第一章、绪论

赵涵

2023 年 3 月 7 日

1　机器学习历史回顾

机器学习实际上已经存在了几十年或者也可以认为存在了几个世纪。追溯到17世纪，贝叶斯、拉普拉斯关于最小二乘法的推导和马尔可夫链，这些构成了机器学习广泛使用的工具和基础。1950年（艾伦.图灵提议建立一个学习机器）到2000年初（有深度学习的实际应用以及最近的进展，比如2012年的AlexNet），机器学习有了很大的进展。从20世纪50年代研究机器学习以来，不同时期的研究途径和目标并不相同，可以划分为四个阶段。

第一阶段是20世纪50年代中叶到60年代中叶，这个时期主要研究“有无知识的学习”。这类方法主要是研究系统的执行能力。这个时期，主要通过对机器的环境及其相应性能参数的改变来检测系统所反馈的数据，就好比给系统一个程序，通过改变它们的自由空间作用，系统将会受到程序的影响而改变自身的组织，最后这个系统将会选择一个最优的环境生存。在这个时期最具有代表性的研究就是Samuet的下棋程序。但这种机器学习的方法还远远不能满足人类的需要。

第二阶段从20世纪60年代中叶到70年代，这个时期主要研究将各个领域的知识植入到系统里，在本阶段的目的是通过机器模拟人类学习的过程。同时还采用了图结构及其逻辑结构方面的知识进行系统描述，在这一研究阶段，主要是用各种符号来表示机器语言，研究人员在进行实验时意识到学习是一个长期的过程，从这种系统环境中无法学到更加深入的知识，因此研究人员将各专家学者的知识加入到系统里，经过实践证明这种方法取得了一定的成效。在这一阶段具有代表性的工作有Hayes-Roth和Winson的对结构学习系统方法。

第三阶段从20世纪70年代中叶到80年代，称为复兴时期。在此期间，人们从学习单个概念扩展到学习多个概念，探索不同的学习策略和学习方法，且在本阶段已开始把学习系统与各种应用结合起来，并取得很大的成功。同时，专家系统在知识获取方面的需求也极大地刺激了机器学习的研究和发展。在出现第一个专家学习系统之后，示例归纳学习系统成为研究的主流，自动知识获取成为机器学习应用的研究目标。1980年，在美国的卡内基梅隆（CMU）召开了第一届机器学习国际研讨会，标志着机器学习研究已在全世界兴起。此后，机器学习开始得到了大量的应用。1984年，Simon等20 多位人工智能专家共同撰文编写的MachineLearning文集第二卷出版，国际性杂志Machine Learning创刊，更加显示出机器学习突飞猛进的发展趋势。这一阶段代表性的工作有Mostow的指导式学习、Lenat的数学概念发现程序、Langley 的BACON程序及其改进程序。

第四阶段20世纪80年代中叶至今，是机器学习的最新阶段。这个时期的机器学习具有如下特点：（1）机器学习已成为新的学科，它综合应用了心理学、生物学、神经生理学、数学、自动化和计算机科学等形成了机器学习理论基础。（2）融合了各种学习方法，且形式多样的集成学习系统研究正在兴起。（3）机器学习与人工智能各种基础问题的统一性观点正在形成。（4）各种学习方法的应用范围不断扩大，部分应用研究成果已转化为产品。

机器学习作为一门课程，其实很早已经在大学开设，通过我的调查来看，早在2007年就在斯坦福大学开设课程，当然可能还有更早的大学开设。**人工智能**（Artificial Intelligence）正在改变世界，斯

坦福大学处于这一趋势的最前沿。多年来，斯坦福已经涌现出许多AI 方面的重大研究突破，斯坦福研究者也是AI 领域的开拓者。所以这门课程，主要参考斯坦福大学的课程CS229[1]，还有Bilibili网站上，白板机器学习推导[2]，并且在它的基础上，我们补充一些有关深度学习等相关的内容。

2 机器学习的分类

机器学习如果按照学习的任务来看，分为回归，聚类，分类。按照学习的方式来看，分为监督学习，无监督学习，强化学习。按照模型的思想，又可以分为频率派机器学习和贝叶斯派机器学习。按照模型的工作方式，又可以分为判别式模型和生成式模型。在之后的章节我们都会一一介绍。

2.1 按学习任务分类

机器学习主要面临三个任务，**回归**（Regression），**分类**（Classification），**聚类**（Cluster）。回归的意思是让模型学习一个具体的实数。比如房价预测问题，对目前市场上已有的二手房和新房，建立一个模型，去拟合房价，当出现新房源时，模型能准确的预测房屋价格，因为房屋价格是一个连续变化的实数，所以这就是回归任务。还有一类数据是需要进行分类的，比如人脸识别，人脸识别是目前深度学习领域很成熟的技术了，当你站在摄像头前，模型能准确地区分出来，是否是你本人。这种任务就属于分类任务，区分每一类地不同。最后一类学习任务是聚类，聚类所要解决的问题是，当数据没有区分度时，并且你还不知道如何准确区分时，你需要选择一个尽量能区分数据的标准，把数据进行聚类，即比较像的数据归为一类。以后再来新的数据，就可以按照之前建立好的模型，很快地进行分类。

2.2 按学习方式分类

我们首先解释一下，什么是学习方式？所谓的学习方式就是观察学习的数据是否具有标签，标签是一个数据的标注。举个例子，比如一张图片，上面是一只小狗，人类看到这张照片很快就会做出分类，把图片上的信息做出判断，但是对于计算机，它是不知道图片上是什么，所以应该给图片添加一个标签，这个标签就是dog，把图片和标签同时输入给计算机，计算机才能知道，这个图片所要表达的意思。带有标签的学习，就称为**监督学习**（Supervised Learning）；如果数据没有标签，就称为**无监督学习**（Unsupervised Learning）。这里可以理解为标签是一个帮助你学习的老师，有老师监督，就是监督学习，没有老师，你只能自己摸索，属于无监督学习。还有一类学习，比如下棋，驾驶，打游戏等，这些从新手到大师，需要不停地与环境进行相互作用，得到环境的反馈，进而改进你的行为与策略，这种学习过程，我们称为**强化学习**（Reinforcement Learning）。

我们按照学习的方式，把三大类下，各有哪些模型，简单说明一下，作为本课程的一个总览。

监督学习：线性回归（Linear Regression），逻辑思蒂回归（Logistic Regression），广义线性模型（General Linear Model），高斯判别模型（Gaussian Discrimination Model），朴素贝叶斯模型（Naive Bayes' Model），支撑向量机（Support Vector Machine），决策树（Decision Tree），神经网络（Neural Network）等。

无监督学习：K-means聚类（K-means Cluster），基于密度的聚类（Density-Based Spatial Clustering of Applications with Noise），混合高斯模型（Mixture Gaussian Model），概率图模型（Probably Graph Model），深度概率图模型（Deep Probability Graph Model），变分推断（Variational Inference）等。

强化学习：策略评估（Policy Evaluation）策略改进（Policy Improvement），深度强化学习（Deep Reinforcement Learning）。

我们以上只是按照学习的方式，简单的说明了一下，在每一类下，大致都有哪些模型，由于课程的时间关系，我们不会把所有的模型都进行介绍，原则是介绍目前最基础，最常用，最流行的一些模型与算法。比如像决策树模型，我们就不再介绍了，相对于其他模型，目前已经不常用这个模型了（在CS229课程也不涉及该模型），如果感兴趣的话，可以去查阅相关的资料与书籍。

2.3 按建模思想分类

对概率的诠释有两大学派，一种是频率派，另一种是贝叶斯派。我们对观测到的数据集采用如下的记号：

$$X_{N\times p} = (x^{(1)}, x^{(2)}, \cdots, x^{(N)})^T,\ x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_p^{(i)})^T \tag{1}$$

这里，X是样本构成的一个矩阵，每一个样本是$x^{(i)}$，它具有p个特征，所以是一个p维的列向量。N代表样本的总数。每个样本都是由$p(x|\theta)$ 生成的，且都是独立同分布的（independently identically distribution，iid）。

频率派观点：$p(x|\theta)$的参数θ是一个常量。对于N个样本来说，样本集的联合概率分布为：$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$。为了求$\theta$的大小，我们采用**最大对数似然估计**（Maximum Log-likelihood Estimation，MLE）的方法：

$$\theta_{MLE} = \arg\max_{\theta} \log p(X|\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \log p(x^{(i)}|\theta) \tag{2}$$

贝叶斯派的观点：$p(x|\theta)$中的θ不是一个常量。这个θ应该满足一个预设的先验分布，即$\theta \sim p(\theta)$。于是根据贝叶斯定理以来样本集参数的后验概率可以写成：

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int_{\theta} p(X|\theta)p(\theta)d\theta} \tag{3}$$

为了求θ的值，我们要最大化这个参数的后验概率$MaximumAPosterior$：

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|X) = \arg\max_{\theta} p(X|\theta)p(\theta) \tag{4}$$

第二个等号是由分母和θ没有关系。求解这个θ值后计算$\int_{\theta} p(X|\theta)p(\theta)d\theta$，就得到了参数的后验概率。其中$p(X|\theta)$叫似然函数，是我们的模型分布。得到了参数的后验分布后，我们可以将这个分布用于预测贝叶斯预测：

$$p(x^{new}|X) = \int_{\theta} p(x^{new}|\theta)p(\theta|X)d\theta \tag{5}$$

频率派和贝叶斯派分别给出了一系列的机器学习模型。频率派的观点导出了统计机器需学习算法，而贝叶斯派导出了概率图模型。在应用频率派的MLE方法时，最优化理论占有重要地位。而贝叶斯派的模型无论是后验概率的建模还是后验概率进行推断时，积分占有重要地位。因此采样积分方法，蒙特卡洛马尔科夫链（Monte Carlo Markov Chain）采样是不可或缺的。

2.4 按模型输出分类

按照模型最后输出的结果来看，又可以把模型分成判别式和生成式。所谓的判别式是在已有的数据上，寻找$p(Y|X)$，这里X是数据，Y是数据所对应的标签。我们建立一个模型，能最大化这个条件概率就行了。而生成式模型，是在对$p(X|Y)$建立模型，即我们知道了数据的标签，然后可以通过已有的标签去生成一个新的数据，这个新数据是之前的数据集没有的。这种模型是目前比较受欢迎的，因为它可以解决数据量过小，人工产生数据成本高等问题。目前比较常用的模型为：变分自动编码器（Variational Auto-Encoder）和生成对抗网络（Generalize Adversarial Network），后面的章节，我们会对这两种模型进行简单的介绍。

3 高斯分布

在介绍机器学习的模型之前，我们先从高斯分布入手，讨论一下它的性质，因为高斯分布在机器学习中占有举足轻重的作用。当数据量很大时，随机变量都会趋近于高斯分布，这是中心极限定理的结果。我们认为存在一个高斯分布，在总体中，我们通过采样得到的样本集，去估计真实的高斯分布的参数，首先我们有：

$$\theta = (\mu, \Sigma) = (\mu, \sigma^2) \tag{6}$$

用MLE方法去进行估计，那么：

$$\theta_{MLE} = \arg\max_{\theta} \log p(X|\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \log p(x^{(i)}|\theta) \tag{7}$$

3.1 一维高斯分布

简单起见，我们首先讨论一维高斯分布，即每个样本是一个标量。一般地，高斯分布的PDF写成如下的形式：

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x^{(i)} - \mu)^2/2\sigma^2) \tag{8}$$

那么把公式代入到方程7，我们得到：

$$\mu_{MLE} = \arg\max_{\mu} \log p(X|\theta) = \arg\max_{\mu} \sum_{i=1}^{N} (x^{(i)} - \mu)^2 \tag{9}$$

于是：

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} (x^{(i)} - \mu)^2 = 0 \rightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \tag{10}$$

接下来对θ中的另一个参数σ做估计，有：

$$\sigma_{MLE} = \arg\max_{\sigma} \log p(X|\theta) = \arg\max_{\sigma} \sum_{i=1}^{N} [-\log \sigma - \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2] = \arg\min_{\sigma} \sum_{i=1}^{N} [\log \sigma + \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2] \tag{11}$$

于是：

$$\frac{\partial}{\partial \sigma} \sum_{i=1}^{N} [\log \sigma + \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2] = 0 \tag{12}$$

可以得到：

$$\sigma^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \mu)^2 \tag{13}$$

值得注意的是，上面的推到中，首先对μ做MLE，然后利用这个结果求σ_{MLE}，因此可以预期的是对数据集求期望时，$E_{[D]}[\mu_{MLE}]$是无偏估计（可以把$\mathcal{D}$认为是总体）：

$$E_{\mathcal{D}}[\mu_{MLE}] = E_{\mathcal{D}}[\frac{1}{N} \sum_{i=1}^{N} x^{(i)}] = \frac{1}{N} \sum_{i=1}^{N} E_{\mathcal{D}}[x^{(i)}] = \mu \tag{14}$$

但是当对σ_{MLE}求期望的时候，由于使用了单个数据集的μ_{MLE}，因此对所有数据集求期望的时候我们

会发现σ_{MLE}是有偏估计：

$$
\begin{aligned}
E_{\mathcal{D}}[\sigma^2_{MLE}] &= E_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}(x^{(i)}-\mu_{MLE})^2] = E_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}(x^{(i)^2}-2x^{(i)}\mu_{MLE}+\mu^2_{MLE})] && (15)\\
&= E_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x^{(i)^2}-\mu^2_{MLE}] = E_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x^{(i)^2}-\mu^2+\mu^2-\mu^2_{MLE}] && (16)\\
&= E_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x^{(i)^2}]-E_{\mathcal{D}}[\mu^2_{MLE}-\mu^2] = \sigma^2-(E_{\mathcal{D}}[\mu^2_{MLE}]-\mu^2) && (17)\\
&= \sigma^2-(E_{\mathcal{D}}[\mu^2_{MLE}]-E^2_{\mathcal{D}}[\mu_{MLE}]) = \sigma^2-Var[\mu_{MLE}] && (18)\\
&= \sigma^2-Var[\frac{1}{N}\sum_{i=1}^{N}x^{(i)}] = \sigma^2-\frac{1}{N^2}\sum_{i=1}^{N}Var[x^{(i)}] && (19)\\
&= \frac{N-1}{N}\sigma^2 && (20)
\end{aligned}
$$

所以：

$$\hat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)}-\mu)^2 \tag{21}$$

3.2 多维高斯分布

对于多维的高斯分布，具有如下的表达式：

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{22}$$

其中$x,\mu\in\mathbb{R}^p,\Sigma\in\mathbb{I}\times\mathbb{I}$，$\Sigma$为协方差矩阵，这里我们认为协方差矩阵为正定的对称阵。首先我们来看指数上的数字，指数上的数字可以记为x和μ之间的马氏距离。对于协方差矩阵可以进行特征值分解，$\Sigma = U\Lambda U^T = (u_1,u_2,...,u_p)diag(\lambda_i)(u_1,u_2,...,u_p)^T = \sum_{i=1}^{p}u_i\lambda_i u_i^T$，于是：

$$\Sigma^{-1} = \sum_{i=1}^{p}u_i\frac{1}{\lambda_i}u_i^T \tag{23}$$

那么指数上的表达式，可以写为：

$$
\begin{aligned}
(x-\mu)^T\Sigma^{-1}(x-\mu) &= \sum_{i=1}^{p}(x-\mu)^Tu_i\frac{1}{\lambda_i}u_i^T(x-\mu) && (24)\\
&= \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i} && (25)
\end{aligned}
$$

这里$y_i = (x-\mu)^Tu_i$，即在特征向量上的投影长度，因此上式就是指数上，取不同值的同心椭圆。在多维高斯分布下，一般会遇到两个问题：1，参数Σ,μ的自由度为$o(p^2)$对于维度很高的数据，不容易处理，我们一般取近似，有两种方法，一种是因子分析，一种是概率主成分分析。2，单独一个高斯分布，是单峰的，对于多峰的数据分布得不到好的拟合效果，解决办法是混合高斯模型。我们对多维高斯分布，在深入分析一下它的性质。

首先引入一些记号：$x = (x_1,x_2,\ldots,x_p)^T = (x_a,x_b)^T$，这里我们把一个随机向量分为了两部分，一部分是$x_a$,一部分是$x_b$，且$a+b=p$，所以我们有$\mu = (\mu_m,\mu_n)$，$\Sigma = \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}$，已知$x\sim N(\mu,\Sigma)$。

在这些记号下，我们给出一个定理：

定理：已知$x \sim N(\mu, \Sigma), y = Ax + b$，那么$y \sim N(A\mu + b, A\Sigma A^T)$。

证明：$E[y] = E[Ax+b] = AE[x]+b = A\mu+b$，$Var[y] = Var[Ax+b] = Var[Ax] = AVar[x]A^T$。

下面我们利用这个定理得到$p(x_a), p(x_b), p(x_b|x_a)$ 这四个概率值。

1，$x_a = (I_{m\times m}\ 0_{m\times n}) \begin{pmatrix} x_a \\ x_b \end{pmatrix}$，带入定理中得到：

$$E[x_a] = (I\ 0) \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_a \tag{26}$$

$$Var[x_a] = (I\ 0) \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} I \\ 0 \end{pmatrix} = \Sigma_{aa} \tag{27}$$

所以$x_a \sim N(\mu_a, \Sigma_{aa})$。

2，同1，得到$x_b \sim N(\mu_b, \Sigma_{bb})$。

3，对于两个条件概率，我们进行构造性证明，即构造如下的三个量：

$$x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1}x_a \tag{28}$$

$$\mu_{b\cdot a} = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a \tag{29}$$

$$\Sigma_{bb\cdot a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \tag{30}$$

特别地，最后一个公式，叫做Σ_{bb}地Schur Complementary。可以看到：

$$x_{b\cdot a} = (-\Sigma_{ba}\Sigma_{aa}^{-1}\ I_{n\times n}) \begin{pmatrix} x_a \\ x_b \end{pmatrix} \tag{31}$$

所以：

$$E[x_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}\ I_{n\times n}) \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_{b\cdot a} \tag{32}$$

$$Var[x_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}\ I_{n\times n}) \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} \\ I_{n\times n} \end{pmatrix} = \Sigma_{bb\cdot a} \tag{33}$$

利用这三个量可以得到$x_b = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a$。这里有定理：

若$x \sim N(\mu, \Sigma)$，则$Mx \perp Nx \Leftrightarrow M\Sigma N^T = 0$。

证明：$\because x \sim N(\mu, \Sigma), \therefore Mx \sim N(M\mu, M\Sigma M^T), Nx \sim N(N\mu, N\Sigma N^T), \therefore Cov(Mx, Nx) = E[(Mx - M\mu)(Nx - N\mu)^T] = E[M(x-\mu)(x-\mu)^T N] = ME[(x-\mu)(x-\mu)^T]N = M\Sigma N$，$\because Mx \perp Nx$且均为高斯，$\therefore Cov(Mx, Nx) = M\Sigma N^T = 0$。那么我们有$M = (-\Sigma_{ba}\Sigma_{aa}^{-1}\ I)$，$N = (I\ 0)$，计算$M\Sigma N^T$，可以得到0。那么我们有如下的结论：

$$x_{b\cdot a} \perp x_a \Rightarrow x_{b\cdot a}|x_a = x_{b\cdot a} \tag{34}$$

进而有：

$$x_b|x_a = x_{b\cdot a}|x_a + \Sigma_{ba}\Sigma_{aa}^{-1}x_a|x_a = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a \tag{35}$$

因此：

$$E[x_b|x_a] = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a \tag{36}$$

$$Var[x_b|x_a] = \Sigma_{bb\cdot a} \tag{37}$$

这里同样用到了定理。

4，可以继续构造一个对称的量：

$$x_{a\cdot b} = x_a - \Sigma_{ab}\Sigma_{bb}^{-1}x_b \tag{38}$$

$$\mu_{a\cdot b} = \mu_a - \Sigma_{ab}\Sigma_{bb}^{-1}\mu_b \tag{39}$$

$$\Sigma_{aa\cdot b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \tag{40}$$

所以：

$$E[x_a|x_b] = \mu_{a\cdot b} + \Sigma_{ab}\Sigma_{bb}^{-1}x_b \tag{41}$$

$$Var[x_a|x_b] = \Sigma_{aa\cdot b} \tag{42}$$

下面我们看一道例题：

已知：$p(x) = N(\mu, \Lambda^{-1}), p(y|x) = N(Ax+b, L^{-1})$，求$p(x), p(x|y)$。

解：令$y = Ax+b+\epsilon$，$\epsilon \sim N(0, L^{-1})$，所以$E[y] = E[Ax+b+\epsilon] = A\mu+b$，$Var[y] = A\Lambda^{-1}A^T+L^{-1}$，因此：

$$p(y) = N(A\mu+b, L^{-1}+A\Lambda^{-1}A^T) \tag{43}$$

引入$z = \begin{pmatrix} x \\ y \end{pmatrix}$，我们计算这个随机向量的协方差$Cov[x,y] = E[(x-E[x])(y-E[y])^T]$。我们直接通过定义进行计算：

$$Cov[x,y] = E[(x-\mu)(Ax-A\mu+\epsilon)^T] = E[(x-\mu)(x-\mu)^TA^T] = Var[x]A^T = \Lambda^{-1}A^T \tag{44}$$

因为协方差具有对称性，所以我们有：

$$p(z) = N(\begin{pmatrix} \mu \\ A\mu+b \end{pmatrix}, \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1}+A\Lambda^{-1}A^T \end{pmatrix}) \tag{45}$$

同样利用之前我们计算得到的结果，直接得到：

$$E[x|y] = \mu + \Lambda^{-1}A^T(L^{-1}+A\Lambda^{-1}A^T)^{-1}(A\mu+b) \tag{46}$$

$$Var[x|y] = \Lambda^{-1} - \Lambda^{-1}A^T(L^{-1}+A\Lambda^{-1}A^T)^{-1}A\Lambda^{-1} \tag{47}$$

这一部分内容，我们会在介绍因子分析的时候，再次讨论。

References

[1] Andrew Ng. cs229. `http://cs229.stanford.edu/syllabus.html`.

[2] shuhuai008. Machine learning. `https://space.bilibili.com/97068901`.