质3 不相关性等价于独立性。

$\mathbf{x}$多元正态分布$\begin{cases}\text{“}\mathbf{X}\text{任意两分量}X_i,X_j\text{间互不相关”}\\ \quad\Leftrightarrow\text{“各分量间相互独立”}\\ \text{协方差矩阵}\Sigma\text{是对角矩阵}\Rightarrow \mathbf{X}\text{各分量相互独立}\end{cases}$

定义：随机变量X_i,X_j之间

$$\begin{cases}X_i,X_j\text{间不相关：}\ \mathbf{E}\{X_iX_j\}=\mathbf{E}\{X_i\}\mathbf{E}\{X_j\}\\ X_i,X_j\text{相互独立：}\ p_{X_iX_j}(x_i,x_j)=p(x_i)p(x_j)\end{cases}$$

并且

$$p_{X_iX_j}(x_i,x_j)=p_{X_i}(x_i)p_{X_j}(x_j)\Rightarrow \mathbf{E}\{X_iX_j\}=\mathbf{E}\{X_i\}\mathbf{E}\{X_j\}$$

性质3 不相关性等价于独立性。

多元正态分布的任意随机变量间不相关性，等价于独立性。

证：$\forall X_i,X_j,i\neq j,\quad i,j=1,\ldots,d$若$X_i,X_j$间互不相关，则

$$\sigma_{ij}=\mathbf{E}\left\{(X_i-\mu_i)(X_j-\mu_j)\right\}=\mathbf{E}\left\{X_iX_j-X_i\mu_j-X_j\mu_i+\mu_i\mu_j\right\}$$
$$=\mathbf{E}\{X_i\}\mathbf{E}\{X_j\}-\mu_j\mathbf{E}\{X_i\}-\mu_i\mathbf{E}\{X_j\}+\mu_i\mu_j=0$$

$$\text{所以：}\Sigma=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\cdots\cdots&\sigma_{1d}\\ \sigma_{21}&\sigma_{22}&\cdots\cdots&\sigma_{2d}\\ \cdots&\cdots&\cdots\cdots&\cdots\\ \sigma_{d1}&\sigma_{d2}&\cdots\cdots&\sigma_{dd}\end{bmatrix}=\begin{bmatrix}\sigma_{11}&&&\\ &\sigma_{22}&&\\ &&\cdots\cdots&\\ &&&\sigma_{dd}\end{bmatrix}$$

因此：若X_i,X_j间互不相关，则协方差矩阵Σ是对角的

性质3 不相关性等价于独立性。

$$\text{证(续1)}:\Sigma^{-1}=\begin{bmatrix}1/\sigma_{11}&&&\\ &1/\sigma_{22}&&\\ &&\cdots\cdots&\\ &&&1/\sigma_{dd}\end{bmatrix}$$

$$|\Sigma|=\prod_{i=1}^{d}\sigma_{ii}\qquad |\Sigma|^{\frac{1}{2}}=\prod_{i=1}^{d}\sqrt{\sigma_{ii}}=\prod_{i=1}^{d}\sigma_i$$

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})=[x_1-\mu_1,\ldots,x_d-\mu_d]\begin{bmatrix}1/\sigma_{11}&&0\\ &\ddots&\\ 0&&1/\sigma_{dd}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ \vdots\\ x_d-\mu_d\end{bmatrix}$$

$$=\left[(x_1-\mu_1)/\sigma_{11}\ \cdots\ (x_d-\mu_d)/\sigma_{dd}\right]\begin{bmatrix}x_1-\mu_1\\ \vdots\\ x_d-\mu_d\end{bmatrix}=\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

性质3 不相关性等价于独立性。

证(续2): $\quad (\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})=\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$

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

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

$$=\prod_{i=1}^{d}p_{X_i}(x_i)$$ 所以$\mathbf{X}$各分量相互独立

所以对于多元正态分布$\mathbf{X}$，

$\begin{cases}\text{各分量}X_i,X_j\text{间互不相关}\Leftrightarrow\text{各分量相互独立。}\\ \text{协方差矩阵}\Sigma\text{是对角的}\Rightarrow \mathbf{X}\text{各分量相互独立，且正态分布}\end{cases}$

性质4 边缘密度分布与条件密度分布仍为正态分布。

边缘概率密度：$p_{X_i}(x_i)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p_X(\mathbf{x})dx_1\ldots dx_{i-1}dx_{i+1}\ldots dx_d$

$$p_{X_i}(x_i)=\frac{1}{\sqrt{2\pi}\sigma_{ii}}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

$\mathbf{i}=1,\ldots,\mathbf{d}$

条件概率密度：给定$\mathbf{X}_i=x_i$条件下，X_j的分布。

$$p_{X_i|X_j}(x_j|x_i)=\frac{p_{X_iX_j}(x_i,x_j)}{p_{X_i}(x_i)}$$

其中：

$$\begin{cases}p_{X_iX_j}(x_i,x_j)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p_X(\mathbf{x})dx_1\ldots dx_{i-1}dx_{i+1}\ldots dx_{j-1}dx_{j+1}\ldots dx_d\\ p_{X_i}(x_i)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}p_X(\mathbf{x})dx_1\ldots dx_{i-1}dx_{i+1}\ldots dx_d\end{cases}$$

性质5 线性变换的正态性

随机向量X经线性变换后，仍为正态分布。

若
$$\begin{cases} \text{随机向量} \boldsymbol{X} = [X_1, \ldots, X_d]^T, \; \boldsymbol{p_X}(\boldsymbol{x}) \sim \boldsymbol{N}(\boldsymbol{\mu}, \Sigma) \\ \text{变换矩阵} \boldsymbol{A} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \cdots \quad \boldsymbol{a}_k] \quad d\text{行} \times k\text{列}; \\ \text{线性变换} \quad \boldsymbol{Y} = \boldsymbol{A}^T \boldsymbol{X} \end{cases}$$

则 $\boldsymbol{p_Y}(\boldsymbol{y}) \sim \boldsymbol{N}\left(\boldsymbol{A}^T \boldsymbol{\mu}, \boldsymbol{A}^T \Sigma \boldsymbol{A}\right)$

性质5 线性变换的正态性

线性变换：$\boldsymbol{Y} = \boldsymbol{A}^T \boldsymbol{X}$

说明1： 适当选择$\boldsymbol{A}$为某种非奇异矩阵，**可使$\boldsymbol{Y}$中各分量相互独立**

即 $\Sigma_{\mathrm{Y}} =$ 对角矩阵

变换矩阵：$\boldsymbol{A} = \Phi$；

协方差矩阵：$\Sigma_{\mathrm{Y}} = \boldsymbol{A}^T \Sigma \boldsymbol{A} = \Phi^T \Sigma \Phi = \Lambda$

说明2：可将坐标变换至超球坐标系 $\Sigma_{\mathrm{Y}} = \boldsymbol{I}$

$\boldsymbol{A} = \boldsymbol{A}_\omega = \Phi \Lambda^{-\frac{1}{2}}$（白化变换）；

$$\Sigma_{\mathrm{Y}} = \boldsymbol{A}_\omega{}^T \Sigma \boldsymbol{A}_\omega = \left(\Phi \Lambda^{-\frac{1}{2}}\right)^T \Sigma \left(\Phi \Lambda^{-\frac{1}{2}}\right) = \Lambda^{-\frac{1}{2}} \Phi^T \Sigma \Phi \Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}} = \boldsymbol{I}$$

性质6 线性组合的正态性

正态分布的$\boldsymbol{X}$各分量的线性组合仍为正态分布。

若：随机向量 $\boldsymbol{X} = [x_1, \ldots, x_d]^T, \; p_X(\boldsymbol{x}) \sim \boldsymbol{N}(\boldsymbol{\mu}, \Sigma)$

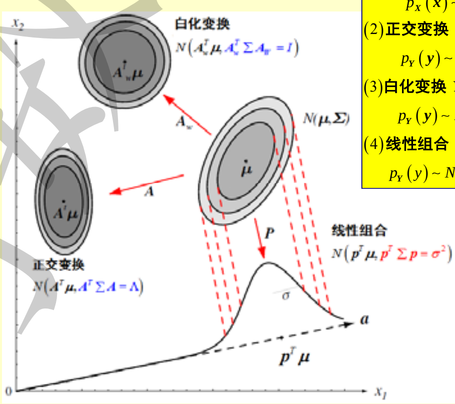
d维向量 $\boldsymbol{a} = [a_1, \ldots, a_d]^T$；

随机变量 $Y = \boldsymbol{a}^T \boldsymbol{X}$

则：$\boldsymbol{p_Y}(y) \sim \boldsymbol{N}\left(\mu_y, \sigma_y^2\right) = \boldsymbol{N}\left(\boldsymbol{a}^T \boldsymbol{\mu}, \boldsymbol{a}^T \Sigma \boldsymbol{a}\right)$

$\Rightarrow$ 通过线性组合可获知观测数据在a方向的整体分散程度。

多元正态分布特征空间的线性变换



(1)**二维正态分布** $\boldsymbol{X}$

$p_X(\boldsymbol{x}) \sim N(\boldsymbol{\mu}, \Sigma)$

(2)**正交变换** $\boldsymbol{Y} = \boldsymbol{A}^T \boldsymbol{X}$

$p_Y(\boldsymbol{y}) \sim N\left(\boldsymbol{A}^T \boldsymbol{\mu}, \boldsymbol{A}^T \Sigma \boldsymbol{A} = \Lambda\right)$

(3)**白化变换** $\boldsymbol{Y} = \boldsymbol{A}_w^T \boldsymbol{X}$

$p_Y(\boldsymbol{y}) \sim N\left(\boldsymbol{A}_w^T \boldsymbol{\mu}, \boldsymbol{A}_w^T \Sigma \boldsymbol{A}_W = I\right)$

(4)**线性组合** $Y = \boldsymbol{p}^T \boldsymbol{X}$

$p_Y(y) \sim N\left(\boldsymbol{p}^T \boldsymbol{\mu}, \boldsymbol{p}^T \Sigma \boldsymbol{p} = \sigma^2\right)$

下一章主题

如何估计类条件概率密度函数？