再看高斯

emacsun

目录

高斯分布 在机器学习中是如此重要，我们必须对其进行透彻的理解和掌握。

1 定义

对于单变量随机变量高斯分布可以写为：

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1.1}$$

我们在 之前的博文 中证明过这个形式的分布确实是概率分布，此处略去证明。

对于D维的高斯随机矢量$\mathbf{x}$，则：

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} \tag{1.2}$$

其中μ是D维的均值矢量，$\mathbf{\Sigma}$是$D\times D$的协方差矩阵。$|\mathbf{\Sigma}|$是$\mathbf{\Sigma}$的行列式。

高斯分布如同幽灵一样无处不在。比如，对于单变量随机变量，最大化熵的分布是高斯分布，同样对于多变量随机矢量亦是如此。再比如，假设我们考虑多个随机变量和的分布，根据中心极限定理，满足一定条件后（这些条件都是很容易满足的条件），这些随机变量的和服从高斯分布。在机器学习中更是

如此。实际上，对高斯分布的处理需要相当的数学基础，但是在以后的学习中，我们会发现这种投入是值得的。

2 高斯分布的形状

对单变量高斯分布我们早已熟悉其钟形曲线。对于高斯矢量，我们从式(1.2)可以看到起形状依赖于二次型：

$$\Delta^2 = (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu) \tag{2.1}$$

这个Δ叫做μ到$\mathbf{x}$的 Mahalanobis距离。当$\mathbf{\Sigma}$是单位阵时，这个距离退化为欧几里得距离。当这个二次型是常量的时候，高斯分布在$\mathbf{x}$面上是常量。

不失一般性，对于$\mathbf{\Sigma}$我们可以假设其为实的对称的。对于实的对称的$\mathbf{\Sigma}$，其特征值也一定是实的。

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{2.2}$$

其中$i = 1, \ldots, D$。因为$\mathbf{\Sigma}$是实的对称的，所以其特征向量可以从一个正交基中选取，即：

$$\mathbf{u}_i^T \mathbf{u}_j = \sigma_{ij} \tag{2.3}$$

协方差矩阵$\mathbf{\Sigma}$可以表示为特征向量的扩展：

$$\mathbf{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \tag{2.4}$$

同理，协方差矩阵的逆矩阵也可以表示为：

$$\mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T \tag{2.5}$$

我们定义：

$$y_i = \mathbf{u}_i^T (\mathbf{x} - \mathbf{u}) \tag{2.6}$$

则式 (2.1)可以简化为：

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \tag{2.7}$$

我们可以把y_i解释为一个新的坐标系统，该坐标由正交矢量$\mathbf{u}_i$通过相对于原来的坐标x_i移位和旋转得到。为了生成$\mathbf{y} = (y_1, \ldots, y_D)^T$，我们有：

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \mu) \tag{2.8}$$

其中U是一个矩阵，其每一行是$\mathbf{u}_i^T$，根据式 (2.3)我们知道$\mathbf{U}$是一个正交矩阵，即$\mathbf{U}\mathbf{U}^T = \mathbf{I}$。从式 (2.6)我们可以知道二次型 (2.1) 在平面上是一个常量。如果所有的特征值λ_i是正数，那么这些二次型是椭圆, 其中心在μ,其轴的方向是$\mathbf{u}_i$的方向，轴的长度与特征值λ_i有关。如图1所示：

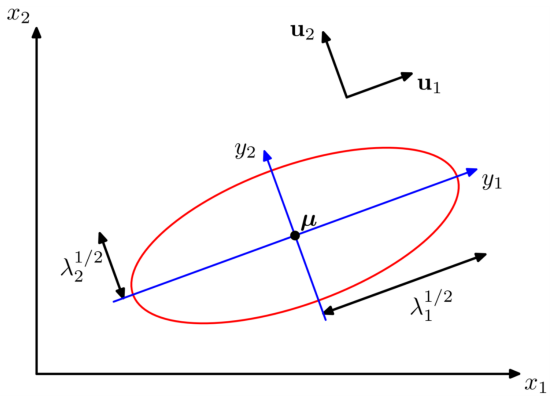

图 1: 二维高斯矢量在$\exp(-1/2)$时的图形

为了使得高斯随机变量有比较严格的定义，要求所有的特征值λ_i是正数，否则这个分布的积分就不会是1，也就是说这个分布没有归一化。对于非正定矩阵（有至少一个特征值为0或者负数），高斯分布会降维为低维的子空间。

现在我们在新的坐标系下考虑高斯分布。从$\mathbf{x}$到$\mathbf{y}$的转化过程，我们有一个Jacobian矩阵$\mathbf{J}$，

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji} \tag{2.9}$$

其中U_{ji}是矩阵$\mathbf{U}^T$的元素。根据$\mathbf{U}$的正交性，我们得到：

$$|\mathbf{J}|^2 = |\mathbf{U}^T|^2 = |\mathbf{U}^T||\mathbf{U}| = |\mathbf{I}| = 1 \tag{2.10}$$

所以：

$$|\mathbf{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2} \tag{2.11}$$

所以在y_j坐标系中，高斯分布的形式是：

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda)^{1/2}} \exp\left\{ -\frac{y_j^2}{2\lambda_j} \right\} \tag{2.12}$$

这个形式非常的简洁可以看做是多个独立的高斯随机变量的乘积。特征向量定义了移位和旋转的规则，基于这个规则联合概率密度分解为独立随机变量的乘积。显然：

$$\int p(\mathbf{y})\mathrm{d}\mathbf{y} = \prod_{j=1}^{D} \frac{1}{(2\pi)^{1/2}} \exp\left\{ -\frac{y_j^2}{2\lambda_j} \right\} dy_j = 1 \tag{2.13}$$

3　多变量高斯分布的期望和方差

接下来我们通过观察多变量高斯分布的矩赋予μ和$\mathbf{\Sigma}$特定的数学意义。高斯随便量$\mathbf{x}$的期望：

$$\mathbb{E}[x] = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}\mathbf{x}\mathrm{d}\mathbf{x} \tag{3.1}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{z})^T\mathbf{\Sigma}^{-1}(\mathbf{z})\right\}(\mathbf{x}-\mu)\mathrm{d}\mathbf{z} \tag{3.2}$$

我们实用了变量替换$\mathbf{z} = \mathbf{x} - \mu$。我们注意到指数函数部分是一个偶函数所以在积分过程中$\mathbf{z}+\mu$中的$\mathbf{z}$会变为0（奇函数在$(-\infty,\infty)$上的积分为0），只剩下μ，因此：

$$\mathbb{E}[\mathbf{x}] = \mu \tag{3.3}$$

我们称μ为$\mathbf{x}$的均值。接下来，我们考虑高斯分布的二阶矩。在单变量场景下，二阶矩定义为$\mathbb{E}[x^2]$，在多变量时，有D^2个二阶矩$\mathbb{E}[x_ix_j]$，我们把这些二阶矩合在一起构成一个矩阵$\mathbb{E}[\mathbf{x}\mathbf{x}^T]$，这个矩阵的计算过程为：

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}\mathbf{x}\mathbf{x}^T\mathrm{d}\mathbf{x} \tag{3.4}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{z})^T\mathbf{\Sigma}^{-1}(\mathbf{z})\right\}(\mathbf{z}+\mu)(\mathbf{z}+\mu)^T\mathrm{d}\mathbf{z} \tag{3.5}$$

这里我们再一次的实用了变量替换$\mathbf{z} = \mathbf{x} - \mu$。$(\mathbf{z}+\mu)(\mathbf{z}+\mu)^T$中的$\mu\mathbf{z}^T$和$\mu^T\mathbf{z}$在积分过程中会变为0，只剩下$\mu\mu^T$和$\mathbf{z}\mathbf{z}^T$。对于$\mu\mu^T$，这是个常量，在积分的过程中仍然会以常量的形式留下来。对于$\mathbf{z}\mathbf{z}^T$我们可以采用$\mathbf{z}$的特征向量展开的形式处理。

$$\mathbf{z} = \sum_{j=1}^{D} y_j\mathbf{u}_j \tag{3.6}$$

其中$y_j = \mathbf{u}_j^T\mathbf{z}$，则：

$$\frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{z})^T\mathbf{\Sigma}^{-1}(\mathbf{z})\right\}(\mathbf{z})(\mathbf{z})^T\mathrm{d}\mathbf{z} \tag{3.7}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathbf{u}_i\mathbf{u}_j^T\int \exp\left\{-\sum_{k=1}^{D}\frac{y_k^2}{2\lambda_k}\right\}y_iy_j\mathrm{d}\mathbf{y} \tag{3.8}$$

$$= \sum_{i=1}^{D}\mathbf{u}_i\mathbf{u}_i^T\lambda_i = \mathbf{\Sigma} \tag{3.9}$$

综上，我们有：

$$\mathbf{E}[\mathbf{x}\mathbf{x}^T] = \mu\mu^T + \mathbf{\Sigma} \tag{3.10}$$

对于单变量，我们通过$\mathbb{E}[(x-\mu)(x-\mu)]$来定义方差。同样的在多变量情况下，仍然可以减去军区，定义一个随机矢量的协方差：

$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T] \tag{3.11}$$

对于高斯分布，我们有：

$$\text{cov}[\mathbf{x}] = \boldsymbol{\Sigma} \tag{3.12}$$

又因为参数$\boldsymbol{\Sigma}$控制着$\mathbf{x}$的协方差。所以我们称$\boldsymbol{\Sigma}$为协方差矩阵。

4 高斯分布的局限

尽管高斯分布用途广泛，但其也不是放之四海而皆准的真理。高斯分布有其独有的缺陷。考虑一个一般的对称协方差矩阵$\boldsymbol{\Sigma}$，其有$D(D+1)/2$个独立的参数。对于μ，又有D个独立的参数。多以一共有$D(D+3)/2$个参数。当D变大时，其独立的参数个数以D的二次方增长。计算这么大矩阵的逆矩阵是一件非常困难的事情。解决办法？一个有效的办法是限制协方差矩阵的形式，如果我们只考虑$\boldsymbol{\Sigma}$是对角阵，即$\boldsymbol{\Sigma} = \text{diag}(\sigma_i^2)$,我们就只用关心$2D$个独立的参数。对应的概率密度图像是一系列同心的且轴与坐标轴平行的椭圆。当我们限制协方差矩阵$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$时，这个协方差矩阵叫做各向同性协方差。这个时候我们只需要处理$D+1$个独立的参数。三个类型的高斯分布，在固定概率密度下的形状如图2所示。

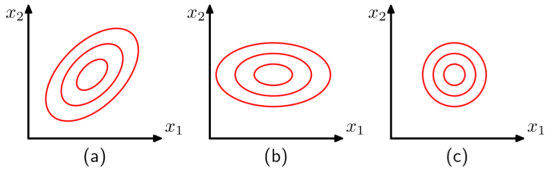

图 2: 高斯分布在固定概率时的形状

有得必有失。图2 中对高斯分布的近似限制了分布中的自由变量个数使得求逆变得异常简单，但是也限制了概率密度函数的形式，使我们很难捕捉到数据中的可能有用的相关性。

另一个高斯分布的限制是：高斯分布是单峰分布，只有一个最大值。所以高斯分布对于多峰分布的模拟能力有限。所以高斯分布一方面可能由于过多的参数而过于灵活，一方面由于其具有单封特性而只有有限的近似能力。我们在以后的学习中会看到，通过引入隐藏变量可以很好的解决这个问题。特别是，

通过引入离散的隐式变量到混合高斯模型（mixtures of Gaussians）中，我们可以获得丰富多彩的多峰分布（multimodal distributions）。通过引入连续的隐式变量，对模型的控制可以独立于D维的数据空间。实际上，通过引入离散的或者连续的隐式随机变量，我们获得了非常多的有用模型，这些模型在机器学习的方方面面都发挥了重要的作用。比如，基于高斯模型的马尔科夫随机场（Markov Random field, MRF）在图像处理方面的应用。比如，线性动态系统（linear dynamical system，LDS）在时间序列建模方面也非常重要。