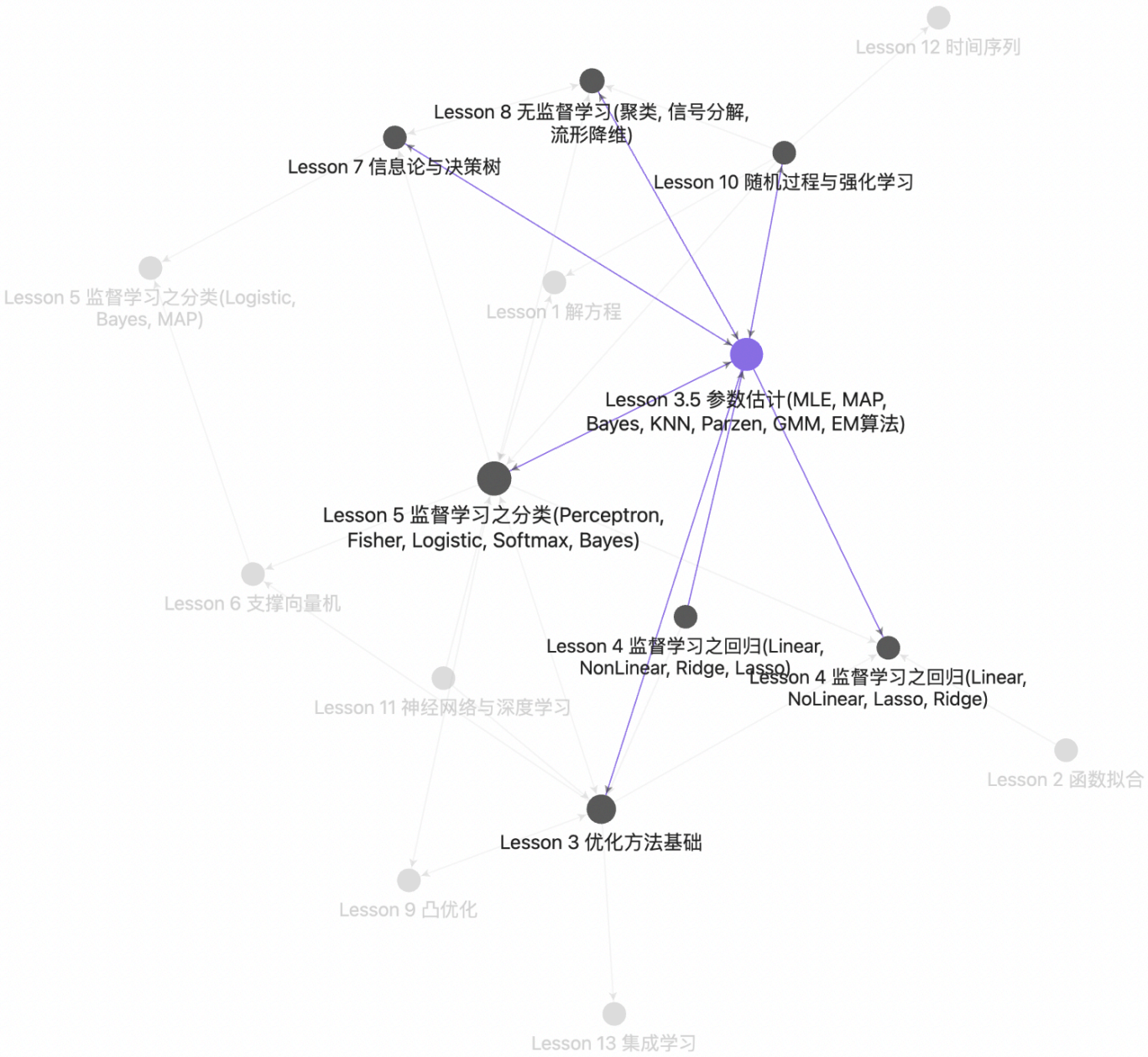
机器学习：概率密度函数的估计(Estimator of Probability Density Function)

第 3.5 章概率密度函数的估计，全部内容都是我之后加的，考虑第三章内容太多，所以就新开了一个 3.5 章，**这一章是我觉得非常非常重要的知识点**。夸张一点说，MLE 最大似然估计撑起了 ML 理论推导的半边天，所以这部分理解很重要，可以多花些时间。变分推断，GMM, EM 比较难，可以看完全部内容后再回来看。



数据科学的课堂上并没有这部分内容，而是把这部分穿插在了 4,5 两章，但是模式识别的老师讲了(虽然模式识别讲的顺序我觉得很有问题)。原则上是选看，但是我觉得不看 3.5 章理解 4,5 两章可能会存在一定困难，**尤其是经典的抛硬币问题，非常推荐大家在老师讲基本的回归分类算法前看一下**。变分推断会比较难，看不懂没关系，因为我也不太懂...

1. 概密函数估计概述

概率密度函数估计的思想，被用在非常多的机器学习目标函数推导中。在基础的线性回归及其正则化，逻辑回归，贝叶斯决策中，处处可见其影子。本章介绍的内容非常重要，是理解之后的监督学习中算法的必要基础。

我们通常假设某些训练集(随机变量)服从某种概率分布 $p(x)$ ，但其分布的参数 θ 是未知的。而参数估计算法，则需要根据一组服从此概率分布的样本，来估计出概率分布的参数。非参数估计算法，则不知道 $p(x)$ 的分布，通过不同的采样策略采样样本，计算采样的 $\tilde{p}(x)$ 来拟合 $p(x)$ 。

1.1. 算法分类

对于已知概率密度函数形式的问题(参数估计)，通过最大似然估计，最大后验估计，贝叶斯估计等算法可以解决。

对于不指定概率密度函数形式的问题(非参数估计)，通过 $K-$ 近邻，以及基于 $Parzen$ 窗的 $Mean-Shift$ 算法等方法可以解决。

1.2. 独立同分布假设(IID)

对于两个随机变量 X,Y ，如果 $f(x,y)=f_X(x)f_Y(y)$ 几乎处处成立(不成立点为有限集或无限不可数集)，则称它们相互独立。

对于 n 维随机向量 $x=(X_1,\cdots,X_n)$ ，如果 $f(x)=f_{X_1}(x_1)\cdots f_{X_n}(x_n)$ 几乎处处成立，则称它们相互独立。

如果一组随机变量之间相互独立，且服从同一种概率分布，则称它们独立同分布($Independent\ And\ Identically\ Distributed, IID$)。

在机器学习中，一般假设各个样本之间独立同分布。如果样本集(观测数据) $x_i, i=1,\cdots,l$ 独立同分布，均服从概率分布 $p(x)$ ，则 $p(x,\theta)=p(x)p(\theta)$， $\therefore p(x|\theta)=p(x)$ 。因此它们的联合概率(似然概率)为：

$$p(x_1,x_2,\cdots,x_l)=\prod_{i=1}^{l}p(x_i|\theta)=\prod_{i=1}^{l}p(x_i)$$

即直接让每次观测到的数据相乘，就得到了似然概率。在概密函数估计，以及各种 ML, DL 算法中，经常使用这个假设，以简化联合概率的计算。

2. 最大似然估计(MLE)

2.1. 原理

最大似然估计($Maximum\ Likelihood\ Estimation, MLE$)为样本构造一个似然函数，通过让似然函数最大化，求解出参数 θ 。

最大似然估计的直观解释为，寻求参数的值使得给定的样本集出现的概率(或概密函数值)最大。有监督学习问题从概率的角度便可以这样描述，即当给出了标签和训练集 (y,X) ，求使得 $p(y|X,\theta)$ 最大的参数 θ 。在Lesson 4 监督学习之回归(Linear, NonLinear, Lasso, Ridge, Generalization)线性回归的均方误差，以及Lesson 5 监督学习之分类(Perceptron, Fisher, Logistic, Softmax, Bayes)逻辑回归的交叉熵误差推导中，目标函数本质上就是一个最大似然过程。

最大似然估计认为使得观测数据(样本集)出现概率最大的参数就是模型的最优参数，这一方法也体现了"存在即合理"地朴素哲学思想：既然这组样本出现了，那么它们出现的概率理应是最大化的。

假设样本服从概率分布为 $p(x;\theta)$ ，其中 x 为随机变量，θ 为需要估计的参数。给定样本集 $x_i, i=1,\cdots,l$ ，已知 $p(x|\theta)$ 是给定参数 θ 时，样本满足的概率分布，也就是样本 x 的概率函数(似然函数)。假设样本满足独立同分布假设，其多个样本的联合概率函数为：

$$\prod_{i=1}^{l}p(x_i;\theta)$$

这个联合概率就是所有样本的似然函数，其中 x_i 为已知量，θ 为待确定的未知数，似然函数是优化变量 θ 的函数：

$$L(\theta)=\prod_{i=1}^{l}p(x_i;\theta)$$

目标是让这个函数的值最大化，这样做的依据是这组样本出现了，因此应该最大化它们出现的概率。即求解如下的最优化问题：

$$\max_{\theta}\prod_{i=1}^{l}p(x_i;\theta)$$

求解驻点方差可以得到问题的解。但是乘积求导不易处理且容易造成浮点数溢出，所以将似然函数取对数，得到对数似然函数：

$$\ln L(\theta)=\ln\prod_{i=1}^{l}p(x_i;\theta)=\sum_{i=1}^{l}\ln p(x_i;\theta)$$

它是一个增函数，因此最大化似然函数等价于最大化对数似然，最后求解的最优化问题为：

$$\max\sum_{i=1}^{l}\ln p(x_i;\theta)$$

这是一个不带约束的最优化问题，一般情况下可以直接求得解析解，也可以用梯度下降或者牛顿法来求解。对于离散型或者连续型概率分布，这种处理方法是统一的。

一般情况下似然函数是凹函数，因此有全局极大值点。

2.2. 伯努利分布最大似然参数估计

对于伯努利分布 $Bern(p)$，有 n 个样本，取值为 1 的有 a 个，0 的有 $n-a$ 个，因此样本集的似然函数为：$L(p)=p^a(1-p)^{n-a}$。

其对数似然函数为：

$$\ln L(p)=a\ln p+(n-a)\ln(1-p)$$

对 p 求导并令导数为 0，可以得到：

$$\frac{a}{p}-(n-a)\frac{1}{1-p}=0\rightarrow p=\frac{a}{n}$$

2.3. 正态分布最大似然参数估计

对于正态分布 $N(\mu,\sigma^2)$：

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

有样本集 $x_1,\cdots,x_n$。其似然函数为：

$$L(\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x_i-\mu)^2}{2\sigma^2})=(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2)$$

其对数似然函数为：

$$\ln L(\mu,\sigma)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

对 μ,σ 求偏导并令其为 0：

$$\begin{cases}\frac{\partial \ln L(\mu,\sigma)}{\partial \mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0\\ \frac{\partial \ln L(\mu,\sigma)}{\partial \sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}$$

解得：

$$\mu=\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\quad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

这和正态分布的结论是类似的，其最大似然估计得到的均值是样本集的均值，方差为样本集的方差。

对于 n 维正态分布 $N(\mu,\Sigma)$：

$$p(x)=\frac{1}{(2\pi)^{\frac{2}{n}}|\Sigma|^{\frac{1}{2}}}\exp(\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

给定一组样本 $x_1,\cdots,x_l$，其对数似然函数为：

$$\begin{aligned}\ln L(\mu,\Sigma)&=\ln\prod_{i=1}^{l}\frac{1}{(2\pi)^{\frac{2}{n}}|\Sigma|^{\frac{1}{2}}}\exp(\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))\\&=-\frac{nl}{2}\ln(2\pi)-\frac{l}{2}\ln|\Sigma|-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))\end{aligned}$$

对 μ 求梯度并令梯度为 0：

$$\nabla_\mu \ln L=\sum_{i=1}^{l}\Sigma^{-1}(x_i-\mu)=0$$

两边左乘 Σ，可以解得：

$$\mu=\frac{1}{l}\sum_{i=1}^{l}x_i$$

对 Σ 的求解更加复杂，因为他要满足对称正定性的约束条件，可以解得：

$$\Sigma=\frac{1}{l}\sum_{i=1}^{l}(x_i-\mu)(x_i-\mu)^T$$

这与一维正态分布的最大似然估计结果在形式上是统一的。

3. 最大后验估计(MAP)

3.1. 原理

最大似然估计将参数 θ 看做确定值(待优化的普通变量)，通过最大化对数似然函数来确定其最优值。而最大后验估计($Maximum\ A\ Posteriori\ Probability\ Estimation, MAP$) 则将 θ 看做随机变量，假设 θ 服从某种概率分布，通过最大化后验概率 $p(\theta|x)$ 来确定其值，其中心思想是是的在样本出现的条件下，参数的后验概率最大化。求解时需要假设参数 θ 服从某种概率分布(即引入了 θ 的先验分布)。

通俗地说，最大似然估计优化的是似然函数 $p(x|\theta)$ ，而最大后验估计优化的是函数 $p(x|\theta)p(\theta)$ 。$p(\theta)$ 为先验知识，也就是 θ 服从的分布。 L_1L_2 正则化采用的就是最大后验估计的思想，即在线性回归 / 逻辑回归最大似然的基础上，引入了 w_i 的先验知识，最大化后验概率。

假设参数服从概率分布 $p(\theta)$ 。根据贝叶斯公式，参数对样本集的后验概率(即已知样本集 x 的条件下参数 θ 的条件概率)为：

$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{p(x)}=\frac{p(x|\theta)p(\theta)}{\int_\theta p(x|\theta)p(\theta)\mathrm{d}\theta}$$

$p(x|\theta)$ 即给定参数值时样本的概率分布，就是 x 的概率密度函数或概率质量函数，可以根据样本的值 x 进行计算，与最大似然估计是一致的。θ 非确定值，而是随机变量。根据Lesson 5 监督学习之分类(Perceptron, Fisher, Logistic, Softmax, Bayes)的贝叶斯决策的思想，可以简化分母计算(与 θ 无关)，因此最大后验概率等价于：

$$\arg\max_\theta p(\theta|x)\Leftrightarrow\arg\max_\theta p(x|\theta)p(\theta)$$

如果 θ 服从均匀分布，该项为常数，最大后验估计等价于最大似然估计。其他的主要区别就是先验项 $p(\theta)$ 。

3.2. 伯努利分布最大后验参数估计

对于上面的伯努利分布，同样假设有 n 个样本，取值为 1 的有 a 个，0 的有 $n-a$ 个。引入先验，假设参数 p 服从正态分布 $N(0.3,0.1^2)$ ，则目标函数为：

$$L(p)=p^a(1-p)^{n-a}\frac{1}{\sqrt{2\pi}\times0.1}\exp(-\frac{(p-0.3)^2}{2\times0.1^2})$$

其对数为：

$$\ln L(p)=a\ln p+(n-a)\ln(1-p)+\ln\frac{1}{\sqrt{2\pi}\times0.1}-50(p-0.3)^2$$

对 p 求导并令导数为 0：

$$\frac{a}{p}-\frac{n-a}{1-p}-100(p-0.3)=0$$

可以得到 p 的值。

3.3. 正态分布最大后验参数估计

假设有正态分布 $N(\mu,\sigma_v^2)$，均值 μ 未知，方差 σ_v^2 已知。由一组采样自该分布的独立同分布样本 $x_1,\cdots,x_l$。假设不确定参数 μ 服从正态分布 $N(\mu_0,\sigma_m^2)$，最大后验概率估计的目标函数为：

$$L(\mu)=\frac{1}{\sqrt{2\pi}\sigma_m}\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_m^2})\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_v}\exp(-\frac{(x_i-\mu)^2}{2\sigma_v^2})$$

其对数为：

$$\ln L(\mu)=\ln\frac{1}{\sqrt{2\pi}\sigma_m}-\frac{(\mu-\mu_0)^2}{2\sigma_m^2}+n\ln\frac{1}{\sqrt{2\pi}\sigma_v}-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma_v^2}$$

最大化此目标函数等价于最小化如下函数：

$$f(\mu)=\frac{(\mu-\mu_0)^2}{2\sigma_m^2}+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma_v^2}$$

对 μ 求导并令导数为 0，可以解得；

$$\mu=\frac{\sigma_m^2(\sum_{i=1}^{n}x_i)+\sigma_v^2\mu_0}{\sigma_m^2 n+\sigma_v^2}$$

这就是均值的最大后验估计结果。如果忽略上式的分子和坟墓的第二部分，及假设 μ 的方差 σ_v 为 0，则均值与最大似然估计的结果相同，此时 μ 退化成一个确定值。

方差的最大后验估计的计算过程，和均值类似。只是此时均值是确定值，方差满足某种概率分布。

4. 贝叶斯估计

贝叶斯估计和最大后验概率估计的思想类似，区别在于它需要考虑分母的积分结果，并且它不求出参数具体的值的表达式，而是求出参数所服从的概率分布。

注意贝叶斯决策(分类)和贝叶斯估计的区别。贝叶斯决策的训练部分，根据训练集给出的 θ 和 x 来得到模型的分布和概密函数，测试部分，在测试集上比较样本对于不同类的后验概率大小来进行分类，可以不考虑分母的积分。贝叶斯估计是参数估计问题，根据给出的 θ 的先验分布和似然函数，可以得到 θ 所服从的后验概率分布，需要考虑分母的积分。

参数 θ 的后验分布为：

$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{\int_\theta p(x|\theta)p(\theta)\mathrm{d}\theta}$$

$p(\theta)$ 为先验分布，$p(x|\theta)$ 为给定参数时样本的概率分布。考虑分母的全概率公式，这里得到的是参数的概率分布，通常取其数学期望作为参数的估计值：

$$E[p(\theta|x)]$$

可以通过变分概率来近似计算参数的后验概率。

5. 变分推断($Variational\ Inference$)

概率图模型常用于通过已知变量向未知变量的推断，这部分介绍的变分推断，多种采样算法，以及Lesson 10 随机过程与强化学习介绍的的隐马尔科夫模型，都是典型的概率图模型。常用于计算隐变量。

统计推断就是在根据观测变量 x 计算隐变量 z 的条件概率，即后验概率 $p(z|x)$ 。根据贝叶斯公式：

$$p(z|x)=\frac{p(x|z)}{p(x)}$$

贝叶斯决策和贝叶斯估计都是基于这个式子。上式的分母计算复杂，离散型概率需要对隐变量求和，连续型概率需要通过对联合概密函数 $p(x,z)$ 求积分得到边缘概率 $p(x)$ 。即：

$$p(z|x)=\frac{p(x|z)p(z)}{\int_z p(x|z)p(z)\mathrm{d}z}$$

前面提到，这里的困难是分母的积分很难计算，尤其是 z 为高维向量、联合概率函数而不是某些特定类型的函数的情况下，这个多重积分难以得到解析解。数值积分法计算高维积分也存在困难。因此可以通过依赖随机数采样的蒙特卡洛算法，和不依赖于随机数的变分推断，对分母近似求解。

变分推断又称为变分贝叶斯，他构造需要计算的概率分布 $p(x)$ 的一个近似分布 $q(x)$ ，最小化二者的 KL 散度(Lesson 7 信息论与决策树)以得到 $q(x)$ ，此即原本需要计算的概率分布的近似值。 即用一个变分分布来近似隐变量 z 的条件概率。

$$q(z)\approx p(z|x)$$

用 KL 散度衡量二者的分布差异。变分推断的目标即找到一个概率分布 $q(z)$，使得它和要计算的后验概率分布 $p(z|x)$ 的 KL 散度最小化：

$$\min_q D_{KL}(q(z)||p(z|x))$$

$\because p(z|x)=p(x,z)/p(x)$ ，并且根据第七章 KL 散度的定义有：

$$D_{KL}(q(z)||p(z|x))=\int_z q(z)\ln\frac{q(z)}{p(z,x)/p(x)}\mathrm{d}z=\int_z q(z)\Big(\ln\frac{q(z)}{p(z,x)}+\ln p(x)\Big)\mathrm{d}z$$

$p(x)$ 与 z 无关，$q(z)$ 作为概率函数积分为 1 ，因此有：

$$\int_z q(z)\ln p(x)\mathrm{d}z=\ln p(x)$$

上式变形为：

$$D_{KL}(q(z)||p(z|x))=\int_z q(z)\ln\frac{q(z)}{p(z,x)}\mathrm{d}z+\ln p(x)$$

继续变形可以得到：

$$\ln p(x)=D_{KL}(q(z)||p(z|x))-\int_z q(z)\ln\frac{q(z)}{p(z,x)}\mathrm{d}z=D_{KL}(q(z)||p(z|x))+L(q(z))$$

其中，$L(q(z))$ 称为变分下界函数，或者证据下界($Evidence\ Lower\ Bound, ELBO$)，进一步可以分解为：

$$\begin{aligned} L(q(z)) &= -\int_z q(z)\ln\frac{q(z)}{p(z,x)}\mathrm{d}z = -\int_z q(z)\ln\frac{q(z)}{p(z)p(x|z)}\mathrm{d}z \\ &= -\int_z q(z)\Big(\ln\frac{q(z)}{p(z)} + \ln\frac{1}{p(x|z)}\Big)\mathrm{d}z \\ &= E_{q(x)}[\ln p(x|z)] - D_{KL}(q(z)||p(z)) \end{aligned}$$

这也是变分自编码器($Variational\ Auto-Encoder$)中的损失函数表达。上式中，$E_{q(x)}[\ln p(x|z)]$ 为自动编码器 $Auto-Encoder$ 中的重建损失，$D_{KL}(q(z)||p(z))$ 即学习 $p(z)$ 和限定分布 $q(z)$ 的相似性。

而 $p(x|z)$ 和 $p(z)$ 通常易于计算。所以根据上式，最小化 $D_{KL}(q(z)||p(z|x))$ 等价于最大化 $L(q(z))$ 。可以限定 $q(z)$ 的分布，比如说正态分布，然后优化 $L(q(z))$ 。这样就将泛函优化问题转为了一个函数优化问题，优化变量为概率分布 $q(z)$ 的参数。

$q(z)$ 使用正态分布是因为它的支撑区间是 $\mathbb{R}^n$，在整个区间的概率密度函数值非 0，且两个正态分布之间的 KL 散度可以得到解析解。

在Lesson 7 信息论与决策树证明了 $D_{KL}\geqslant 0$，当且仅当两个概率分布完全一致时，值为 0，所以 $L(q(z))$ 是 $\ln p(x)$ 的变分下界：

$$\ln p(x) \geqslant L(q(z))$$

总结：变分推断的目标是使已知分布的 $q(z)$ 能接近想要的后验概率分步 $p(z|x)$，等价于最小化 $q(z), p(z|x)$ 之间的 KL 散度。而最小化散度等价最大化 $ELBO, L(q(z))$，这个优化问题通常更好求解。这也是变分自编码器的思路。

最大化 $ELBO$ 可以最大化对数似然 $\ln p(x)$ 的值，以实现最大似然估计。这也是 EM 算法的思路，$ELBO$ 项的 $-E_{q(z)}\left[\ln\frac{q(z)}{p(z,x)}\right]$ 即为 EM 算法 E 步中构造的数学期望 $E_{Q(x)}\left[\ln(\frac{p(x,z;\theta)}{Q(z)})\right]$ 。

6. 经典的抛硬币例题(三个参数估计算法的关系)

假设一个抛硬币实验，假设正面向上设为 U 的概率为 ρ，反面向上设为 D 概率为 $(1-\rho)$。我们进行了 3 次实验，得到两次正面，一次反面，即序列为 $'UUD'$ 。

1. 使用最大似然估计求 ρ 的估计值。
2. 假设 ρ 的先验概率是服从 $f(\rho)=6\rho(1-\rho)$，用最大后验概率估计 ρ 的值。
3. 假设 ρ 的先验概率是服从 $f(\rho)=6\rho(1-\rho)$，求贝叶斯估计的 ρ 的分布。

解：贝叶斯公式为：

$$p(\rho|x) = \frac{p(x|\rho)p(\rho)}{\int_\rho p(x|\rho)p(\rho)\mathrm{d}\rho}$$

样本的条件概率密度为满足伯努利分布，即：

$$p(x|\rho) = \begin{cases} \rho, & x\in U \\ 1-\rho, & x\in D \end{cases}$$

假设样本满足独立同分布假设，即 $p(x,\rho)=p(x)p(\rho)$，$\therefore p(x|\rho)=p(x)$ 。则根据观测数据，可计算联合概率分布(条件似然概率)为：

$$L(\rho)=\prod_{i=1}^{3}p(x_i|\rho)=\prod_{i=1}^{3}p(x_i)=\rho\times\rho\times(1-\rho)$$

1).最大似然估计，即求解似然概率取最大值时，ρ 的值：

$$\arg\max_{\rho}p(x|\rho)=\arg\max_{\rho}L(\rho)=\arg\max_{\rho}\rho^2(1-\rho)$$

令 $\frac{\partial L(\rho)}{\partial\rho}=0$，得到 $\hat{\rho}_{MLE}=\frac{2}{3}$ 。

2).最大后验估计，即在最大似然估计基础上，引入 ρ 的先验分布。或者说在贝叶斯估计的基础上，简化分母求解：

$$\begin{aligned}\arg\max_{\rho}p(\rho|x)&\Leftrightarrow\arg\max_{\rho}p(x|\rho)p(\rho)\\&=\arg\max_{\rho}L(\rho)f(\rho)\\L(\rho)f(\rho)&=\rho^2(1-\rho)\cdot6\rho(1-\rho)=6\rho^3(1-\rho)^2\end{aligned}$$

令 $\frac{\partial L(\rho)f(\rho)}{\partial\rho}=0$，得到 $\hat{\rho}_{MAP}=\frac{3}{5}$ 。

3).贝叶斯估计，即在最大后验估计基础上，考虑分母积分，不求出 ρ 的确切值，而值求出参数所服从的概率分布：

$$p(\rho|x)=\frac{p(x|\rho)p(\rho)}{\int_\rho p(x|\rho)p(\rho)\mathrm{d}\rho}=\frac{6\rho^3(1-\rho)^2}{\int_\rho 6\rho^3(1-\rho)^2\mathrm{d}\rho}=\frac{\rho^3(1-\rho)^2}{\int_\rho\rho^3(1-\rho)^2\mathrm{d}\rho}$$

这就是 ρ 的分布表达式，可证明 ρ 服从 $Beta$ 分布。

对于 $Beta$ 分布：

$$p(\rho|\alpha,\beta)=\frac{1}{\mathrm{B}(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1},\quad\text{其中 }\mathrm{B}(\alpha,\beta)=\frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}$$

所以，当 $\alpha=2,\beta=2$ 时，$p(\rho)=\frac{1}{\frac{1}{6}}\rho(1-\rho)$，因此可证先验概率 $p(\rho)$ 满足 $Beta$ 分布。

求解分母积分：

$$\int_0^1\rho^3(1-\rho)^2\mathrm{d}\rho=\frac{1}{60}$$

所以：

$$p(\rho|x)=\frac{\rho^3(1-\rho)^2}{\frac{1}{60}}\sim Beta(\rho|4,3)$$

因此，根据 $Beta$ 分布的公式，此时根据样本估计的 $\hat{\rho}_{Bayes}=\frac{\alpha}{\alpha+\beta}=\frac{4}{7}$，同时可证参数 ρ 服从 $Beta$ 分布。

7. 非参数密度估计方法的选择

前面介绍的参数估计方法需要已知概率密度函数的形式，算法值确定概率密度函数的参数。但是对于很多应用情况，我们无法给出概率密度函数的显式表达式，所以此时可以用概率密度的非参数估计法。一般有两种方法，$Parzen$ 窗法和近邻算法。

非参数估计方法主要解决的问题是，从服从 $p(x)$ 分布的总体中抽取一定样本，使得估计出的 $\hat{p}(x)$ 能近似收敛到 $p(x)$，即：

$$P=\int_{\mathbb{R}^d}p(x)dx=\hat{p}(x)V$$

其中，$\hat{p}(x)=\frac{k}{nV}$。n 为样本点总数，V 为包含 x 的一个小区域 $\mathbb{R}$ 的体积，k 为落在此区域里的样本数。

设有一系列包含样本的区域 $\mathbb{R}_1,\cdots,\mathbb{R}_n$，某个区域 $\mathbb{R}_n$ 包含 k_n 个样本，V_n 为 $\mathbb{R}_n$ 的体积，$\hat{p}_n(x)=\frac{k_n}{nV_n}$ 为 $p(x)$ 的第 n 次估计。

如果 $\lim\limits_{n\to\infty}V_n=0$，$\lim\limits_{n\to\infty}k_n=\infty$，$\lim\limits_{n\to\infty}\frac{k_n}{n}=0$，则可以认为 $\hat{p}_n(x)\to p(x)$。

针对非参数密度估计，有两种选择策略，固定体积 V_n，和固定某个 x_i 的近邻 k_n 的个数。

$Parzen$ 窗法通过固定窗口大小，即 $V_n=\frac{1}{h^d}$，同时使用不同的核函数 K 对窗中样本个数的 k_n 和 $\frac{k_n}{n}$ 加以限制，以保证收敛。

而 K 近邻则不同，它通过固定每个窗中拥有样本的个数 k_n，让窗口大小 V_n 刚好包含样本 x_i 的 k_n 个近邻，以进行非参数密度估计。

8. 核密度估计(KDE，$Parzen$ 窗法)

核密度估计($Kernel\ Density\ Estimation,KDE$)也称为 $Parzen$ 窗技术，是一种典型的非参数估计法。它无需求解概率密度函数的参数，而是用一组标准函数的叠加来表示概率密度函数。

有 d 维空间中的样本点 $x_i,i=1,\cdots,n$，它们服从某一未知的概率分布。给定核函数 $K(x)$，在任一点 x 处的概率密度函数的估计值根据所有的样本点计算：

$$p(x)=\frac{1}{nh^d}\sum_{i=1}^{n}K(\frac{x-x_i}{h})$$

h 是核函数的窗口半径，是人工设定的正参数。核函数要保证函数值 $K(\frac{x-x_i}{h})$ 随着待估计点 x 离样本点 x_i 的距离增加而递减。根据这一原则，如果 x 附近的样本点秘籍，则该点处的概率密度函数的估计值更大；如果 x 附近的样本点稀疏，则概率密度函数的估计值小。这也符合对概率密度函数的直观要求。系数 $\frac{1}{nh^d}$ 是为了确保 $p(x)$ 的积分为 1，使得它是一个合法的概率密度函数。其中 $\frac{1}{n}$ 对应着 n 个求和项；$\frac{1}{h^d}$ 是为了确保核函数进行了 d 维换元之后的积分制为 1，即：

$$\int_{\mathbb{R}^d}\frac{1}{h^d}K(\frac{x-x_i}{h})\mathrm{d}x=1$$

核函数能确保积分值为 1。

$$\int_{\mathbb{R}^d} K(y)\mathrm{d}y = 1$$

证明：如果令 $\frac{x-x_i}{h} = y$，其逆变换为 $x = hy + x_i$，此换元的雅克比行列式(每个元素为对应的元素偏导)

$$|\frac{\partial x}{\partial y}| = \begin{vmatrix} h & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & h \end{vmatrix} = h^d$$

使用雅克比行列式计算多重积分的换元法，因此有：

$$\frac{1}{h^d}\int_{\mathbb{R}^d} K(\frac{x-x_i}{h})\mathrm{d}x = \frac{1}{h^d}\int_{\mathbb{R}^d} K(y)|\frac{\partial x}{\partial y}|\mathrm{d}y = h^d\int_{\mathbb{R}^d} K(y)h^d\mathrm{d}y = 1$$

8.1. 常用窗函数

常用核函数是径向对称核($Radially\ Symmetric\ Kernel$)：

$$K(x) = c_{k,d}k(||x||^2)$$

其中 $k(x)$ 为核的剖面($profile$)函数，是 $||x||$ 的减函数且对点 x 关于原点径向对称，这也是径向对称核这一名字的又来。归一化常数 $c_{k,d}$ 确保 $K(x)$ 的积分为 1，此常数根据具体的核函数而定。

$Epanechnikov$ 剖面函数定义为：

$$k(x) = \begin{cases} 1-x & 0 \leqslant x \leqslant 1 \\ 0 & x > 1 \end{cases}$$

其对应的径向对称核称为 $Epanechnikov$ 核，定义为：

$$K(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-||x||^2), & ||x|| \leqslant 1 \\ 0 & ||x|| > 1 \end{cases}$$

其中 c_d 是 d 维单位球的体积。$Epanechnikov$ 剖面函数在 $x = 1$ 处不可导。

高斯核的剖面函数定义为：

$$k(x) = \exp(-\frac{1}{2}x)$$

其对应的多变量高斯核($Multivariate\ Gaussian\ Kernel$)为：

$$K(x) = (2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}||x||^2)$$

它的归一化系数为 $(2\pi)^{-\frac{d}{2}}$，因为 $\int_{\mathbb{R}^d}\exp(-\frac{1}{2}||x||^2) = (2\pi)^{\frac{d}{2}}$。

这就是标准的多维正态分布。

借助剖面函数，$p(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$ 可以写成：

$$f_{h,K}(x)=\frac{c_{k,d}}{nh^d}\sum_{i=1}^{n}k\Big(\Big|\Big|\frac{x-x_i}{h}\Big|\Big|^2\Big)$$

对于计算机视觉的目标跟踪，聚类，图像分割等，需要计算核密度函数的极大值点。如果使用梯度上升法求解该问题，就能得到著名的 $Mean\ Shift$ 均值漂移算法(第八章Lesson 8 无监督学习(聚类, 信号分解, 流形降维))。

9. K 近邻法(KNN)

KNN 其实是一种典型的监督学习算法。但是它并没有学习的过程，而是直接让某一个测试集基于到不同类别的训练集的距离的相似性度量来分类。

一个样本与数据集中的 k 个样本最相似， 如果这 k 个样本中的大多数属于某一个类别， 则该样本也属于这个类别。具体来说即每个样本都可以用它最接近的k个邻居来代表。这就是 KNN 的核心思想

KNN 分类算法的分类预测过程十分简单并容易理解：对于一个需要预测的输入向量 x_i ，我们只需要在训练数据集中寻找 k 个与向量 x 距离最近的样本(带标签)的集合，然后把 x_i 的类别预测为这 k 个样本中类别数最多的那一类。结构化描述如下：

- 计算待分类点与已知类别的点之间的距离。
- 按照距离递增次序排序。
- 选取与待分类点距离最小的 k 个点。
- 确定前 k 个点所在类别的出现次数。
- 返回前 k 个点出现次数最高的类别作为待分类点的预测分类。

有关距离的度量，有多种方式，L_1L_2 范数也可以作为距离的度量标准。

而 KNN 密度估计法，即根据 x_i 的近邻数，将样本分入不同的窗中，进行密度估计。根据总样本数，确定参数 k_n ，即确定即在中样本数为 n 时我们要求每个区域内拥有的样本的个数。在求 x 处的密度估计 $\hat{p}(x)$ 时，我们调整包含 x 的区域的体积，直到区域内恰好落入 k_n 个样本，并用下式来估计 $\hat{p}(x)$ ：

$$\hat{p}(x)=\frac{k_n/n}{V_n}$$

这样，在样本密度比较高的区域的体积就会比较小，而在密度低的区域的体积则会自动增大，这样就能够较好的兼顾在高密度区域估计的分辨率和在低密度区域估计的连续性。

为了取得好的估计效果，需要选择合适的 k_n 和 n 的关系，比如可以选择 $k_n=a\sqrt{n}$ ，a 为可变参数。

半监督学习中的标签传播算法(LP)使用的核函数(距离度量的方式)就有 KNN 和高斯窗(RBF 径向基)。

LP 算法更改标签的原则是选择与其相连的节点中所属标签距离最近的社区标签为自己的社区标签。即对没有标签的样本，通过核函数的密度估计来打上标签。不同的核函数代表距离度量的方式不同。详细的过程可以参考个人专栏的 $PyTorch$ 的图卷积神经网络教程。

11. 高斯分布($Normal\ Distribution$)与高斯混合模型(GMM)

一维正态分布

也称高斯分布($Gaussian\ Distribution$)，其概密函数为：

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

其中 μ 和 σ 分别为均值和方差，令 $t=\frac{(x-\mu)^2}{2\sigma^2}$ ，可得该函数在 $(-\infty,\infty)$ 上积分为 1 。

显然，其关于数学期望 μ 对称，且在该点存在极大值。在远离该点数学期望时，值单调递减。
$\lim\limits_{x\to+\infty}=0$，$\lim\limits_{x\to-\infty}=0$ 。现实中的很多数据，比如人的体重身高寿命等，近似服从正态分布。

如果随机变量 X 服从均值为 μ，方差为 σ^2 的正态分布，记为 $X\sim N(\mu,\sigma)$ 。当 $\mu=0,\sigma=1$，此时为标准正态分布，概密函数为

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

这是一个偶函数，其形状像钟，因此也称为钟形分布。

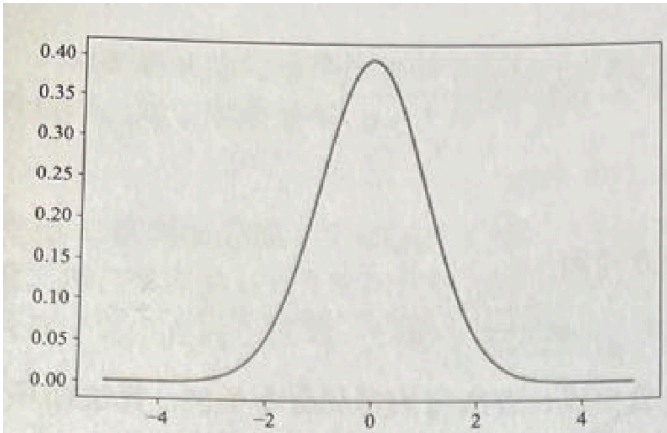


图 5.14　标准正态分布的概率密度函数

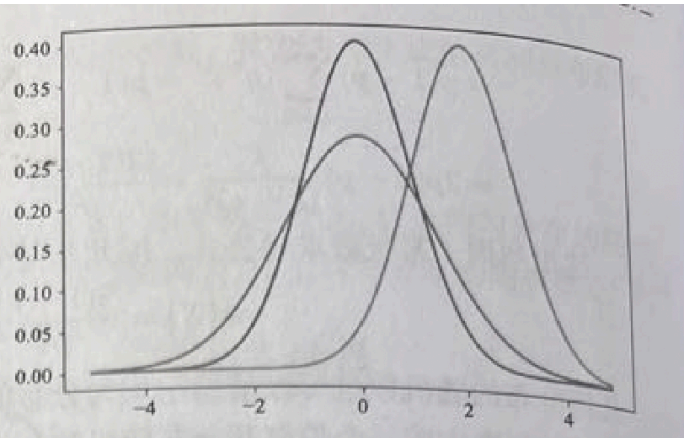


图 5.15　各种均值与方差的正态分布的概率密度函数

均值决定了峰值出现的位置，而方差决定了曲线的宽和窄，方差越大，曲线越宽。

正态分布 $N(\mu,\sigma^2)$ 的分布函数为

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(u-\mu)^2}{2\sigma^2}}\mathrm{d}u$$

e^{-x^2} 的不定积分不是初等函数，因此该函数无解析表达式。

根据分布变换，假设 Z 服从标准正态分布 $N(0,1)$，则 $X=\sigma Z+\mu$ 服从 $N(\mu,\sigma^2)$ 。相反，假设 X 服从 $N(\mu,\sigma^2)$，则 $Z=\frac{X-\mu}{\sigma}$ 服从标准正态分布。

正态分布的 $k-\sigma$ 置信区间定义为 $[\mu-k\sigma,\mu+k\sigma]$，随机变量落入该区间概率为 $p(\mu-k\sigma<X<\mu+k\sigma)=F(\mu+k\sigma)-F(\mu-k\sigma)$。在 $\sigma,2\sigma,3\sigma$ 的概率分别为 $0.6827,0.9545,0.9973$。

使用换元法和分布积分法，令 $z=\frac{x-\mu}{\sigma}$，根据数学期望定义：$E[X]=\int_{-\infty}^{\infty}xf(x)\mathrm{d}x$ 可得期望为 μ，根据方差定义 $var[X]=\int_{-\infty}^{\infty}(x-\mu)^2f(x)\mathrm{d}x$ 可得方差为 σ^2，标准差为 σ。

中心极限定理指出，正态分布是某些概率分布的极限分布。其具有 $(-\infty,\infty)$ 的支撑区间，且在所有定义于此区间的连续型概率分布中，正态分布的熵是最大的。多个正态分布的加权组合可构成高斯混合模型，它可以逼近任意连续型的概率分布。正态分布可以生成随机数。其熵和 KL 散度的计算在Lesson 7 信息论与决策树中会详细阐述。

多维正态分布(*Multivariate Normal Distribution*)

它将一维正态分布推广到高维，可得多维正态分布的概密函数

$$p(x)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)\Big)$$

其中 x 为 n 维随机向量，μ 为 n 维均值向量，Σ 为 n 阶协方差矩阵，通常要求其正定。其表达计作 $N(\mu,\Sigma)$。如果 $n=1,\mu=\mu,\Sigma=\sigma^2$，则该分布退化为一维正态分布。同样可证明其在 $\mathbb{R}^n$ 内的积分值为 1。

如果 $\mu=0,\Sigma=I$，则为标准正态分布，记为 $N(0,I)$，联合概密函数为

$$p(x)=\frac{1}{(2\pi)^{\frac{n}{2}}}\exp\Big(-\frac{1}{2}x^Tx\Big)$$

此时随机向量的各分量相互独立，且各自服从一维标准正态分布 $N(0,1)$。

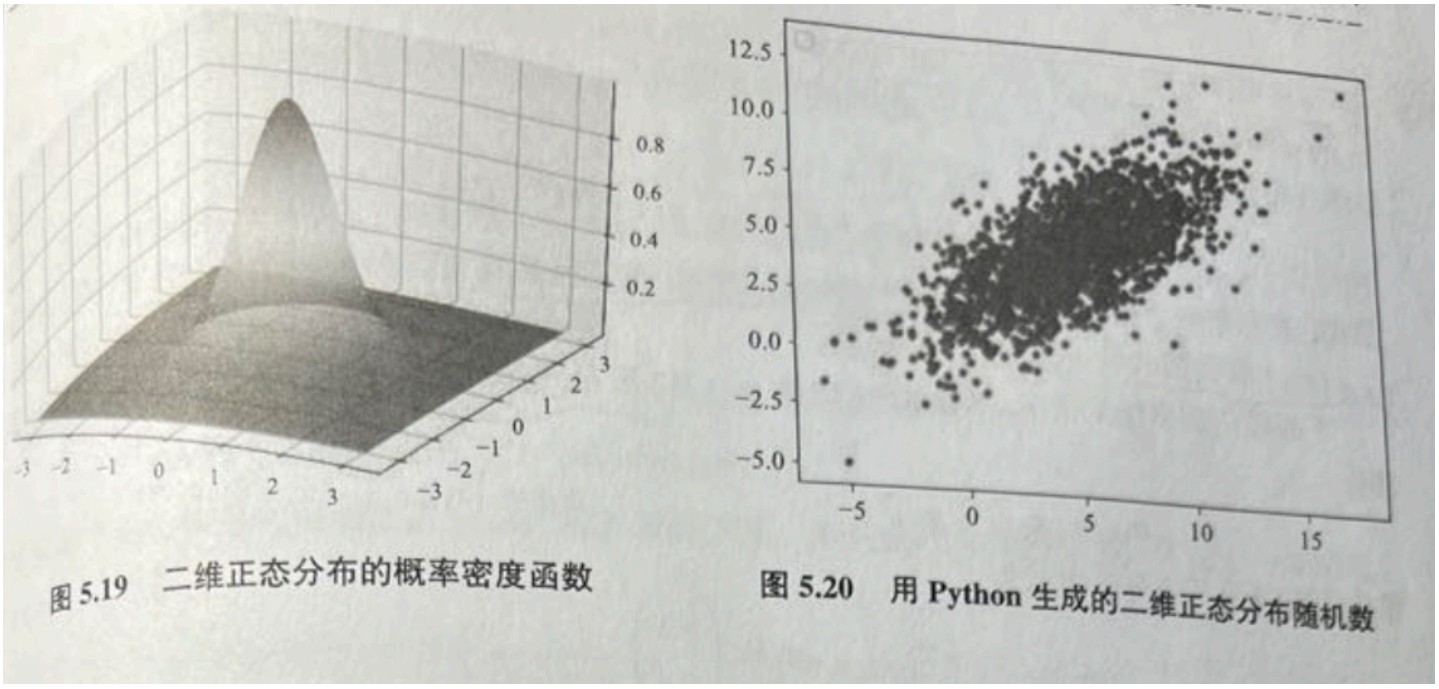


图 5.19　二维正态分布的概率密度函数

图 5.20　用 Python 生成的二维正态分布随机数

如图，其为钟形曲面，均值点有极大值，远离时递减。python 的 random.randn 可生成正态分布的随机数。

考虑二维正态分布，其均值向量和协方差矩阵为：

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

其概密函数可写为

$$p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\Big(-\frac{1}{2(1-\rho^2)}\Big(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\Big)\Big)$$

$0 \leqslant \rho \leqslant 1$ 称为相关系数，如果值为 0，则 X, Y 相互独立。下面计算边缘密度函数，如果令 $u = \frac{x-\mu_1}{\sigma_1}, v = \frac{y-\mu_2}{\sigma_2}$，则有

$$p_X(x) = \int_{-\infty}^{+\infty} p(x,y)\mathrm{d}y = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

因此 X 服从正态分布 $N(\mu_1, \sigma_1^2)$，类似得到 Y 服从正态分布 μ_2, σ_2^2。如果二者相互独立，$\rho = 0$，则有

$$p(x,y) = p_X(x)p_Y(y)$$

下面计算条件概密函数，根据定义

$$p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)} = \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)\sigma_2^2}\Big(y - \big(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\big)\Big)^2\right)$$

服从正态分布 $N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2))$，$p_{Y|X}(y|x)$ 仍然是正态分布，且均值与另一个变量 x 有关，$p_{X|Y}(x|y)$ 同理。

推广到多维，假设 $x \in \mathbb{R}^n$ 服从正态分布 $N(\mu, \Sigma)$。将该向量拆成两部分 $x_A = (x_1 \cdots x_\tau)^T, x_B = (x_{\tau+1} \cdots x_n)^T$，整个向量可以分块表示为 $x = \begin{pmatrix} x_A \\ x_B \end{pmatrix}$。

相应的均值向量和协方差向量可拆分为

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$$

高斯混合模型

12. EM 算法($Expectation\ Maximization$，期望最大化算法)

首先介绍 $Jensen$ 不等式，它将应用于后续算法推导。

$Jensen$ 不等式

回顾第 Lesson 3 优化方法基础 中的凸优化定义，如果 $f(x)$ 是一个凸函数，$0 \leqslant \theta \leqslant 1$，则有：

$$f(\theta x_1 + (1-\theta)x_2 \leqslant \theta f(x_1) + (1-\theta)f(x_2))$$

将上式从两个点推广到 m 个点，如果

$$a_i \geqslant 0, i = 1, 2, \cdots, m \qquad\qquad a_1 + \cdots + a_m = 1$$

可以得到，对于 $\forall x_1, \cdots, x_m$ 有

$$f(a_1 x_1 + \cdots + a_m x_m) \leqslant a_1 f(x_1) + \cdots + a_m f(x_m)$$

如果将 x 看作一个随机变量，$p(x = x_i) = a_i$ 是其概率分布，则有

$$E[x] = a_1 x_1 + \cdots + a_m x_m \qquad\qquad E[f(x)] = a_1 f(x_1) + \cdots + a_m f(x_m)$$

从而得到 $Jensen$ 不等式：

$$E[f(x)] \geqslant f(E[x])$$

对于凹函数，不等式反号。

如果 $f(x)$ 是严格凸函数且 x 不是常数，则有：

$$E[f(x)] > f(E[x])$$

如果 $f(x)$ 是严格凸函数，当且仅当随机变量 x 是常数时，不等式取等号：

$$E[f(x)] = f(E[x])$$

$Jensen$ 不等式可以推广到随机向量的情况，我们将利用该不等式取推导求解含有隐变量的最大似然估计问题的 EM 算法。

算法原理

算法过程

13. 一些仅供参考奇奇怪怪的个人想法

想了两天，结合一些学习过的算法，论文，做深度学习的模型构造经验，以及朋友毕设的问题，想了一些仅供参考的奇奇怪怪的想法？不一定对，仅供参考。

很多学习问题及其应用，其实都可以看成相似度的度量问题。举例来说，用最大似然估计(当给出了标签和训练集 (y, X) 求使得 $p(y|X, \theta)$ 最大的参数 θ)求联合概率密度推导出的线性回归，用最大似然估计和交叉熵联合推导出的 $Logistic$ 回归和 $Softmax$ 回归，最大后验估计(引入后验)推导出的拉索回归(L_1)、岭回归(L_2)和贝叶斯决策，基于距离度量的 KNN ，支持向量机的最胖类间 $margin$，条件熵的计算以及决策树的生成，基于 KL 散度的流形降维，基于类间相似性度量的聚类算法，使用卷积算子度量数据之间互相关程度(为什么卷积是离散的乘法 $*$ ，因为本来就可以看成两个概率矩阵，数据 p_{ij} 和 q_{ij} 的最大似然问题)，残差网络的恒等映射，稠密网络的特征融合，自注意力机制里的输出特征图的矩阵哈达玛积，基于 JS 散度的无监督对抗网络。这些算法，其实都是度量的不同类样本间，或者说模型的输入和输出之间距离和相似度，或者说，模式之间的匹配程度，。

如果这一段你觉得有点陌生，或者难以理解，可以把全部的内容看完再回头看这一段，相信你也许会和我有些共鸣。虽然我觉得这样总结稍微有些牵强，但是我觉得就像是所有的学习问题本质上都是最优化问题，其实所有的学习算法的目标从概率的角度解释，都是共通的。这也是我觉得机器学习中理论的魅力所在吧。