MLAPP（Machine Learning A Probabilistic Perspective）读书会第四次活动

第一章 Introduction 第二章　Probability

主讲人 网络上的尼采

（新浪微博:@Nietzsche_复杂网络机器学习）

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

网络上的尼采(813394698) 16:32:00
MLAPP 第一章 Introduction。
先是一句文艺的台词：

Machine learning: what and why?

We are drowning in information and starving for knowledge. — John Naisbitt.

即使我们拥有海量数据，也不一定能获得感兴趣的东西，在长尾分布中大部分数据对象都是非平凡的：

It should be noted, however, that even when one has an apparently massive data set, the effective number of data points for certain cases of interest might be quite small. In fact, data across a variety of domains exhibits a property known as the **long tail**, which means that a few things (e.g., words) are very common, but most things are quite rare (see Section 2.4.6 for details). For example, 20% of Google searches each day have never been seen before[4]. This means that the core statistical issues that we discuss in this book, concerning generalizing from relatively small samples sizes, are still very relevant even in the big data era.

有监督学习，可以看成输入到输出的一种映射，输出离散的就是分类，连续的就是回归。有监督学习可以理解为学习一个最佳的逼近函数。
无监督学习和有监督的区别是没有标注好的训练数据，也可以叫知识发现，从概率角度看是一个密度估计问题，有监督的学习则是一个条件概率密度估计问题。书上提到无监督的学习更接近人和动物的学习，对此深表赞同。

引用 Hinton 的一段话

When we're learning to see, nobody's telling us what the right answers are — we just look. Every so often, your mother says "that's a dog", but that's very little information. You'd be lucky if you got a few bits of information — even one bit per second — that way. The brain's visual system has 10^{14} neural connections. And you only live for 10^9 seconds. So it's no use learning one bit per second. You need more like 10^5 bits per second. And there's only one place you can get that much information: from the input itself. — Geoffrey Hinton, 1996 (quoted in (Gorder 2006)).

无监督能做的事情：发现簇结构也就是聚类，主成分分析，发现稀疏图的结构，协同过滤等。这本书把关联规则的挖掘也归结过来了，关于这方面大家可以看 HAN 的那本《数据挖掘概念与技术》。
此外这本书无监督的任务少归结了一项：outliers detection
下面介绍一些基本概念：参数模型和非参模型的区别，主要体现在参数数目是否固定，是否随数据集的规模变化，举个例子：高斯过程的协方差矩阵就会随着数据的不断到来而增长。以后我们还会具体讲非参模型比如狄利克雷过程等。非参模型这章讲了 KNN 分类的例子。
维灾: 高维是个问题 ,以 k 近邻索引结构为例 x-tree 的变种也最多到 50 多维 ,再多时间复杂度变 o(n^2)；在超高维空间，欧氏距离失效，所有的数据对象都会变成 outliers，一些算法比如 KNN 自然也不能用了。
下面是参数模型，以线性回归为例：
这个大家都比较熟悉了，PRML MLAPP 读书会各讲了一次

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

这本书名字很好 ,Machine Learning A Probabilistic Perspective ,现在我们从概率的角度来看线性回归，

如果我们把上面的误差 ϵ 看成高斯分布，原来的线性回归就会变成下面的高斯分布形式：

$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$$

其中μ(x)是 x 的线性函数， $\mu(\mathbf{x}) = w_0 + w_1 x = \mathbf{w}^T \mathbf{x}$

上面线性回归的高斯分布形式的均值便是μ(x)，方差是随机误差分布的方差。

shock(21638731) 17:12:21

然后呢

用 maximum likelihood

是么🤔

网络上的尼采(813394698) 17:13:53

加上非线性的基函数可以有非线性的表达能力 $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2)$

参数估计不一定用最大似然，也有贝叶斯线性回归，看对 w 的态度了， $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2)$

如果对 w 加上高斯先验会引出高斯过程。

van19(493010846) 17:15:29

我有一个问题，就是如果我的线性模型欠拟合（或过拟合）的情况严重，那么这个误差就会不符合 u=0 的高斯分布，所以，为何可以认为误差符合 u=0 的高斯分布？

网络上的尼采(813394698) 17:17:07

u 不等于 0 的高斯分布可以转化成 u=0 的形式，另外随机误差符合高斯分布只是一种假设，线性高斯模型比较好处理。

van19(493010846) 17:18:07

嗯，好的

网络上的尼采(813394698) 17:19:18

再看逻辑回归，这是分类，加了 sigmoid 函数。

因为是二分类问题可以写成伯努利分布的形式：

$$p(y|\mathbf{x}, \mathbf{w}) = \text{Ber}(y|\mu(\mathbf{x}))$$

其中 $\mu(\mathbf{x}) = \text{sigm}(\mathbf{w}^T \mathbf{x})$

sigmoid 函数的形式：

$$\text{sigm}(\eta) \triangleq \frac{1}{1 + \exp(-\eta)} = \frac{e^\eta}{e^\eta + 1}$$

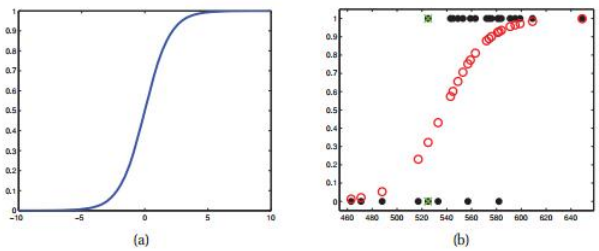


Figure 1.19 (a) The sigmoid or logistic function. We have $\text{sigm}(-\infty) = 0$, $\text{sigm}(0) = 0.5$, and $\text{sigm}(\infty) = 1$. Figure generated by `sigmoidPlot`. (b) Logistic regression for SAT scores. Solid black dots are the data. The open red circles are the predicted probabilities. The green crosses denote two students with the same SAT score of 525 (and hence same input representation x) but with different training labels (one student passed, $y = 1$, the other failed, $y = 0$). Hence this data is not perfectly separable using just the SAT feature. Figure generated by `logregSATdemo`.

这个函数很优美

自然界中的很多规律比如细菌的群体增长都符合这个函数。
关于参数的估计会在逻辑回归那章详细讲。

最后做预测时超过阈值便可以归为一类 $\hat{y}(x) = 1 \iff p(y = 1|\mathbf{x}) > 0.5$

过拟合：过拟合和很多因素有关，模型的复杂度、训练数据的多少、outliers 的干扰等都是造成过拟合的原因，现在随着计算能力的增强，数据的增长，一些复杂的模型比如 Deep Learning 的过拟合问题得以解决。

最后 No free lunch 定理：不存在通用的最好的模型，不考虑时间因素的话无穷猴子也能打出一部莎士比亚全集，有时必须在速度，精度，复杂度上做折中。

MLAPP 第二章 Probability

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

先来句

拉普拉斯的名言。

PRML 第二章的内容到了 MLAPP 分散到了几章的内容，这一章和 PRML 第二章有很多重复的内容。

arxov517(1193235126) 14:36:19

概率沉思录第一章恰好解释了拉普拉斯那句话。。。

网络上的尼采(813394698) 14:38:03

一开始介绍了频率派和贝叶斯学派的区别

贝叶斯逝世后沉寂了一百多年，然后贝叶斯的思想被人们重新发现，看《数理统计简史》从上世纪 30 年代开始和频率派展开论战，当时是被几个频率派的牛人打压，书里还写到一统计学家准备写一本书客观的介绍频率派和贝叶斯，最后把自己写成了贝叶斯狂热分子。

当时贝叶斯在应用上受限，因为求 marginalization 算积分很难直接得出来 ，后来有了蒙特卡洛算法才开始有了较大发展。

接下来都是基本的东西

针对离散变量的一些定理 ，大家可以看到最下面的贝叶斯公式

2.2.2.1 Probability of a union of two events

Given two events, A and B, we define the probability of A or B as follows:

$$p(A \vee B) = p(A) + p(B) - p(A \wedge B) \tag{2.1}$$

$$= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \tag{2.2}$$

2.2.2.2 Joint probabilities

We define the probability of the joint event A and B as follows:

$$p(A, B) = p(A \wedge B) = p(A|B)p(B) \tag{2.3}$$

This is sometimes called the **product rule**. Given a **joint distribution** on two events $p(A, B)$, we define the **marginal distribution** as follows:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b) \tag{2.4}$$

where we are summing over all possible states of B. We can define $p(B)$ similarly. This is sometimes called the **sum rule** or the **rule of total probability**.

The product rule can be applied multiple times to yield the **chain rule** of probability:

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3)\ldots p(X_D|X_{1:D-1}) \tag{2.5}$$

where we introduce the Matlab-like notation $1 : D$ to denote the set $\{1, 2, \ldots, D\}$.

2.2.2.3 Conditional probability

We define the **conditional probability** of event A, given that event B is true, as follows:

$$p(A|B) = \frac{p(A, B)}{p(B)} \text{ if } p(B) > 0 \tag{2.6}$$

2.2.3 Bayes rule

Combining the definition of conditional probability with the product and sum rules yields **Bayes rule**, also called **Bayes Theorem**[2]:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')} \tag{2.7}$$

下面是针对连续变量的，概率分布函数，概率密度的定义：

2.2.5 Continuous random variables

So far, we have only considered reasoning about uncertain discrete quantities. We will now show (following (Jaynes 2003, p107)) how to extend probability to reason about uncertain continuous quantities.

Suppose X is some uncertain continuous quantity. The probability that X lies in any interval $a \leq X \leq b$ can be computed as follows. Define the events $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$. We have that $B = A \vee W$, and since A and W are mutually exclusive, the sum rules gives

$$p(B) = p(A) + p(W) \tag{2.17}$$

and hence

$$p(W) = p(B) - p(A) \tag{2.18}$$

Define the function $F(q) \triangleq p(X \leq q)$. This is called the **cumulative distribution function** or **cdf** of X. This is obviously a monotonically increasing function. See Figure 2.3(a) for an example. Using this notation we have

$$p(a < X \leq b) = F(b) - F(a) \tag{2.19}$$

Now define $f(x) = \frac{d}{dx}F(x)$ (we assume this derivative exists); this is called the **probability density function** or **pdf**. See Figure 2.3(b) for an example. Given a pdf, we can compute the probability of a continuous variable being in a finite interval as follows:

$$P(a < X \leq b) = \int_a^b f(x)dx \tag{2.20}$$

As the size of the interval gets smaller, we can write

$$P(x \leq X \leq x + dx) \approx p(x)dx \tag{2.21}$$

We require $p(x) \geq 0$, but it is possible for $p(x) > 1$ for any given x, so long as the density integrates to 1. As an example, consider the **uniform distribution** $\text{Unif}(a, b)$:

$$\text{Unif}(x|a, b) = \frac{1}{b-a}\mathbb{I}(a \leq x \leq b) \tag{2.22}$$

If we set $a = 0$ and $b = \frac{1}{7}$, we have $p(x) = 2$ for any $x \in [0, \frac{1}{7}]$.

上面这些基本的东西看浙大的《概率论和数理统计》那本小书就行。

接着讲各种分布，大部分上午在 PRML 第二章已经讲过了，大家可以看那一章的讲课记录，现在再回顾下。

二项式分布：简答的例子就是抛硬币，抛 n 次有 k 次是正面或反面的概率

Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial** distribution, written as $X \sim \text{Bin}(n, \theta)$. The pmf is given by

$$\text{Bin}(k|n, \theta) \triangleq \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{2.28}$$

where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!} \tag{2.29}$$

is the number of ways to choose k items from n (this is known as the **binomial coefficient**, and is pronounced "n choose k"). See Figure 2.4 for some examples of the binomial distribution. This distribution has the following mean and variance:

$$\text{mean} = \theta, \quad \text{var} = n\theta(1-\theta) \tag{2.30}$$

伯努利分布：是二项式分布的特例，抛一次正反面的概率，theta 是参数

Now suppose we toss a coin only once. Let $X \in \{0, 1\}$ be a binary random variable, with probability of "success" or "heads" of θ. We say that X has a **Bernoulli** distribution. This is written as $X \sim \text{Ber}(\theta)$, where the pmf is defined as

$$\text{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \tag{2.31}$$

多项式分布：抛硬币变成了掷骰子

The binomial distribution can be used to model the outcomes of coin tosses. To model the outcomes of tossing a K-sided die, we can use the **multinomial** distribution. This is defined as follows: let $\mathbf{x} = (x_1, \ldots, x_K)$ be a random vector, where x_j is the number of times side j of the die occurs. Then $\mathbf{x}$ has the following pmf:

$$\text{Mu}(\mathbf{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j} \tag{2.33}$$

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!} \tag{2.34}$$

泊松分布：泊松分布是离散分布中仅次于二项式分布最重要的分布，在现实中广泛存在，此外还有泊松过程。

2.3.3 The Poisson distribution

We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$, written $X \sim \text{Poi}(\lambda)$, if its pmf is

$$\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \tag{2.39}$$

The first term is just the normalization constant, required to ensure the distribution sums to 1.

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents. See Figure 2.6 for some plots.

高斯分布：

Gaussian (normal) distribution

The most widely used distribution in statistics and machine learning is the Gaussian or normal distribution. Its pdf is given by

$$\mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{2.43}$$

Here $\mu = \mathbb{E}[X]$ is the mean (and mode), and $\sigma^2 = \text{var}[X]$ is the variance. $\sqrt{2\pi\sigma^2}$ is the normalization constant needed to ensure the density integrates to 1 (see Exercise 2.11).

Student t 分布：其实是无限个均值一样，方差不同的高斯分布混合而成，高斯分布是它的特例。

is the **Student** t **distribution**[5] Its pdf is as follows:

$$\mathcal{T}(x|\mu, \sigma^2, \nu) \propto \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\frac{\nu+1}{2})} \tag{2.51}$$

下面这个图将高斯分布、 t 分布 、拉普拉斯分布放在一块比较对孤立点干扰的鲁棒性，t 分布依然很淡定。

Figure 2.7 (a) The pdf's for a $\mathcal{N}(0, 1)$, $\mathcal{T}(0, 1, 1)$ and $\text{Lap}(0, 1/\sqrt{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$. (b) Log of these pdf's. Note that the Student distribution is not log-concave for any parameter value, unlike the Laplace distribution, which is always log-concave (and log-convex...) Nevertheless, both are unimodal. Figure generated by `studentLaplacePdfPlot`.

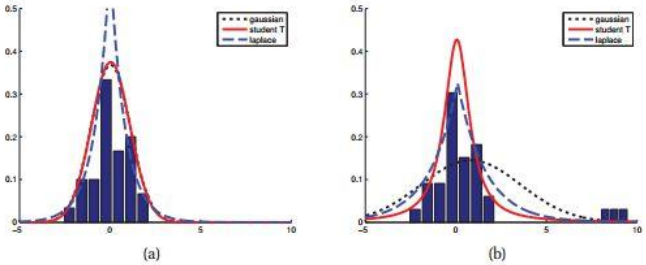


Figure 2.8 Illustration of the effect of outliers on fitting Gaussian, Student and Laplace distributions. (a) No outliers (the Gaussian and Student curves are on top of each other). (b) With outliers. We see that the Gaussian is more affected by outliers than the Student and Laplace distributions. Based on Figure 2.16 of (Bishop 2006a). Figure generated by `robustDemo`.

拉普拉斯分布：这是拉普拉斯在找随机误差的分布形式时，没找到正态分布找到了这个。

Another distribution with heavy tails is the **Laplace distribution**[6], also known as the **double sided exponential** distribution. This has the following pdf:

$$\text{Lap}(x|\mu, b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{2.53}$$

Here μ is a location parameter and $b > 0$ is a scale parameter. See Figure 2.7 for a plot. This distribution has the following properties:

$$\text{mean} = \mu,\ \text{mode} = \mu,\ \text{var} = 2b^2 \tag{2.54}$$

gamma 分布：可以看到里面有 gamma 函数，gamma 函数可以说在数理统计里面无处不在。

The **gamma distribution** is a flexible distribution for positive real valued rv's, $x > 0$. It is defined in terms of two parameters, called the shape $a > 0$ and the rate $b > 0$:[7]

$$\text{Ga}(T|\text{shape} = a, \text{rate} = b) \triangleq \frac{b^a}{\Gamma(a)}T^{a-1}e^{-Tb} \tag{2.55}$$

beta 分布：这是二项式的共轭分布，讲 PRML 时我们已经很熟悉了，里面也有 gamma 函数：

The **beta distribution** has support over the interval $[0, 1]$ and is defined as follows:

$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)}x^{a-1}(1-x)^{b-1}$$

Here $B(p, q)$ is the beta function,

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

beta 分布形式灵活多变：

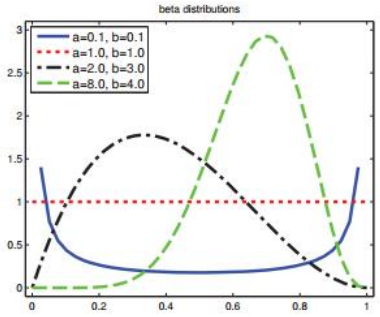


Figure 2.10 Some beta distributions. Figure generated by `betaPlotDemo`.

a and b are both less than 1, we get a bimodal distribution with "spikes" at 0 and 1; if a and b are both greater than 1, the distribution is unimodal. For later reference, we note that the distribution has the following properties (Exercise 2.16):

$$\text{mean} = \frac{a}{a+b},\ \text{mode} = \frac{a-1}{a+b-2},\ \text{var} = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.62}$$

下面讲帕累托分布，PRML 里面没有这个分布，这个分布不属于指数族

我们平时说的长尾分布，幂律分布可以表达成这种幂函数的形式：

The Pareto pdf is defined as follow:

$$\text{Pareto}(x|k, m) = km^k x^{-(k+1)}\mathbb{I}(x \geq m)$$

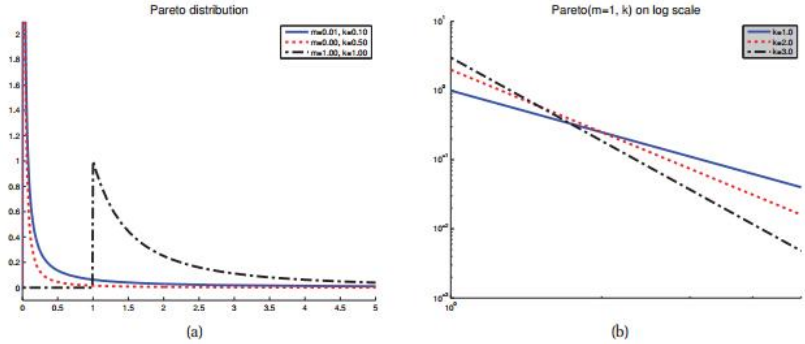
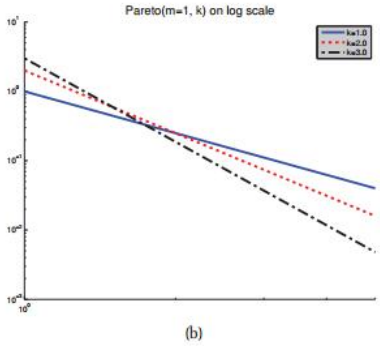


Figure 2.11 (a) The Pareto distribution Pareto$(x|m,k)$ for $m=1$. (b) The pdf on a log-log scale. Figure generated by `paretoPlot`.

大家看上面这个图的左边，长尾分布还是非常形象的。
有时这种分布和指数分布不容易区分，指数分布在下降过程中要陡一些。
如何判断幂律分布最简单的办法就是取双对数看是否能拟合成直线，就像下面这个图: :



为什么叫帕累托分布是因为著名的二八法则，帕累托发现意大利 20%的人口拥有 80%的财产。
幂律分布是一种最偏离正态分布的分布，正态分布大部分都集中在均值附近，所以有个均值做标度表示，但是幂律分布没有，所以我们经常听到人均。。。有时是没有代表性的。
尘绳聋(523201603) 15:16:06
get
东方朔(569920936) 15:16:44
所以我们经常听到人均。。。有时是没有代表性的。骤然明白了什么
van19(493010846) 15:17:15
屡屡被平均
网络上的尼采(813394698) 15:17:29
既然偏离正态分布这么厉害，从直觉上讲，产生正态分布的中心极限定理这时已经失效了，背后肯定隐藏着显著因素。关于幂律分布是如何生长出来的，这方面的解释有马太效应。圣经《新约·马太福音》：“凡有的，还要加给他叫他多余；没有的，连他所有的也要夺过来。”，也就是强者愈强，中国古代《道德经》里的“天之道，损有余而补不足；人之道则不然，损不足以奉有余。”也是说的这个原因。
黄浩军<littleblack1988@foxmail.com> 15:21:17
总裁班开课~
尘绳聋(523201603) 15:23:07
好厉害，一个分布也能讲出这么多所以然，我们学渣就只会记住公式是啥

东方朔(569920936) 15:23:30

神解释 

黄浩军<littleblack1988@foxmail.com> 15:23:31

dxhml(601765336) 15:23:42

是啊 

颜延(260930916) 15:23:56
网神的人文造诣也很高嘛
南(287663401) 15:24:00
学习了
BSS-DL(475795274) 15:24:30
神解释，服了
Shin-Chong(191162272) 15:24:34
这 2 句话 记下了
大头娃娃(283664823) 15:24:55
果然是尼采呀
好奇心害死薛定谔的猫(337025583) 15:25:20

gump(915537522) 15:26:06

网络上的尼采(813394698) 15:26:21
幂律分布在复杂网络研究里也很重要，一开始大家都认为现实中的网络度分布也和随机网络一样是指数分布，巴拉巴西发现无标度网的度分布是呈现幂律的，靠这个一举奠定了自己的地位。关于幂律分布的生长模型有很多，可以用一些随机过程的方法比如马氏链来证明最后收敛到幂律，大家可以找来看看。
<(523723864) 15:28:10

 我只知道一个 BA 模型...

网络上的尼采(813394698) 15:28:21
嗯，就是 BA
继续
多元高斯分布：

The multivariate Gaussian

The **multivariate Gaussian** or **multivariate normal** (**MVN**) is the most widely used joint probability density function for continuous variables. We discuss MVNs in detail in Chapter 4; here we just give some definitions and plots.

The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2.70}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ is the mean vector, and $\boldsymbol{\Sigma} = \mathrm{cov}[\mathbf{x}]$ is the $D \times D$ covariance matrix. Sometimes we will work in terms of the **precision matrix** or **concentration matrix** instead. This is just the inverse covariance matrix, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. The normalization constant $(2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}$ just ensures that the pdf integrates to 1 (see Exercise 4.5).

Figure 2.13 plots some MVN densities in 2d for three different kinds of covariance matrices. A full covariance matrix has $D(D+1)/2$ parameters (we divide by 2 since $\boldsymbol{\Sigma}$ is symmetric). A diagonal covariance matrix has D parameters, and has 0s in the off-diagonal terms. A **spherical** or **isotropic** covariance, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_D$, has one free parameter.

多元 t 分布：

Multivariate Student t distribution

A more robust alternative to the MVN is the **multivariate Student t** distribution, whose pdf is given by

$$\mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{\nu^{D/2}\pi^{D/2}} \times \left[1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \tag{2.71}$$

$$= \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)}|\pi\mathbf{V}|^{-1/2} \times \left[1 + (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})} \tag{2.72}$$

多项式分布的共轭分布狄利克雷分布：

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}\mathbb{I}(\mathbf{x} \in S_K)$$

生成高斯分布的中心极限定理：

Now consider N random variables with pdf's (not necessarily Gaussian) $p(x_i)$, each with mean μ and variance σ^2. We assume each variable is **independent and identically distributed** or **iid** for short. Let $S_N = \sum_{i=1}^{N} X_i$ be the sum of the rv's. This is a simple but widely used transformation of rv's. One can show that, as N increases, the distribution of this sum approaches

$$p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s-N\mu)^2}{2N\sigma^2}\right) \tag{2.96}$$

Hence the distribution of the quantity

$$Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \tag{2.97}$$

converges to the standard normal, where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the sample mean. This is called the **central limit theorem**. See e.g., (Jaynes 2003, p222) or (Rice 1995, p169) for a proof.

下面是 Monte Carlo 方法：

MC 是以随机为基础的，最早的 MC 方法是二战时美国研制原子弹算积分时发明的。

这章介绍的是一个容易理解的近似求圆周率的例子：在单位矩形里随机撒豆子，用落在圆里的与所有的比例估算出圆的面积从而得到 pi 的近似值。

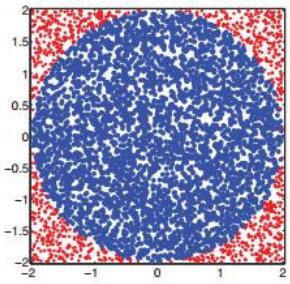


Figure 2.19 Estimating π by Monte Carlo integration. Blue points are inside the circle, red crosses are outside. Figure generated by `mcEstimatePi`.

MCMC（Markov Chain Monte Carlo）方法我在 PRML 第十一章 Sampling 时已经详细讲了，感兴趣的同学可以看下原来的记录。

信息熵的公式：

The **entropy** of a random variable X with distribution p, denoted by $\mathbb{H}(X)$ or sometimes $\mathbb{H}(p)$, is a measure of its uncertainty. In particular, for a discrete variable with K states, it is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) \tag{2.107}$$

KL 散度，表示两个分布间的差别:

KL divergence

One way to measure the dissimilarity of two probability distributions, p and q, is known as the **Kullback-Leibler divergence** (**KL divergence**) or **relative entropy**. This is defined as follows:

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \tag{2.110}$$

where the sum gets replaced by an integral for pdfs.[10] We can rewrite this as

$$\mathbb{KL}(p||q) = \sum_k p_k \log p_k - \sum_k p_k \log q_k = -\mathbb{H}(p) + \mathbb{H}(p,q) \tag{2.111}$$

where $\mathbb{H}(p,q)$ is called the **cross entropy**,

$$\mathbb{H}(p,q) \triangleq -\sum_k p_k \log q_k \tag{2.112}$$

可以用 jensen 不等式来证明 KL 是非负的，这个性质会在以后的 EM 和变分推断中用到:

Theorem 2.8.1. (**Information inequality**) $\mathbb{KL}(p||q) \geq 0$ *with equality iff* $p = q$.

Proof. To prove the theorem, we need to use **Jensen's inequality**. This states that, for any convex function f, we have that

$$f\left(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{n} \lambda_i f(\mathbf{x}_i) \tag{2.113}$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. This is clearly true for $n = 2$ (by definition of convexity), and can be proved by induction for $n > 2$.

Let us now prove the main theorem, following (Cover and Thomas 2006, p28). Let $A = \{x : p(x) > 0\}$ be the support of $p(x)$. Then

$$-\mathbb{KL}(p||q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \tag{2.114}$$

$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} = \log \sum_{x \in A} q(x) \tag{2.115}$$

$$\leq \log \sum_{x \in \mathcal{X}} q(x) = \log 1 = 0 \tag{2.116}$$