第二章 贝叶斯决策理论

2009.09.29

基于最小风险的贝叶斯决策

- **条件期望损失**：对于特定的观察样本 $\mathbf{x}$（特征向量），决策 α_i 造成的损失对 $\mathbf{x}$ 实际所属类别的各种可能的平均，也叫做条件风险：

$$R(\alpha_i \mid \mathbf{x}) = E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j \mid \mathbf{x})$$

- **期望风险**：对所有 $\mathbf{x}$ 取值所作的决策 $\alpha(\mathbf{x})$ 所带来的平均风险，即条件风险对 $\mathbf{x}$ 的数学期望。

$$R(\alpha) = E[R(\alpha(\mathbf{x}) \mid \mathbf{x})] = \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

基于最小风险的贝叶斯决策

- **目标**：决策带来的损失的平均值——（平均）风险最小。
- **决策规则**

$$Decide \quad \alpha_k, \quad \text{if} \quad R(\alpha_k \mid \mathbf{x}) = \min_{j=1,\cdots,a} R(\alpha_i \mid \mathbf{x})$$

  - 通过保证对于每个观测值下的条件风险最小，使得决策的数学期望——平均风险最小。基于最小风险的贝叶斯决策是一致最优决策。

基于最小风险的贝叶斯决策

- 最小风险决策的**计算步骤**
  - 在已知 $P(\omega_i)$，$p(\mathbf{x} \mid \omega_i)$，$i=1,\ldots,c$，以及给定待识别样本 $\mathbf{x}$ 的情况下，根据贝叶斯公式计算后验概率；
  - 利用后验概率及决策表（或损失矩阵），算出每个决策的条件风险 $R(\alpha_i \mid \mathbf{x})$；
  - 按照最小的条件风险进行决策。

◆ 损失矩阵在某些特殊问题，存在简单的解析表达式。
◆ 实际问题中得到合适的损失矩阵不容易。

基于最小风险的贝叶斯决策

- 两类别问题
  - 定义（符号简化）
    - 行为 α_1： deciding ω_1;
    - 行为 α_2： deciding ω_2;
    - 损失 $\lambda_{ij} = \lambda(\alpha_i, \omega_j), \quad i, j = 1, 2.$
  - 条件风险

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

基于最小风险的贝叶斯决策

- 两类别问题
  - 决策规则

$$\text{Decide} \begin{cases} \omega_1, & \text{if } (\lambda_{21} - \lambda_{11}) P(\omega_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2 \mid \mathbf{x}) \\ \omega_2, & \text{otherwise} \end{cases}$$

用贝叶斯公式展开➔

$$\text{Decide} \begin{cases} \omega_1, & \text{if } \dfrac{P(\mathbf{x} \mid \omega_1)}{P(\mathbf{x} \mid \omega_2)} > \dfrac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \dfrac{P(\omega_2)}{P(\omega_1)} \\ \omega_2, & \text{otherwise} \end{cases}$$

基于最小风险的贝叶斯决策

- 例解：两类细胞识别问题：正常类(ω_1)和异常类(ω_2)
  - 根据已有知识和经验，两类的先验概率为：

    正常(ω_1)：　$P(\omega_1)=0.9$

    异常(ω_2)：　$P(\omega_2)=0.1$

    对某一样本观察值 $\mathbf{x}$，通过计算或查表得到：

    $p(\mathbf{x}|\omega_1)=0.2$，　$p(\mathbf{x}|\omega_2)=0.4$

    $\lambda_{11}=0$，　$\lambda_{12}=6$，　$\lambda_{21}=1$，　$\lambda_{22}=0$
  - 按最小风险决策如何对细胞 $\mathbf{x}$ 进行分类？

基于最小风险的贝叶斯决策

- 利用贝叶斯公式计算两类的后验概率：

$$P(\omega_1|\mathbf{x})=\frac{0.9\times0.2}{0.9\times0.2+0.1\times0.4}=0.818,$$

$$P(\omega_2|\mathbf{x})=\frac{0.4\times0.1}{0.2\times0.9+0.4\times0.1}=0.182;$$

$$R(\alpha_1|\mathbf{x})=\sum_{j=1}^{2}\lambda_{1j}P(\omega_j|\mathbf{x})=\lambda_{12}(\omega_2|\mathbf{x})=1.092,$$

$$R(\alpha_2|\mathbf{x})=\sum_{j=1}^{2}\lambda_{2j}P(\omega_j|\mathbf{x})=\lambda_{21}(\omega_1|\mathbf{x})=0.818;$$

$$\because R(\alpha_1|\mathbf{x})>R(\alpha_2|\mathbf{x}),\quad \therefore \text{Decide } \alpha_2,\ \mathbf{x}\in\omega_2.$$

两种决策方法之间的关系

- 基于最小错误率的Bayes决策可作为最小风险Bayes决策的一种特殊情形。
- 定义损失为

$$\lambda(\alpha_i,\omega_j)=\begin{cases}0, & i=j\\ 1, & i\neq j\end{cases},\quad i,j=1,\cdots,c.$$

  - 不考虑“拒绝”等其他决策；
  - 决策正确时没有损失；决策错误时损失为1；即0-1损失函数。

两种决策方法之间的关系

- 条件风险

$$R(\alpha_i|\mathbf{x})=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j|\mathbf{x})$$

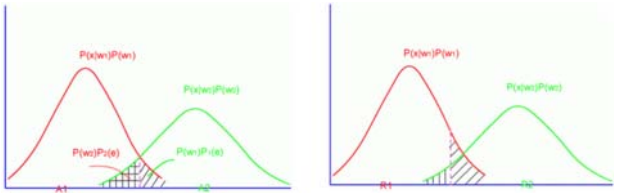
$$=\sum_{j=1,\,j\neq i}^{c}P(\omega_j|\mathbf{x})=1-P(\omega_i|\mathbf{x})$$

- 决策规则

$$\min_{\alpha}R(\alpha_i|\mathbf{x})\Leftrightarrow\max_{\alpha}P(\omega_i|\mathbf{x})$$

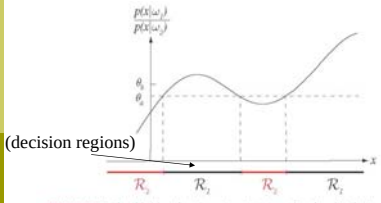
两种决策方法之间的关系

- 图例一



两种决策方法之间的关系

- 图例二
  - 由0-1损失函数确定阈值 θ_a；
  - 给予假设 $\lambda_{12}>\lambda_{21}$ 确定阈值 θ_b。(R_1变小)



$$\theta_a=P(\omega_2)/P(\omega_1)$$

$$\theta_b=\frac{P(\omega_2)(\lambda_{12}-\lambda_{22})}{P(\omega_1)(\lambda_{21}-\lambda_{11})}$$

FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Neyman-Pearson 决策

- **问题的提出**
  - 某些两类判决问题，某一类错误较另一类错误更为重要 — 损失更为严重。例如在癌细胞识别问题中，把异常误判为正常的损失更为严重。
  - 先验概率未知。
- **基本思想**
  - 严格限制较重要的一类错误概率，在令其等于某常数的约束下使另一类误判概率最小。

Neyman-Pearson 决策

- **两类错误率**
  - 令 R 是整个特征空间，R_1 是类别 ω_1 的决策域，R_2 是类别 ω_2 的决策域：$R_1 + R_2 = R$。

$$\begin{aligned} P(error) &= \int_{R_1} P(\omega_2 \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \int_{R_2} P(\omega_1 \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int_{R_1} p(\mathbf{x} \mid \omega_2) P(\omega_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x} \mid \omega_1) P(\omega_1) d\mathbf{x} \\ &= P(\omega_2) \int_{R_1} p(\mathbf{x} \mid \omega_2) dx + P(\omega_1) \int_{R_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} \\ &= P(\omega_2) P_2(error) + P(\omega_1) P_1(error) \end{aligned}$$

  - $P_1(error)$，$P_2(error)$即两类错误率。

Neyman-Pearson 决策

- **决策目标：** 在$P_2(error) = \varepsilon_0$条件下，求$P_1(error)$极小值。
  - 根据Lagrange乘子法，建立数学模型

$$\gamma = P_1(error) + \lambda (P_2(error) - \varepsilon_0),$$

其中 λ 是Lagrange乘子，目标是求 γ 的极小值。

注意：$P_1(error) = \int_{R_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} = 1 - \int_{R_1} p(\mathbf{x} \mid \omega_1) d\mathbf{x};$

$$P_2(error) = \int_{R_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} = 1 - \int_{R_2} p(\mathbf{x} \mid \omega_2) d\mathbf{x} = \varepsilon_0.$$

Neyman-Pearson 决策

- **决策目标：** 极小化 γ

$$\gamma = (1 - \lambda \varepsilon_0) + \int_{R_1} [\lambda p(\mathbf{x} \mid \omega_2) - p(\mathbf{x} \mid \omega_1)] d\mathbf{x};$$

或者

$$\gamma = (1 - \varepsilon_0) \lambda + \int_{R_2} [p(\mathbf{x} \mid \omega_1) - \lambda p(\mathbf{x} \mid \omega_2)] d\mathbf{x};$$

  - 对于固定的 λ，要使得 γ 最小，应满足

$$\forall \mathbf{x} \in R_1, \quad \lambda p(\mathbf{x} \mid \omega_2) - p(\mathbf{x} \mid \omega_1) < 0;$$

$$\forall \mathbf{x} \in R_2, \quad p(\mathbf{x} \mid \omega_1) - \lambda p(\mathbf{x} \mid \omega_2) < 0;$$

Neyman-Pearson 决策

- **决策准则**

$$\text{if } \lambda p(\mathbf{x} \mid \omega_2) \underset{>}{<} p(\mathbf{x} \mid \omega_1), \quad \text{then } \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

or

$$l(\mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \underset{<}{>} \lambda, \quad \text{then } \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

➔N-P决策规则归结为找阈值 λ，使得

$$\int_{R_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} = \varepsilon_0.$$

  - λ 的显式解不易求解，可用试探法。

Neyman-Pearson 决策

- 求决策准则的**方法二**
  - 令 t 是 R_1 和 R_2 的分界点（面），将 γ 分别对 t 和 λ 求偏导，γ 极值点存在的必要条件是：

$$\frac{\partial \gamma}{\partial t} = 0 \Rightarrow \lambda = \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)};$$

$$\frac{\partial \gamma}{\partial \lambda} = 0 \Rightarrow \int_{R_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x} = \varepsilon_0;$$

    - 方程式确定一个分界面，使得$P_2(error) = \varepsilon_0$，同时又使得$P_1(error)$尽可能小。该分界面上 $\mathbf{x}$ 值具有一个特点，即它们的两类条件密度函数之比是一个常数，该比值就是Lagrange乘子 λ 。

Neyman-Pearson 决策

- **例解**：一个两类问题中，模式均为二维正态分布，其均值矢量和协方差阵分别为：

$$\mu_1=(-1,0)^T,\ \mu_2=(1,0)^T,\ \Sigma_1=\Sigma_2=I.$$

设 $\varepsilon_0=0.09$，求Neyman-Pearson的决策阈值。

- 解：

$$p(\mathbf{x}\mid\omega_1)=\frac{1}{2\pi}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)^T(\mathbf{x}-\mu_1)\right]=\frac{1}{2\pi}\exp\left[-\frac{1}{2}\left((x_1+1)^2+x_2^2\right)\right];$$

$$p(\mathbf{x}\mid\omega_2)=\frac{1}{2\pi}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_2)^T(\mathbf{x}-\mu_2)\right]=\frac{1}{2\pi}\exp\left[-\frac{1}{2}\left((x_1-1)^2+x_2^2\right)\right];$$

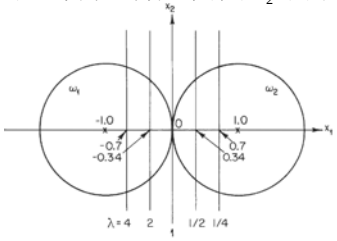
$$\therefore\ \frac{p(\mathbf{x}\mid\omega_1)}{p(\mathbf{x}\mid\omega_2)}=\exp(-2x_1).$$

Neyman-Pearson 决策

- **例解**
  - 判决准则：

if $\exp(-2x_1)\ \begin{matrix}>\\<\end{matrix}\ \lambda$, i. e. $x_1\ \begin{matrix}<\\>\end{matrix}\ -\frac{1}{2}\ln\lambda$, then $\mathbf{x}\in\begin{cases}\omega_1\\ \omega_2\end{cases}$;

∴对于不同的λ，决策边界是平行于x_2的不同直线。（如图）



Neyman-Pearson 决策

- **例解**
  - 通过计算$P_2(error)=\varepsilon_0$求解λ：

$$P_2(error)=\int_{R_1}p(\mathbf{x}\mid\omega_2)d\mathbf{x}$$

$$=\int_{-\infty}^{-\frac{1}{2}\ln\lambda}\int_{-\infty}^{\infty}\frac{1}{2\pi}\exp\left[-\frac{(x_1-1)^2+x_2^2}{2}\right]dx_2dx_1$$

$$=\int_{-\infty}^{-\frac{1}{2}\ln\lambda}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(x_1-1)^2}{2}\right]dx_1.$$

λ	4	2	1	1/2	1/4
ε_0	0.046	0.089	0.0159	0.258	0.378

λ与ε_0的关系表

Neyman-Pearson 决策

- 最小错误率的Bayes决策与N-P决策
  - 均以似然比为基础;
  - 最小错误率的Bayes决策的阈值是先验概率之比

$$\frac{P(\omega_2)}{P(\omega_1)};$$

  - Neyman-Pearson决策的阈值是Lagrange乘子（和先验概率无关）。

其他决策方法（自学）

- **最大最小决策**
  - 基本思想：类先验概率未知，考查先验概率变化对错误率的影响，找出使最小风险贝叶斯决策的风险最大的先验概率，以这种最坏情况设计分类器。
- **序贯分类方法**
  - 基本思想：除考虑分类造成的损失外，还考虑特征获取所造成的代价。先用一部分特征分类，然后逐步加入新特征以减少分类损失，同时衡量总的损失，以求得最优的效益。

分类器设计

- **分类器**(classifier)：能够将每个样本都分到某个类别中去（或者拒绝）的计算机算法。
  - 是从特征空间到决策空间的映射。
- **决策域**(decision region)：分类器将 *d* 维特征空间划分为若干区域。
- **决策面**(decision boundary)：不同类别区域之间的边界，又叫作分类边界、决策边界或分类面。数学上用解析形式表示成决策面方程。

分类器设计

- **判别函数**(discriminant functions)：是模式（或特征向量）$\mathbf{x}$ 的函数，用于表述决策规则。
  - 对于c类别问题，相应于每一类别定义一个函数，构成一组判别函数 $g_i(\mathbf{x})$, $i = 1,2,\ldots,c$，使得

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \;\Rightarrow\; \mathbf{x} \in \omega_i \quad j = 1,\cdots,c,\quad j \neq i;$$

  即将 $\mathbf{x}$ 分类到有最大判别函数值的类别。
- 判别函数的选择不唯一。如果 $f(\cdot)$ 是一个单调递增函数（如logarithm），将 $g_i(\mathbf{x})$ 替换成 $f(g_i(\mathbf{x}))$ 不改变判决结果。➔ **简化分析和计算!**

分类器设计

- **最小错误率Bayes决策**
  - 决策规则：将x归于 ω_i类，如果

(1) $P(\omega_i \mid \mathbf{x}) = \max_{j=1,\cdots,c} P(\omega_j \mid \mathbf{x})$;

or (2) $p(\mathbf{x} \mid \omega_i)P(\omega_i) = \max_{j=1,\cdots,c} p(\mathbf{x} \mid \omega_j)P(\omega_j)$;

or (3) $l(\mathbf{x}) = \dfrac{p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x} \mid \omega_j)} > \dfrac{p(\mathbf{x} \mid \omega_j)}{p(\mathbf{x} \mid \omega_i)},\ j = 1,\cdots,c, j \neq i;$

or (4) $\ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i) = \max_{j=1,\cdots,c} \left(\ln p(\mathbf{x} \mid \omega_j) + \ln P(\omega_j)\right);$

  - 判别函数

(1) $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$

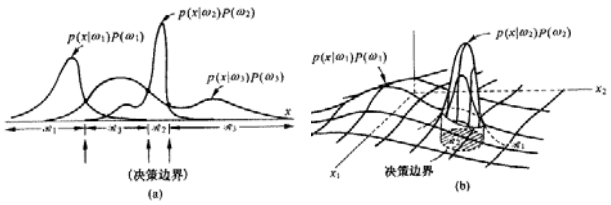
(2) $g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$

(3) $g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$

分类器设计

- **最小错误率Bayes决策**
  - 决策面方程：相邻的两个决策域在决策面上的判别函数值相等，即

$$g_i(\mathbf{x}) = g_j(\mathbf{x}).$$



分类器设计

- **最小错误率Bayes决策**
  - 分类器：一个计算 c 个判别函数并选取与最大判别函数值相对应的类别的网络或机器。

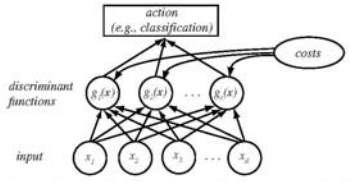


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

分类器设计

- **两类别**的最小错误率Bayes决策
  - 判决函数：可只定义一个判别函数

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}),$$

此时的决策规则是

$$\text{if } g(\mathbf{x}) \begin{matrix} > \\ < \end{matrix} 0, \text{ then decide } \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}.$$

(1) $g(\mathbf{x}) = P(\omega_1 \mid \mathbf{x}) - P(\omega_2 \mid \mathbf{x})$

(2) $g(\mathbf{x}) = p(\mathbf{x} \mid \omega_1)P(\omega_1) - p(\mathbf{x} \mid \omega_2)P(\omega_2)$

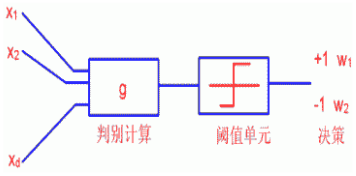
(3) $g(\mathbf{x}) = \ln \dfrac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \dfrac{P(\omega_1)}{P(\omega_2)}$

分类器设计

- **两类别**的最小错误率Bayes决策
  - 决策面方程

$$g(\mathbf{x}) = 0.$$

  - 分类器

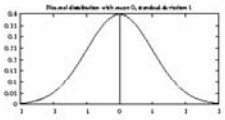


正态分布

- 目的：结合一种比较典型的概率分布来进一步研究基于最小错误率Bayes决策分类器。
- Bayes决策的三个前提：
  - ①类别数确定；②各类的先验概率$P(\omega_i)$已知；③各类的条件概率密度函数$p(\mathbf{x}|\omega_i)$已知。
- Bayes决策中，类条件概率密度的选择要求：
  - 模型合理性
  - 计算可行性
- 最常用概率密度模型：正态分布
  - 观测值通常是很多种因素共同作用的结果，根据中心极限定理，它们（近似）服从正态分布。
  - 计算、分析最为简单的模型。

单变量的正态分布

A bell-shaped distribution defined by the probability density function



$$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

If the random variable X follows a normal distribution, then

- The probability that X will fall into the interval (a,b) is given by $\int_a^b p(x)dx$
- Expected, or mean, value of X is $E[X]=\int_{-\infty}^{\infty}xp(x)dx=\mu$
- Variance of X is $Var(x)=E[(x-\mu)^2]=\int_{-\infty}^{\infty}(x-\mu)^2p(x)dx=\sigma^2$
- Standard deviation of X, σ^2, is $\sigma_x=\sigma$

单变量的正态分布

- $p(x)$完全由μ与σ^2确定，常记作$N(\mu,\sigma^2)$。
- 正态分布的熵(entropy)在所有的已知均值及方差的分布中最大。

$$H(p(x))=-\int_{-\infty}^{\infty}p(x)\ln p(x)dx;$$

- $p(x)$关于均值对称，最大值位于$x=\mu$处，

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

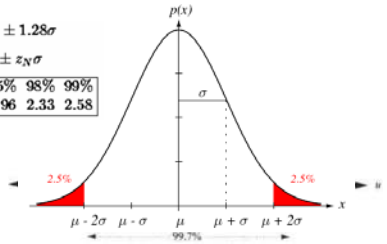
单变量的正态分布

- 样本主要集中分布在其均值附近，其分散程度用标准差来衡量，σ愈大分散程度也越大。从分布的总体中抽取样本，约有95%的样本都落在区间$(\mu-2\sigma,\mu+2\sigma)$内。

80% of area (probability) lies in $\mu\pm1.28\sigma$

N% of area (probability) lies in $\mu\pm z_N\sigma$

N%:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58



多元正态分布

- **概率密度函数**

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

其中：

$\mathbf{x}=[x_1,x_2,\cdots,x_d]^T$是$d$维列向量($T$表示向量的转置)；

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

$\mathbf{\Sigma}$是$d\times d$协方差矩阵；

$\mathbf{\Sigma}^{-1}$是$\mathbf{\Sigma}$的逆矩阵，$|\mathbf{\Sigma}|$是$\mathbf{\Sigma}$的行列式。

多元正态分布

- 如$\mathbf{x}$服从多元正态分布，则有

$$\boldsymbol{\mu}=E[\mathbf{x}]=\int\mathbf{x}p(\mathbf{x})d\mathbf{x};$$

$$\mathbf{\Sigma}=E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T]=\int(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^Tp(\mathbf{x})d\mathbf{x};$$

- 具体的，如x_i是$\mathbf{x}$的第i个分量，μ_i是$\boldsymbol{\mu}$的第i个分量，σ_{ij}^2是$\mathbf{\Sigma}$的第ij个元素，则有

$$\mu_i=E[x_i]=\int x_ip(x_i)dx_i;$$

$$\sigma_{ij}^2=E[(x_i-\mu_i)(x_j-\mu_j)^T]$$

$$=\iint(x_i-\mu_i)(x_j-\mu_j)^Tp(x_i,x_j)dx_idx_j.$$

多元正态分布

- **协方差矩阵 Σ**
  - 是对称非负定阵，这里严格限定成正定阵，即 $|\boldsymbol{\Sigma}| > 0$。
  - 对角线元素 σ_{ii}^2 是 $\mathbf{x}$ 相应分量 x_i 的方差，即 σ_i^2。
  - 非对角线元素 σ_{ij}^2 是 x_i 和 x_j 的协方差，衡量了分量间的相关性。
    - 如果 x_i 和 x_j 统计独立，则 $\sigma_{ij}^2 = 0$。
    - 如 $\sigma_{ij}^2 = 0$（$i \neq j$），则 $p(\mathbf{x})$ 是 n 个各元素的单变量正态密度函数的内积。

$$\boldsymbol{\Sigma} = \begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & E[(x_1-\mu_1)(x_2-\mu_2)] \\ E[(x_2-\mu_2)(x_1-\mu_1)] & E[(x_2-\mu_2)(x_2-\mu_2)] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{bmatrix}.$$

多元正态分布的性质

1. 参数 μ 和 Σ 对分布具有决定性

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma});$$

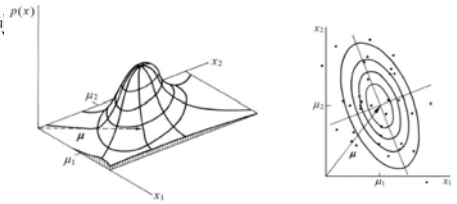
- 多元正态分布由 d+d(d+1)/2 个参数完全确定。

2. 等密度点的轨迹为超椭球面
   - $p(\mathbf{x})$ 是指数函数，因此等概率密度点对应于指数项为常数，即

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \text{常数};$$

多元正态分布的性质

2. 等密度点的轨迹为超椭球面
   - 在二维情况下，方程的解是一个椭圆轨迹，其长短轴方向由协方差矩阵 Σ 的特征向量决定；在三维时则是一个椭球面；超过三维则是超椭球面，主轴方向由协方差矩阵的特征向量决定，各主轴的长度则与相应的特征值成正比。
   - 例



多元正态分布的性质

2. 等密度点的轨迹为超椭球面
   - 马氏距离(Mahalanobis distance)：随机向量 $\mathbf{x}$ 偏离均值向量 $\boldsymbol{\mu}$ 的距离

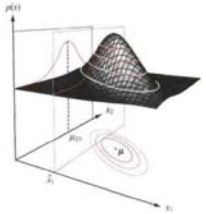
$$r = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})};$$

     - 在数理统计中，常用来确定未知样本集和已知样本集的相似性；考虑到各种特性之间的联系(c.f. 欧式距离)；与尺度无关，即独立于测量尺度；
     - 也可衡量两个服从同一分布并且其协方差矩阵为 Σ 的随机变量的差异程度。
       - $\boldsymbol{\Sigma} = \mathbf{I}$，即为欧式距离；
       - $\boldsymbol{\Sigma} = diag(\sigma_1^2, \cdots, \sigma_d^2)$，即为归一化的欧式距离。

多元正态分布的性质

3. 分布的离散程度
   - 由参数 $|\boldsymbol{\Sigma}|^{1/2}$ 决定，与单变量时由标准差 σ 决定相一致。
4. 边缘分布和条件分布的正态性
   - 多元正态分布的边缘分布和条件分布仍然是正态分布；

$$p(x_i) \sim N(\mu_i, \sigma_{ii}^2)$$



多元正态分布的性质

5. 不相关性等价于独立性
   - x_i 和 x_j 相互独立：$p(x_i, x_j) = p(x_i)p(x_j)$；
   - x_i 和 x_j 不相关：$E[x_i x_j] = E[x_i] \cdot E[x_j]$；
   - 如多元正态分布的任意两个分量互不相关，则它们一定独立。
   - 如多元正态随机向量 $\mathbf{x}$ 的协方差阵 Σ 是对角阵，则 $\mathbf{x}$ 各分量之间是相互独立的正态分布随机变量。

$$\sigma_{ij}^2 = 0, \ \forall i \neq j$$

$$\Rightarrow \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{nn}^2 \end{bmatrix}$$

$$\Rightarrow p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i)$$

多元正态分布的性质

6. 线性变换的正态性

- 多元正态分布的随机向量的线性变换仍然是多元正态分布的随机向量，即

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\mathbf{y} = \mathbf{A}^T\mathbf{x}, \mathbf{A} \in \Re^{d\times k}$$
$$\Rightarrow p(\mathbf{y}) \sim N(\mathbf{A}^T\boldsymbol{\mu}, \mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A});$$

  - 由于协方差矩阵$\boldsymbol{\Sigma}$是对称矩阵，因此总可以找到某个线性变换$\mathbf{A}$，使变换后的协方差矩阵$\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}$成为对角矩阵，这就意味着在某一个新的坐标系统中，可以做到使各分量之间相互独立。

多元正态分布的性质

6. 线性变换的正态性

- **白化变换**

$$\mathbf{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2} \Rightarrow p(\mathbf{y}) \sim N(\mathbf{A}_w^T\boldsymbol{\mu}, \mathbf{I});$$

  其中，矩阵$\boldsymbol{\Phi}$的列向量是$\boldsymbol{\Sigma}$的正交特征向量，矩阵$\boldsymbol{\Lambda}$由$\boldsymbol{\Sigma}$相应的特征根构成的对角矩阵。
  - 将任意的多元正态分布变换成球形分布，即变换后分布的协方差矩阵是单位矩阵。

- **线性组合的正态性**
  - 当 $k=1$ 时(即$\mathbf{A}$是 d 维向量 $\mathbf{a}$)，则 $y=\mathbf{a}^T\mathbf{x}$ 是一个标量，是 $\mathbf{x}$ 的线性组合

$$p(y) \sim N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a});$$

多元正态分布的性质

6. 线性变换的正态性

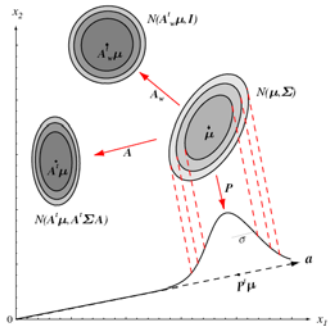


图 2-8 特征空间中的一个线性变换将一个任意正态分布变换成另一个正态分布。一个变换，$\mathbf{A}_w$，将原分布变成分布 $N(\mathbf{A}_w^t\boldsymbol{\mu}, \mathbf{A}_w^t\boldsymbol{\Sigma}\mathbf{A}_w)$；另一个线性变换，即由向量 $\mathbf{a}$ 决定的向某条直线的投影 $\mathbf{P}$，产生沿该直线方向的 $N(\mu, \sigma^2)$ 分布。尽管这些变换产生一个不同空间中的分布，我们还是将它们显示在原 x_1-x_2 空间中。一种白化变换，$\mathbf{A}_w$，将产生一个圆周对称的高斯分布