INT104W11_非监督学习2-多元高斯分布与GMM

高斯分布

> 高斯分布=正态分布

为什么许多机器学习算法都采用高斯分布?

机器学习中的许多模型的正确性都假设数据（近似）服从高斯分布，如逻辑回归，朴素贝叶斯和GMM。理论上当数据高斯分布的情况下，这些模型得到的结果才能说是稳健的。

自然界中的许多数据都呈现高斯分布；根据中心极限定理，当随机变量的数量足够大时，样本平均数的抽样分布也近似于正态分布；可以发现，现实中的很多随机变量是由大量相互独立的随机因素的综合影响所形成的，而其中每一个因素在总的影响中所起的作用都是微小的，故它们往往近似服从高斯分布。

从信息熵的角度来看，当数据的均值和方差一定的情况下，高斯分布是所有分布中信息熵最大的，也就是说高斯分布的假设可以让我们的假设最具有一般性，从最坏（最普适）的角度估计数据，得出结论。熵越大的高斯分布方差越大，在实轴上也越接近“均匀”。

涉及极大似然估计，EM算法的逻辑回归，朴素贝叶斯，GMM，k-means；涉及计算方差最大（最大熵分布）的PCA，LDA。它们都假设数据服从高斯分布。

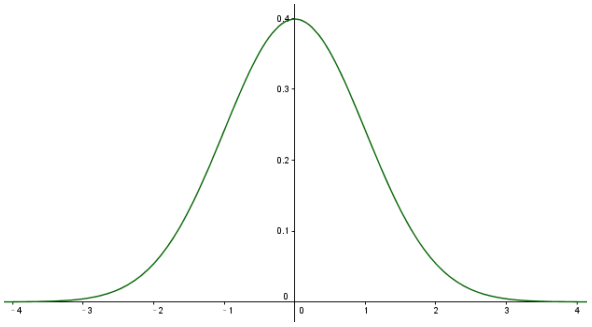
PS:消除偏度的一些方法

正态性是许多统计方法的重要前提假设；如果我们的数据不符合正态分布，强制开展统计分析结果可能会产生偏倚。为了使我们的数据趋向高斯分布，我们首先需要使数据对称，即消除偏度。

有一些可以使用的方法：对数变换，平方根变换，倒数变换，平方根正反旋变换，BOX-COX变换……

从高斯分布到多元高斯分布

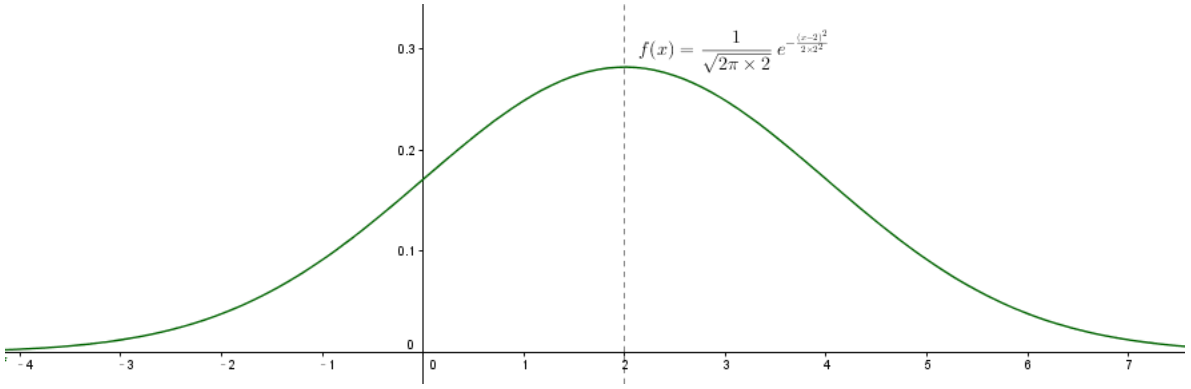
标准高斯函数（标准正态分布）



$$\text{标准高斯函数：}f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{1}$$

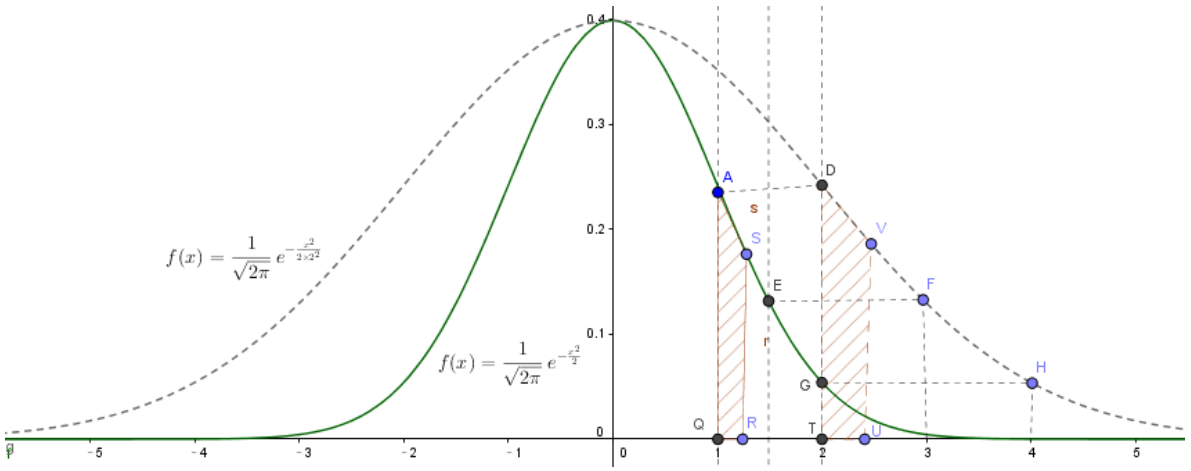
标准高斯函数均值=0，方差=1，概率密度和=1。

一元高斯函数一般形式



$$\text{一元高斯函数：}f_{\mu,\sigma}(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

可见，正态分布一共含有两个参数：μ和σ。μ是正态分布的均值，是分布的中心；σ是正态分布的标准差，越大分布越分散，反之集中，σ^2是方差。由此我们可将一个正态分布简单记为$N(\mu,\sigma^2)$。



可以通过标准化：令$z=\frac{x-\mu}{\sigma}$，将 x 向右移动 μ 个单位，再将密度函数伸展 σ 倍。把不同的正态分布转变为$N(0,1)$。

但需要注意，为了确保函数总面积始终为1，$\frac{1}{\sqrt{2\pi}\sigma}$分母中需添加σ。即图中x 轴方向做 σ 倍延拓的同时， y 轴应该压缩 σ 倍（乘以 1/σ）。

独立（维度不相关）多元正态分布

$$
\begin{aligned}
\text{若}n\text{个变量}x &= [x_1, x_2, ..., x_n]^T\text{互不相关且服从正态分布} && (3)\\
\text{各维度均值}E(x) &= [\mu_1, \mu_2, ..., \mu_n]^T && (4)\\
\text{各维度方差}\sigma(x) &= [\sigma_1, \sigma_2, ..., \sigma_n]^T && (5)\\
\text{根据联合概率密度公式：}\ f(x) &= p(x_1, x_2, ..., x_n) && (6)\\
&= p(x_1)p(x_2)...p(x_n) && (7)\\
&= \frac{1}{(\sqrt{2\pi})^n\sigma_1\sigma_2...\sigma_n}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}...-\frac{(x_n-\mu_n)^2}{2\sigma_n^2}} && (8)\\
\text{类似地，令}z^2 &= \frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}+...+\frac{(x_n-\mu_n)^2}{\sigma_n^2} && (9)\\
\sigma_z &= \sigma_1\sigma_2...\sigma_n && (10)\\
\text{可得：}\ f(z) &= \frac{1}{(\sqrt{2\pi})^n\sigma_z}e^{-\frac{z^2}{2}} && (11)
\end{aligned}
$$

用矩阵形式表示z^2：

$$
\begin{aligned}
z^2 &= \frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}+...+\frac{(x_n-\mu_n)^2}{\sigma_n^2} && (12)\\
&= z^Tz = \left[x_1-\mu_1, x_2-\mu_2, \cdots, x_n-\mu_n\right][\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, ..., \frac{1}{\sigma_n}]^T[\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, ..., \frac{1}{\sigma_n}]\left[x_1-\mu_1, x_2-\mu_2, \cdots, x_n-\mu_n\right]^T && (13)\\
&= \left[x_1-\mu_1, x_2-\mu_2, \cdots, x_n-\mu_n\right]\begin{bmatrix}\frac{1}{\sigma_1^2} & 0 & \cdots & 0\\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0\\ \vdots & \cdots & \cdots & \vdots\\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2}\end{bmatrix}\left[x_1-\mu_1, x_2-\mu_2, \cdots, x_n-\mu_n\right]^{\mathrm{T}} && (14)
\end{aligned}
$$

从矩阵的角度，$\sum$代表变量 X 的协方差矩阵， i行j列的元素值表示x_i与x_j的协方差（由于题设特征间相互独立，对角线外为0，本身的协方差就等于方差）。

$$记：\sum = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{15}$$

$$(\sum)^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} \tag{16}$$

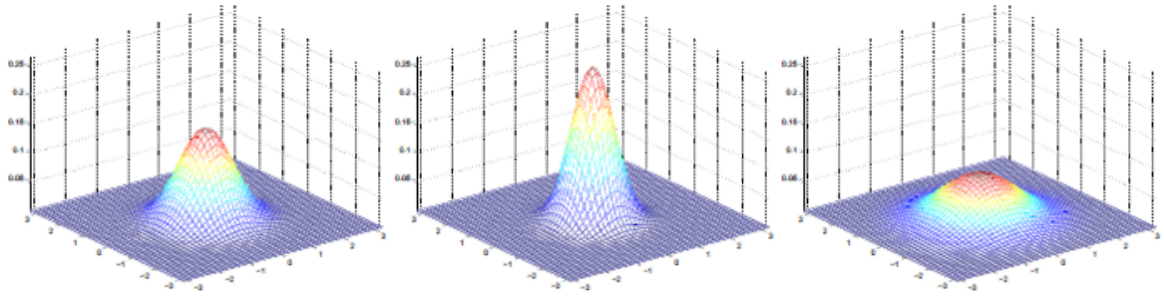
$$记 x - \mu_x = \left[x_1 - \mu_1, x_2 - \mu_2, \cdots, x_n - \mu_n\right]^{\mathrm{T}} \tag{17}$$

$$(14)简化为：\ z^2 = z^{\mathrm{T}} z = (x - \mu_x)^{\mathrm{T}} \sum^{-1} (x - \mu_x) \tag{18}$$

$$得：\ f(z) = \frac{1}{(\sqrt{2\pi})^n \left|\sum\right|^{\frac{1}{2}}} e^{-\frac{(x-\mu_x)^{\mathrm{T}} (\sum)^{-1} (x-\mu_x)}{2}} \tag{19}$$

从非标准正态分布→标准正态分布需要将概率密度函数的高度压缩$\left|\sum\right|^{\frac{1}{2}}$倍； 从一维→n维的过程中，每增加一维，高度将压缩$\sqrt{2\pi}$倍。

以二元分布函数为例，维度不相关高斯分布函数图像：二元正态分布（高斯分布）的等概率曲线是一个椭圆，而三元正态分布的等概率曲面是一个椭球。



二元分布，维度不相关高斯分布可视化

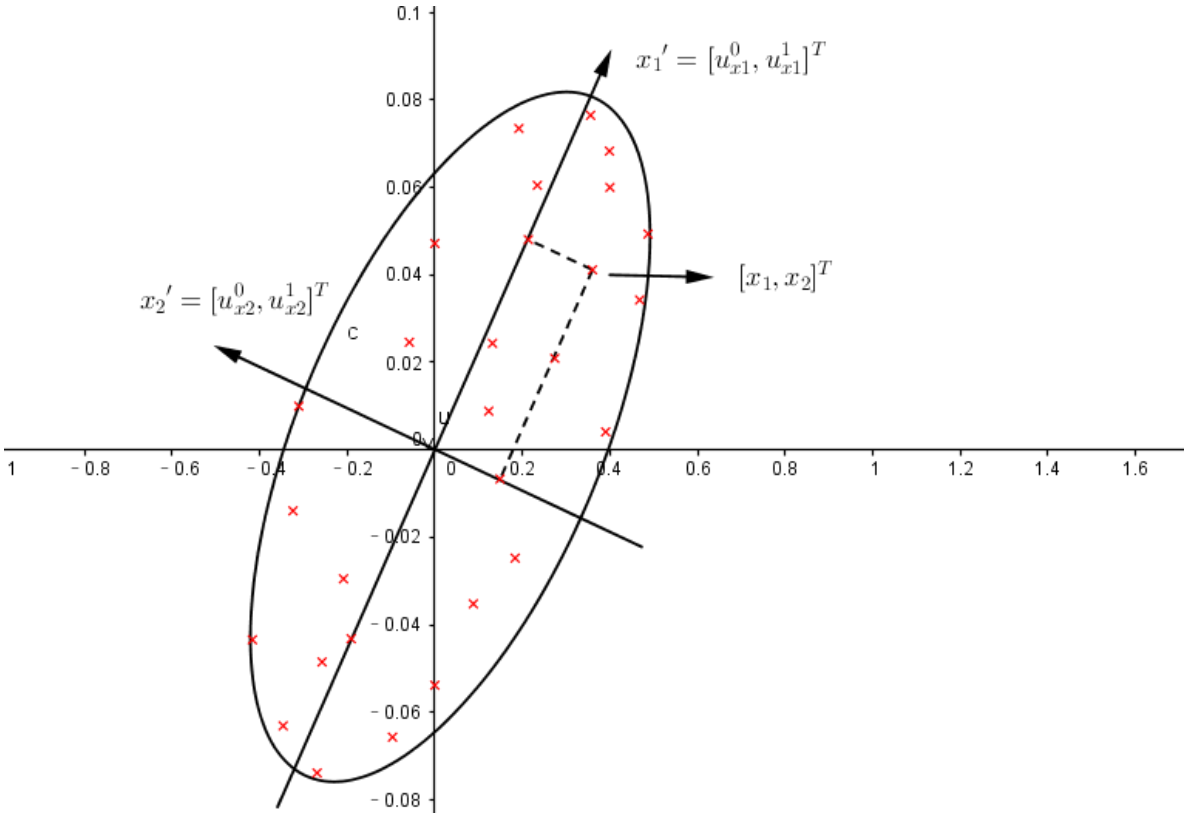
相关多元正态分布

当多元变量间相互关联

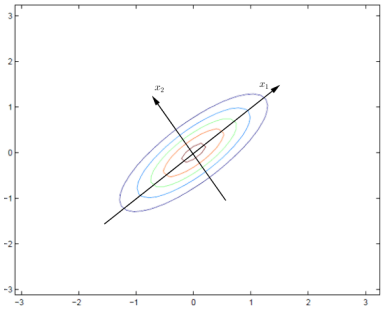
$$多元高斯分布概率密度函数：\ f_{\mu,\sum}(x) = \frac{1}{(\sqrt{2\pi})^D \left|\sum\right|^{\frac{1}{2}}} e^{-\frac{(x-\mu)^{\mathrm{T}} (\sum)^{-1} (x-\mu)}{2}}$$

- $f_{\mu,\Sigma}(x)$ 是表示多维高斯分布的概率密度函数，其中 x 是一个 D 维向量（$x \in \mathbb{R}^D$）。
- D 是数据的维度，表示向量 x 包含了 D 个随机变量。
- μ（mu）是一个 D 维向量，表示多维高斯分布的均值向量。
- Σ（Sigma）是一个 DxD 的协方差矩阵，表示不同维度之间的协方差关系。
- $(x-\mu)^T$ 表示向量 $(x-\mu)$ 的转置（transpose）。
- $|\Sigma|$ 表示协方差矩阵 Σ 的行列式（determinant）。
- Σ^{-1} 表示协方差矩阵 Σ 的逆矩阵（inverse）。

（以二元为例）



通过去相关性，在新的坐标系中，变量间就互不相关了。通过线性代数的手段，只需要旋转+平移。（有点类似于PCA中的计算）



$$\text{假设这个图中：协方差矩阵} \sum = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \tag{20}$$

$$\text{均值向量} \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{21}$$

$$(\sum)_{new} = U^T \sum U \tag{22}$$

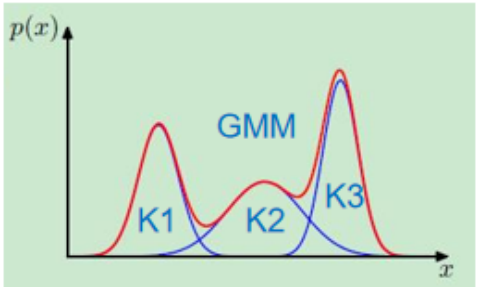
$$= \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \tag{23}$$

$$\Rightarrow \theta = \frac{\pi}{4} \tag{24}$$

$$(\sum)_{new} = \begin{bmatrix} 1.8 & 0 \\ 0 & 0.2 \end{bmatrix} \tag{25}$$

高斯混合模型（GMM）

高斯混合模型（Gaussian Mixture Model, GMM）该算法假设数据集中的所有数据点都由一个或多个高斯分布产生；认为每个高斯分布就是一个簇，代表了一类样本的分布，如下图中的数据集（a）你可以发现数据似乎是由3个不同位置的高斯分布混合得到的；那么GMM的任务就是将这个过程逆向，通过EM方法求解构成原数据集的每个高斯分布的参数，从而确定簇的位置与特性，完成聚类。



高斯混合模型是单一高斯机率密度函数的延伸

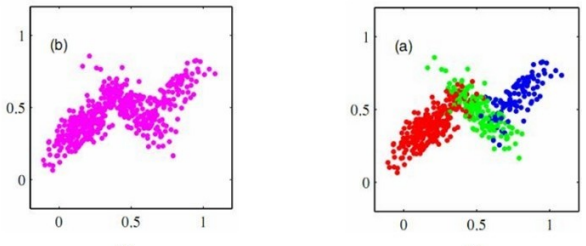


图 1 高斯混合模型图示，(a)表示所有样本数据；(b)表示已经明确了样本的分类

GMM能够平滑地近似任意形状的密度分布

聚类步骤

1. 初始化k个高斯分布$N(\mu_k, \sigma_k^2)$和对应分布在预测中的权重a_k
2. 使用EM算法更新每个高斯分布参数μ_k, σ_k与权重a_k

3. 重复步骤2，直到参数不再变化或达到最大迭代值（类似k-means终止条件，实际上k-means也是一种硬EM）

感性认识最大期望算法(Expectation-Maximum, EM)

给定数据集$D=\{x_1,x_2,...,x_n\}$，混合高斯分布$N(\mu_k,\sigma_k^2)$和对应权重a_k

$$对D中任意样本x_i的出现概率：\ p(x_i)=\sum_{k=1}^{k}a_kN(x_i;\mu_k,\sigma_k^2) \tag{26}$$

$$整个数据集出现的概率：\ T(x)=\prod_{i=1}^{n}p(x_i) \tag{27}$$

$$概率连乘过小，为防溢出两边取对数：\ L=\sum_{i=1}^{n}log(p(x_i)) \tag{28}$$

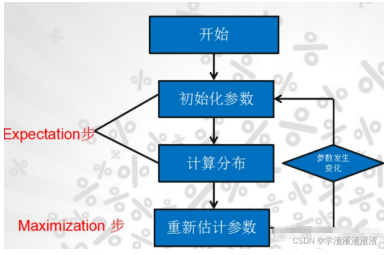
$$=\sum_{i=1}^{n}log(\sum_{k=1}^{k}a_kN(x_i;\mu_k,\sigma_k^2)) \tag{29}$$

$$先固定参数\mu_k,\sigma_k;更新a_k：\ max_{a_k}(L) \tag{30}$$

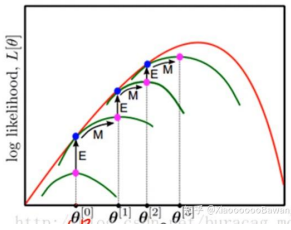
$$再固定参数a_k;更新\mu_k,\sigma_k：\ max_{\mu_k,\sigma_k}(L) \tag{31}$$

（26）~（29）属于E步骤（Expectation），计算联合分布的条件概率期望。

（30）~（31）属于M步骤（Maximization），通过调整参数最大化整个数据集出现的概率（此部分计算展开较复杂）。



EM算法迭代流程



EM算法迭代可视化，极大化似然函数

关于**隐变量**，推荐阅读这些文章：

如何感性地理解EM算法？ https://www.jianshu.com/p/1121509ac1dc

隐变量是什么? - 麋路的回答 - 知乎
https://www.zhihu.com/question/43216440/answer/156368711

机器学习中的隐变量和隐变量模型
https://blog.csdn.net/Ding_xiaofei/article/details/80207084

因子分析（一）- EM算法求解 - XiaooooooBawang的文章 - 知乎
https://zhuanlan.zhihu.com/p/512936256

总之，我们找到了最佳的模型参数，使得数据集出现的概率（期望）最大。

在GMM中，模型参数就是各个高斯分布的均值向量μ_k和协方差矩阵σ_k，它们可以确定每一个高斯分布，也就是聚类的簇。

参考

- 高斯混合模型（GMM）介绍以及学习笔记

  https://blog.csdn.net/jojozhangju/article/details/19182013

- 聚类分析Kmean,GMM,DBSCAN一网打尽 - 欹逸的文章 - 知乎

  https://zhuanlan.zhihu.com/p/695380904

- 【EM算法】期望最大化算法

  https://blog.csdn.net/weixin_42468475/article/details/123061007

- AI大语音（六）——混合高斯模型（GMM）（深度解析）

  https://blog.csdn.net/qq_42734492/article/details/108225927

EM算法理解的九层境界：（搞得玄之又玄的，但好像确实那么回事）

1. EM 就是 E + M
2. EM 是一种局部下限构造
3. K-Means是一种Hard EM算法
4. 从EM 到 广义EM
5. 广义EM的一个特例是VBEM
6. 广义EM的另一个特例是WS算法
7. 广义EM的再一个特例是Gibbs抽样算法
8. WS算法是VAE和GAN组合的简化版
9. KL距离的统一