PRML（Pattern Recognition And Machine Learning）读书会

# 第一章　Introduction

主讲人　常象宇

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

# Introduction to Machine Learning

Likrain

May 4, 2013



大家好，我是 likrain，本来我和网神说的是我可以作为机动，大家不想讲哪里我可以试试，结果大家不想讲第一章。估计都是大神觉得第一章比较简单，所以就由我来吧。我的背景是统计与数学，稍懂些计算机，大家以后有问题可以讨论。

今天我们来讲一下 PRML 第一章，这一章的内容是基于一些简单的例子对于机器学习中的基本概念给与介绍。这是为后续章节的介绍给一个铺垫。

我今天讲的内容包括以下几个部分：

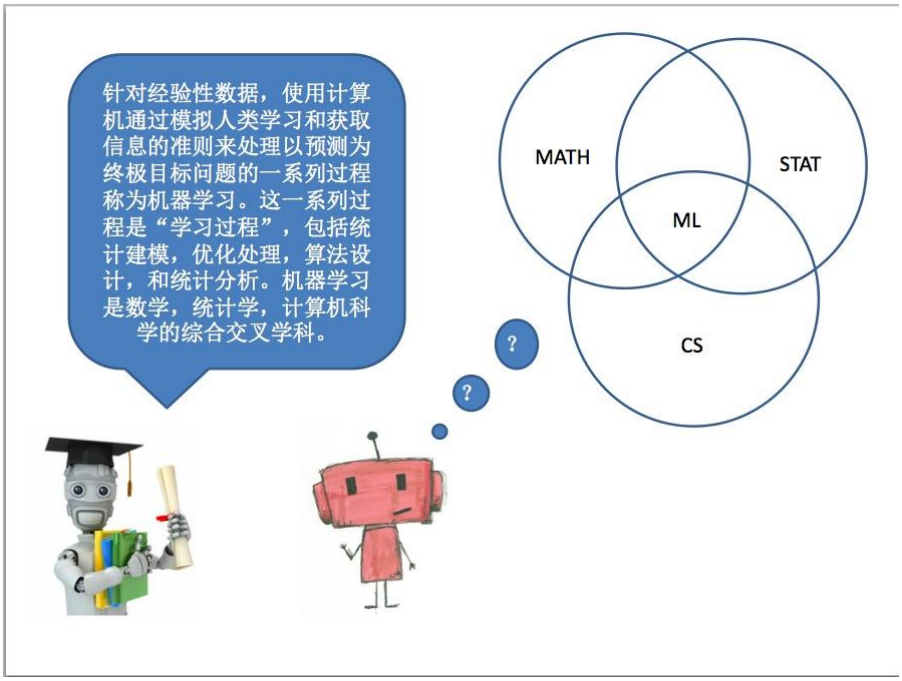
## Outline





把书上的知识点做了个总结大概。

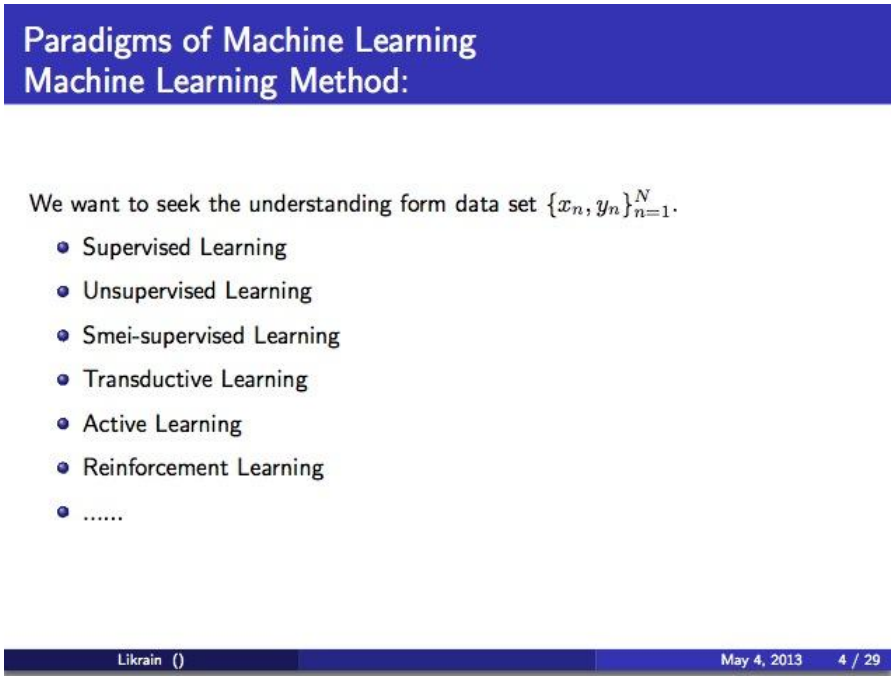
首先我们来看一下，我个人理解的机器学习的定义：




机器学习的分类有很多种，一般是基于两点：数据类型与学习过程。

是否有标签->监督（分类，回归），半监督，无监督（聚类）；

学习过程不同->主动学习，强化学习，转导学习。




==========讨论==========

这里我可以稍微停一下

是否有问题？

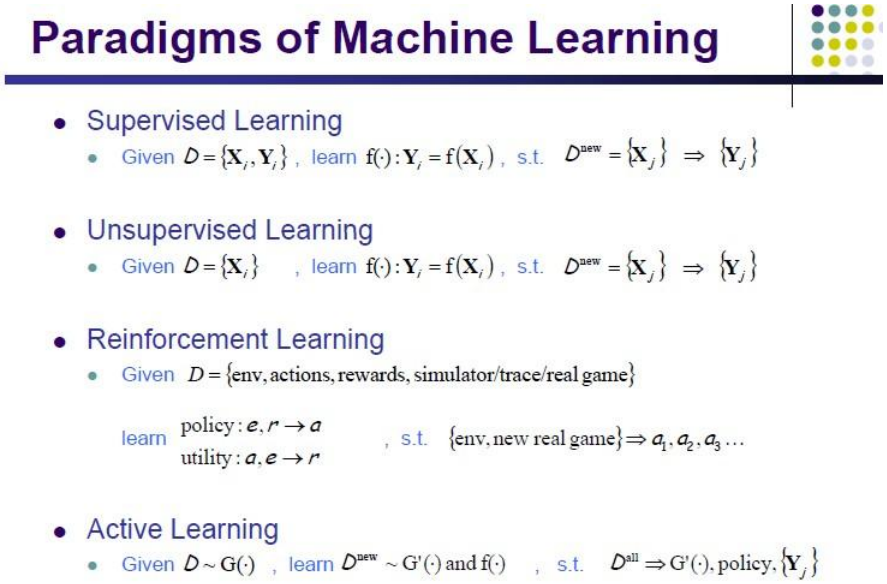
planktonli(1027753147) 19:01:30

这里讲的要更清楚些哦，learning 的区别啊

从(373242125) 19:01:36

$\{x_n, y_n\}_{n=1}^{N}$.

HX(458728037) 19:01:57
学习过程不同->主动学习，强化学习，转导学习是怎么分的呢？
likrain(261146056) 19:02:22
提问结束否？
大家如果没问题了我就开始回答
planktonli(1027753147) 19:02:50




likrain(261146056) 19:02:59
第一问题@planktonli 我不是很清楚哈哈
这样大家如果问题都说完了，我就开始解释，确定可以开始了？
从(373242125) 19:04:38

Unsupervised Learning 无监督学习,只能用于类聚?

网神(66707180) 19:04:43
如果回答的内容在接下来的讲课计划里，可以先不回答。
likrain(261146056) 19:05:32
第一问题@planktonli 我没明白他想问什么，
第二个@丛：特指我有 N 个样本，用 x_n,y_n 代表其中一个
第三个@HX：学习过程可以想象成一种学习机制，这种机制是模拟人类的学习或者想法的一种过程
然后，我对于第三个问题举个例子
planktonli(1027753147) 19:06:42
第二个@丛: 无监督学习包括: clustering, dimension reduction, density estimization
likrain(261146056) 19:07:02
例如转导学习：大家可以这样想象，我们对于一个学习过程总是希望得到一个规则，数学上与形式上，而我们人经常判断不是这样的
你是看到了一些苹果之后，给你了个苹果你去判断，你不是个机器需要有个函数规则，而是通过你看到的以前的苹果直接判断苹果，基于这种想法的学习过程叫做转导学习，例如你可以
google, trunsductive SVM

planktonli(1027753147) 19:08:33
inductive learning 是和 tranductive learning 区别的
likrain(261146056) 19:08:36
我先发言到这里，这个。。。。区别很明显
planktonli(1027753147) 19:09:09
我补充下，inductive learning 主要用于 semi-supervised leraning
likrain(261146056) 19:09:26
不好意思我看错了哈哈，我理解这是一个东西，然后换了个词，不知道其他人意见

==========讨论结束==========
继续
学习理论：一套标准的框架，用统计学，概率论，数学的严格化语言去解释（收敛速率与泛化性能）或者比较不同学习方法与模型的性能。其中最经典的例子：统计学习理论。
统计学习理论的目标：去研究所谓的泛化误差界（Generalization Bounds）

## Paradigms of Machine Learning
## Learning Theory:

Generalization error:

$$\sup_{f \in H} \{\mathcal{E}(f) - \mathcal{E}_N(f)\} \leq B(N, H, \beta, \sigma, \cdots)$$

- Consistency
- Bias vs Variance
- Enough Sample Size
- Learning Rate
- Convergence
- Stability
- Confidence
- .....

Likrain () May 4, 2013 5 / 29

这是一些理论涉及到的 topic，这里面涉及到的数学技巧与数理统计工具比较多，我没法全部解释

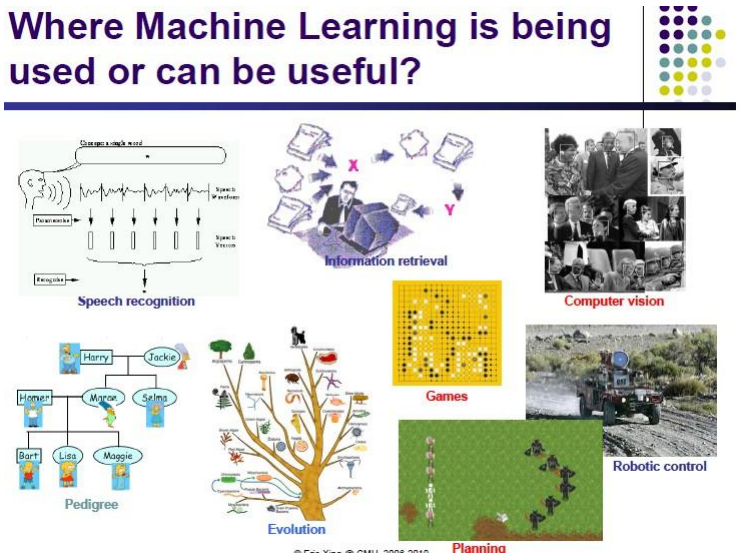
## International Machine Learning Groups
## Machine Group:

- Berkeley
- MIT
- CMU
- CSML
- Groups in China
- .....

Likrain () May 4, 2013 6 / 29

机器学习的研究小组：美国，欧洲，中国例子。伯克利的研究小组比较厉害，当然大家主要知道出名的，

所谓的乔丹哈哈，MIT 的人工智能实验室、CMU 的 machine learning deparment、欧洲的 CSML 放在 UCL，还有我们国内的是吧。。。哈哈
机器学习的例子已经深入到我们每天的生活，这里我们可以讨论一下生活中机器学习的例子。大家发言呵呵
planktonli(1027753147) 19:15:43

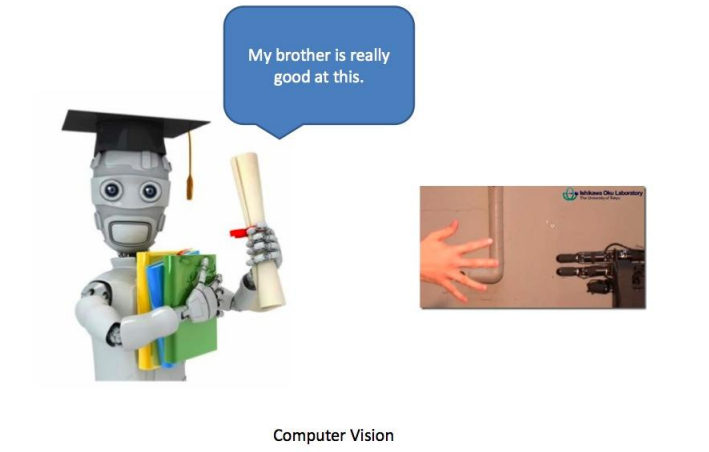



要不
likrain(261146056) 19:17:17
那我讲个例子吧：

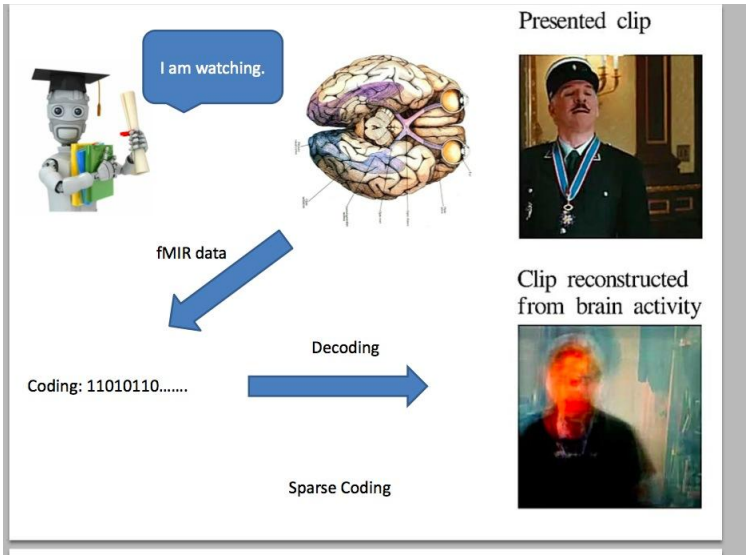
## Advanced Examples

- 野球拳
- 读心术
- Siri

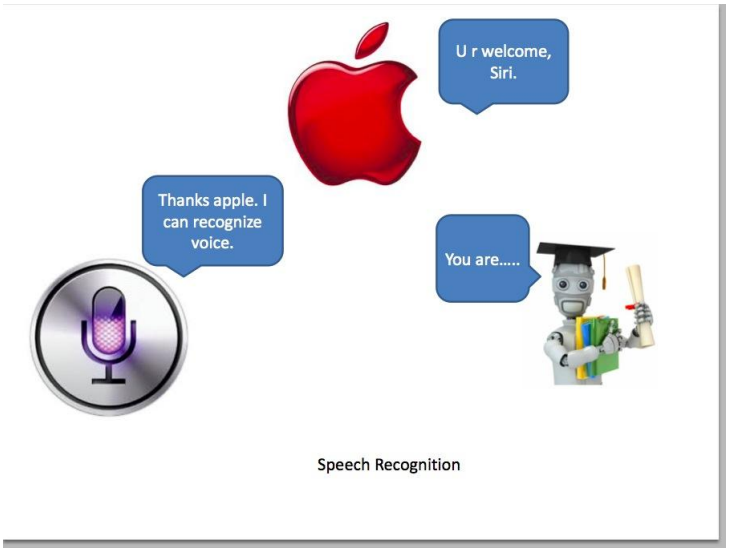
我比较推崇的第一个野球拳：




读心术：




siri：

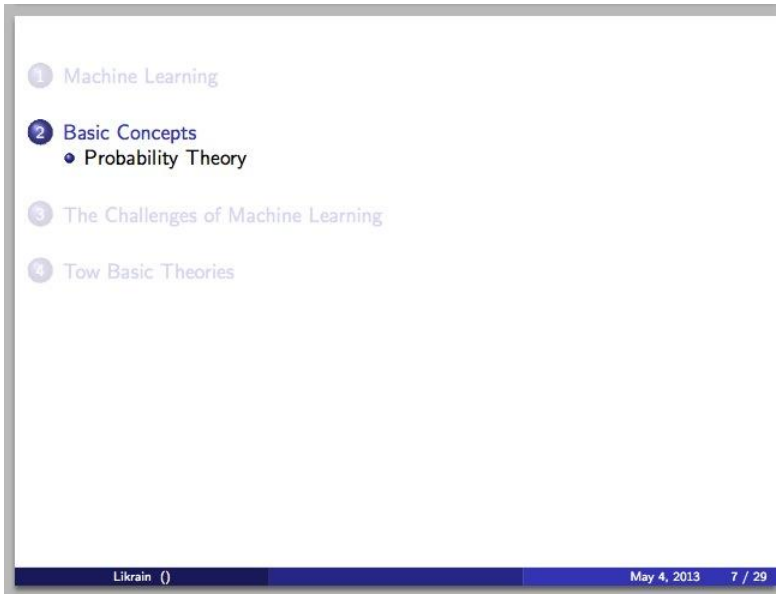


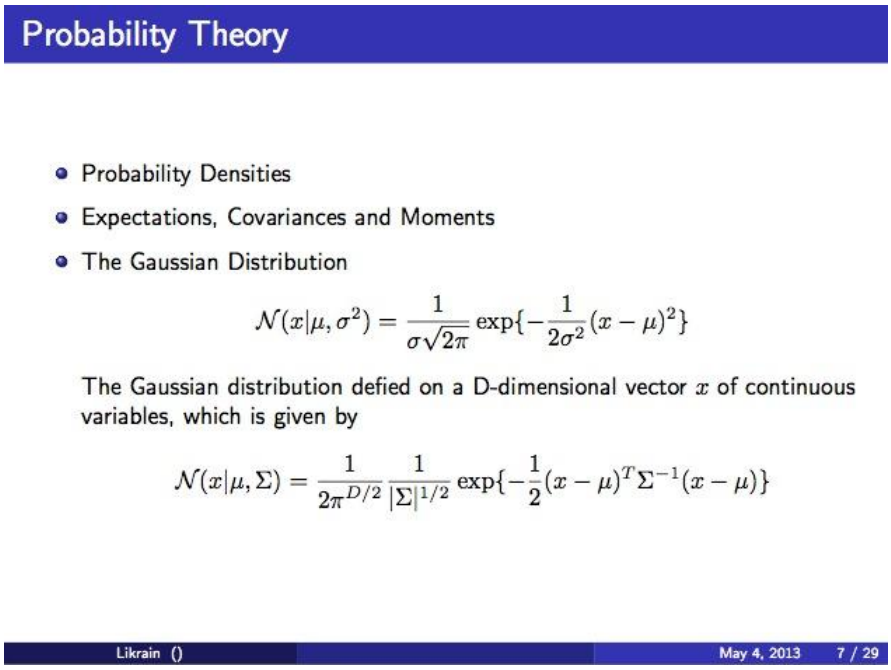
这是我以前讲东西的课件，我觉得这三个例子比较有意思。第一个野球拳，男生懂的哈哈，游戏，通过计算机视觉来完成的高速相机拍摄；第二个是最近伯克利刚刚完成的，使用 fMIR 的数据，预测你看到的电影图像。也就说建立一种预测机制，从你的脑部的 fMIR 数据，把你看到的视频解码，所以说读心术，具体方法以后可以介绍。一个潜在的应用就是说，以后不用测谎仪了，我直接可以根据你的脑部的 fMRI 数据去看你看过什么东西。

进入正题

第一部分：概率基本概念

我认为比较简单，我把一些基本知识点贴上来，然后大家讨论。




我就把书上的抄了下哈

有问题可以提哈

满阶落叶(1316061080) 19:25:06

怎么把高斯分布转到 D 维的？

likrain(261146056) 19:25:33

这个需要推导，我举个例子

planktonli(1027753147) 19:25:46

**Expectation**: the average value of a function $f(x)$ under a pdf or pmf $p(x)$.

$$\mathbb{E}[f] = \sum_x p(x)f(x), \text{ or } \mathbb{E}[f] = \int p(x)f(x)dx.$$

Given $\{x_n\} i.i.d. \sim p(x)$, the expectation can be approximated by

$$\mathbb{E}[f] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n).$$

**Conditional expectation** w.r.t. a conditional distribution

$$\mathbb{E}_x[f|y] = \int_x p(x|y)f(x)dx,$$

which is a function of $y$.

A nice property: $\mathbb{E}[\mathbb{E}[x|y]] = \mathbb{E}[x]$.

likrain(261146056) 19:25:50

首先你有 x1,x2,x3...xn，独立同分布是高斯

planktonli(1027753147) 19:26:17

**Variance** of $f(x)$ is defined by

$$\text{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right]$$

which can be rewritten as

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

In particular, variance of a random variable

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

likrain(261146056) 19:26:28

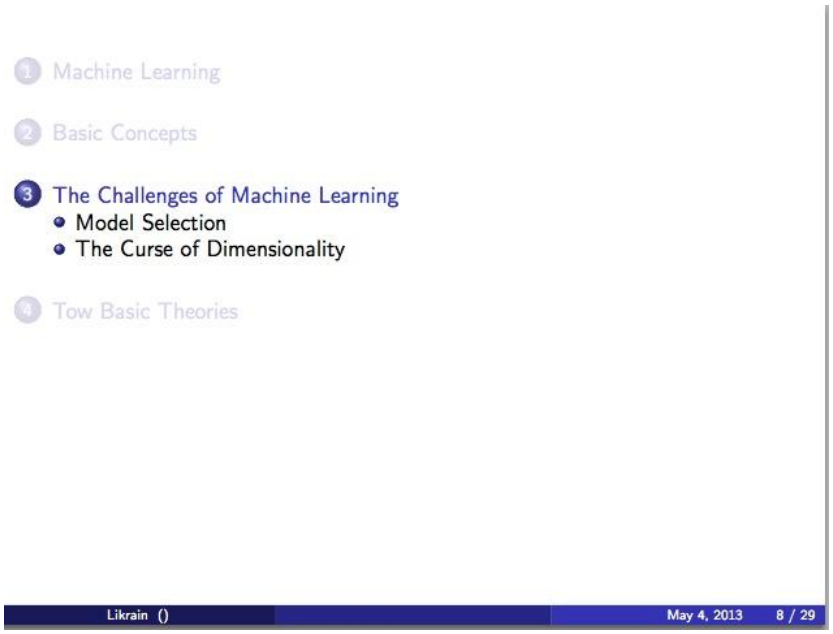
你把他们的密度函数都乘起来就是他们的密度函数，然后如果他们不是标准正态分布，那么你做个变换 就可以，下面就是会出现相关性怎么处理，也可以做个变换转化为独立的。。。我大概说了下过程。。。。

满阶落叶(1316061080) 19:28:35

好的

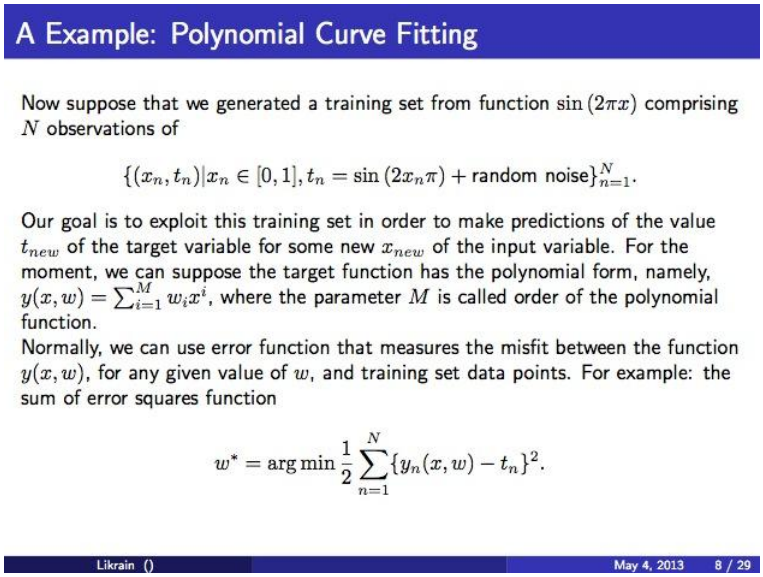
likrain(261146056) 19:28:56

ok，那我继续哈

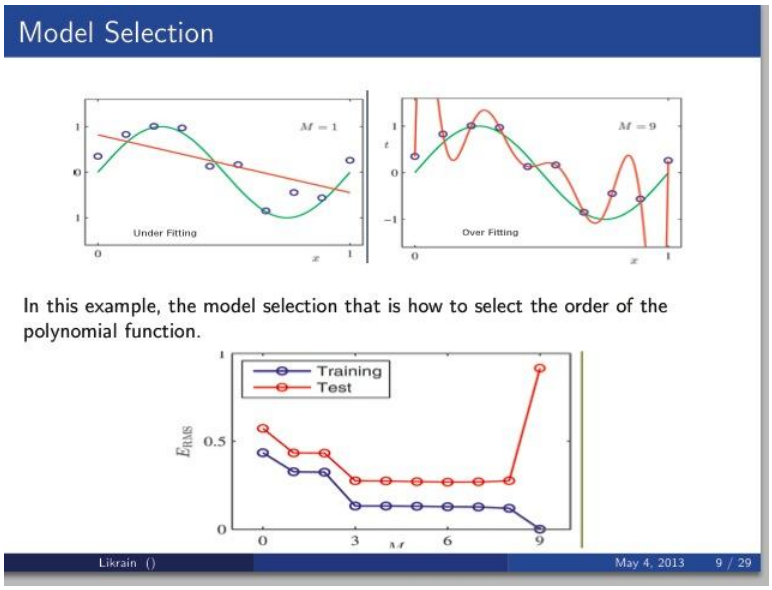



第二部分：模型选择与高维灾难

PRML 中使用多项式拟合的例子引出了模型选择的概念。这里大家可以从两个方面来理解。对于这个例子存在两个模型选择问题：第一函数空间的选择：即去拟合三角函数为何使用了多项式函数。第二确定使用多项式拟合后，多项式的阶数问题。



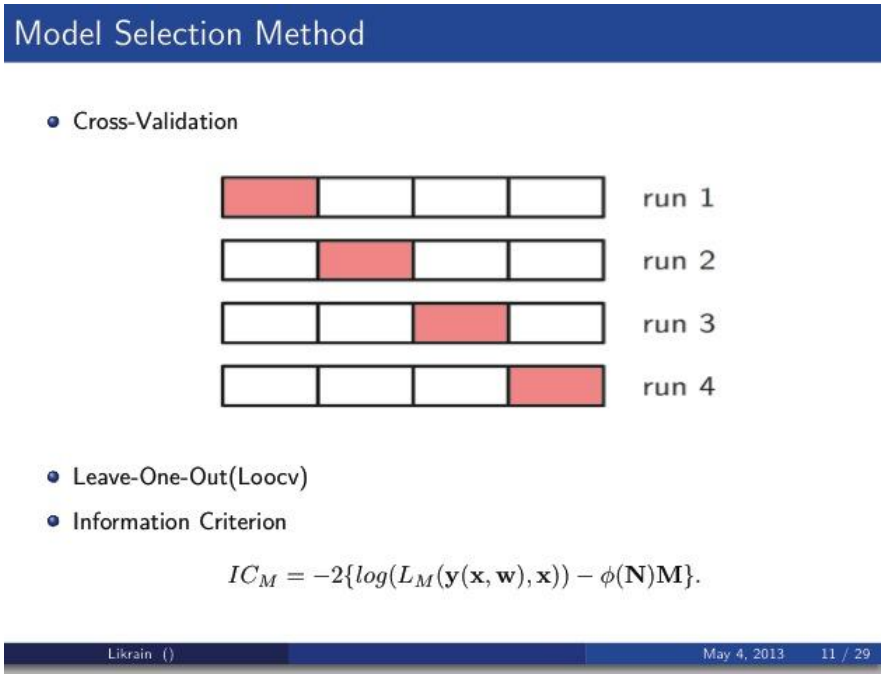

进一步的，例子中给出了使用不同阶数去拟合该三角函数的如何产生过拟合现象。避免过拟合现象的的基本方法：正则化方法，regularization.




理解正则化方法有很多，这里我解释其中一种，正则化方法你可以理解为，我对于我选择函数空间的函数做了一种限制。使得这个函数空间比原来的函数空间小，所以不会把过分拟合的函数选择进入需要的函数空间。

例如：例子中使用了多项式空间，但是加入正则化之后等价于对于一些函数空间的限制，那些过分拟合的 9 次多项式不会被选择进来。

模型选择方法：必须要懂的至少要有交叉验证，AIC。




==========讨论==========

这里我先讲到这里哈哈，大家可以开始发问了哈，这里比较模糊。。。。。很难文字说的特别清楚，如果不做预习，可能以前没看过的人就比较困难了。

范涛@推荐系统(289765648) 19:35:33

做回归分析也有必要做交叉验证吗？

满阶落叶(1316061080) 19:35:47
交叉验证能稍微解释一下嘛？举个例子。。
范涛@推荐系统(289765648) 19:35:57
回归分析时候，不是有假设检验码，看 p-value 不可以吗？
HX(458728037) 19:36:14
cross validation 是一种正则化的方法吗？
lost(549294286) 19:37:49
AIC？
likrain(261146056) 19:38:23
@范涛：交叉验证只是模型选择的一种方法，如果你有模型选择问题，你就可以用交叉验证。例如你做线性回归，你有 10 个变量，你就有 1024 个模型需要选择，你就可以使用交叉验证或者 AIC，做任何一件事情，都会从不同的目的去做，使用交叉验证是从预测的角度去做，使用 AIC 是从模型的复杂度与模型的拟合角度去做。
范涛@推荐系统(289765648) 19:40:20
那我们回归分析时候，经常会看变量的 p-value
likrain(261146056) 19:40:23
使用 p-value 使用假设检验的角度去做，模型选择都是选择方法。
范涛@推荐系统(289765648) 19:40:36
哦，明白
likrain(261146056) 19:40:43
只是传统的统计学里面对于线性模型给出了这种方式
lost(549294286) 19:40:58
AIC，最小信息准则。。
likrain(261146056) 19:41:07
可能是这个名字吧，还有 BIC、EBIC 反正很多
DM-福(864914991) 19:41:30
$$IC_M = -2\{log(L_M(\mathbf{y}(\mathbf{x},\mathbf{w}),\mathbf{x})) - \phi(\mathbf{N})\mathbf{M}\}.$$ 这个是信息熵吗
likrain(261146056) 19:41:37
不是，第一项是似然函数，第二项是模型复杂度
范涛@推荐系统(289765648) 19:42:01
交叉验证，现在有没有加速方法， 变量多的话，模型成指数增长啊
DM-福(864914991) 19:42:10
信息增益好像也是这样写的
likrain(261146056) 19:42:39
指数增长？不对吧哈哈，加速方法我没感觉到有哈哈
DM-福(864914991) 19:43:00
2^n
范涛@推荐系统(289765648) 19:43:27
说错了，恩 2^n
likrain(261146056) 19:43:35
信息熵是有的，交叉验证没有
范涛@推荐系统(289765648) 19:44:09
2^n 个模型，都验证，貌似不现实

likrain(261146056) 19:44:13
所以信息熵可以有办法改进，我可以告诉你们一些 paper
范涛@推荐系统(289765648) 19:44:37
现在在互联网上都是至少几百万维阿
likrain(261146056) 19:44:57
所以都不是这么做的，互联网哈哈，那个啥。。。。。家里吃饭中哈哈，咱们稍微暂停哈哈，  抱歉
网络上的尼采(813394698) 19:45:51
好，你先去吃饭，现在自由讨论
likrain(261146056) 19:45:56
嗯我回来解释哈哈
苍并空指导老师(794717493) 19:46:14
@likrain　例如：例子中使用了多项式空间，但是加入正则化之后等价于对于一些函数空间的限制，那些过分拟合的 9 次多项式不会被选择进来。9 次项也出现吧。。。只是系数比较小
HX(458728037) 19:46:26
这个公式估计就是交叉验证希望的目标函数最小的公式吧
苍并空指导老师(794717493) 19:48:51
大家对机器学习研究的定位怎么看。。。。很多问题在数学上统计上都是很古老的问题了
范涛@推荐系统(289765648) 19:49:24
的确是这样的，我只关注互联网上应用
网神(66707180) 19:49:44
机器学习热主要是现在互联网应用引起的，因为互联网有大量的数据，这是以前的机器学习或数据挖掘缺少的。
范涛@推荐系统(289765648) 19:50:19
还有互联网数据规模带来的一系列新问题
HX(458728037) 19:50:32
你意思是说常规的很多机器学习算法在 互联网的领域根本没法用？
范涛@推荐系统(289765648) 19:50:45
很多事这样的，尤其那种时间复杂度高的
HX(458728037) 19:51:33
那互联网一般都是怎么处理的呢？降维吗？
范涛@推荐系统(289765648) 19:52:26
这个得找群里那些资深人士解读下了
HX(458728037) 19:52:26
还有就是那你说的维数很大 会有多大呢
丛(373242125) 19:52:49
文本分类中,有几万维吧.
planktonli(1027753147) 19:52:59
恩,google 的有这么大，他们采用的是 naive bayes + distributed computing
范涛@推荐系统(289765648) 19:53:26
文本几万维不止吧
Xiaoming(50777501) 19:53:35
既然是第一章，我们说说大方向问题吧。现在很多 ML 方法都涉及统计。大家总体感觉，说说统计究竟是不是一个正确的方向？

范涛@推荐系统(289765648) 19:53:42
几万维现在貌似真不算什么
planktonli(1027753147) 19:53:53
google 一次要跑 1TB 的数据，要一次性载入
HX(458728037) 19:54:21
我没做过文本 都是做图像，图像里面的特征 最大也就是原图的像素的数量，为什么文本的会有那么大的维数呢？
从(373242125) 19:55:01
词向量
阿邦(1549614810) 19:55:04
n gram
苍并空指导老师(794717493) 19:55:06
因为词汇量
HX(458728037) 19:55:30
文本 怎么转化为向量的？
苍并空指导老师(794717493) 19:55:42
词频
从(373242125) 19:55:42
中文分词
planktonli(1027753147) 19:55:45
To Xiaoming: 统计的东西现在占主流，还有一些 优化+几何的东西
网神(66707180) 19:56:22
请问 HX 图像提取像素特征，性能上有没有测试过，比如提取特征平均时间在什么量级，几百微妙还是几十毫秒？
范涛@推荐系统(289765648) 19:56:44
最优化蛮有用的
HX(458728037) 19:57:06
图像特征有简单的也有复杂的，然后还要看原图的分辨率大小，但是 还真没测试到微妙级别过
范涛@推荐系统(289765648) 19:57:43
彩色图像维度也挺大的吧
网神(66707180) 19:58:05
都在几十毫秒-几百毫秒范围吗
planktonli(1027753147) 19:58:44
图像特征看是什么了，RGB histgoram,纹理的快些，SIFT\shape context 这些速度都不快
HX(458728037) 19:58:54
假如是一张 800*600 像素的图像，那么最大的特征也就是 800*600*3 这么大，一般不可能再大了
Xiaoming(50777501) 20:00:26
To planktonli：是啊。感觉统计和几何是能不是喔解决问题，但是“Learning" 的本质并不是一定基于统计，这个很难探讨。现在是，统计能解决到一些问题，就说机器可以学习，就盖上“学习”的帽子了。但本质呢，不一定的。
HX(458728037) 20:01:24
我怎么感觉机器学习 全都是建立在统计上面做的
从(373242125) 20:01:43
神经网络训练出来的模型,人类容易理解么?

planktonli(1027753147) 20:01:57
To Xiaoming: 不过统计本身在表达很多现实世界的时候有一定优势，因此 出现了 统计物理\统计金融 ，其实 机器学习现在可以看成 统计计算机的东西
苍并空指导老师(794717493) 20:02:20
人类为什么要理解 @从
Xiaoming(50777501) 20:02:27
To HX：如果这样，应该叫"机器统计 “，而不是机器学习。
从(373242125) 20:02:58
希望计算出来的模型,是否合符人类的逻辑.
HX(458728037) 20:03:03
可以举出一个例子 某种分类器的算法不是基于统计的？
苍并空指导老师(794717493) 20:03:17
感觉 机器学习 就是 高级数据拟合
planktonli(1027753147) 20:03:27
很多啊，neural network/support vector machine
从(373242125) 20:03:48
或者说,人类希望从计算出来的模型提取有用的信息.
planktonli(1027753147) 20:03:49
可以看成优化的
Xiaoming(50777501) 20:03:52
@从 神经网络 是不基于统计，但并不代表现有的 NN 就是机器学习的正解。
planktonli(1027753147) 20:04:02
其实数学的很多分枝也是交融的
HX(458728037) 20:04:06
SVM 不是典型的基于统计的吗？
苍并空指导老师(794717493) 20:04:36
没有吧。。。
likrain(261146056) 20:04:47
我回来了 哈哈
逆风飞扬(374350284) 20:04:52
统计学习理论
likrain(261146056) 20:04:53
大家可以总结一下争论的问题吗？
planktonli(1027753147) 20:05:03
SVM 只是统计的解释多些，真正的求解和 model 是优化的东西的
likrain(261146056) 20:05:07
主持可以发个言 ，哈哈
网神(66707180) 20:06:14
水平有限，大家好像在讨论机器学习完全是基于统计的吗
likrain(261146056) 20:06:19
看了一下大家争论的问题
范涛@推荐系统(289765648) 20:06:19
按这样说，各种回归，分类器的参数计算，哪个不是通过最优化训练求救出来的

likrain(261146056) 20:06:28

我觉得本身的争论可能不需要，因为我刚才写的定义，大家可能需要看看。

针对经验性数据,使用计算机通过模拟人类学习和获取信息的准则来处理以预测为终极目标问题的一系列过程称为机器学习。这一系列过程是“学习过程”,包括统计建模,优化处理,算法设计和统计分析，几乎所有的机器学习的模型都有统计解释，所以叫做统计建模。

有了模型，就要去学习一些目的，这些目的得到了优化模型，解优化模型 ，就要设计算法，完成算法就需要写 code，所以本来就是交叉学科。最后分析理论。

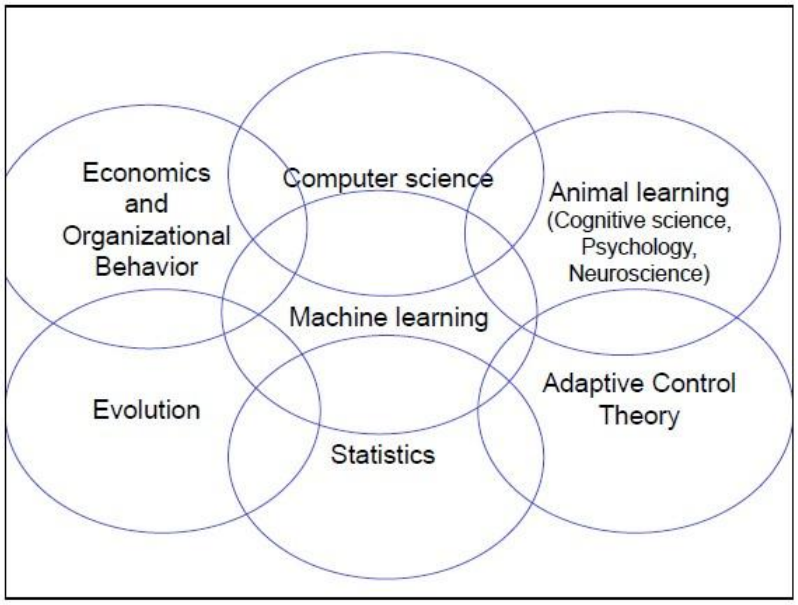
苍井空指导老师(794717493) 20:08:39

理解成高级数据拟合可以么？

likrain(261146056) 20:08:52

远远超出拟合，因为遇到的问题，已经远远超出了拟合的范畴。

planktonli(1027753147) 20:09:21




==========讨论结束==========

likrain(261146056) 20:10:28

## The relationship of the Lease Squares and Maximum Likelihood

First, we assume that, given the value of $x$, the corresponding value of $t$ has a Gaussian distribution with a mean equal to the value $y(x, w)$ of the polynomial curve. We have

$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1}).$$

Likelihood Function:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{\mathbf{n}=1}^{\mathbf{N}} \mathcal{N}(\mathbf{t_n}|\mathbf{y}(\mathbf{x_n}, \mathbf{w}), \beta^{-1}).$$

Maximum Likelihood:

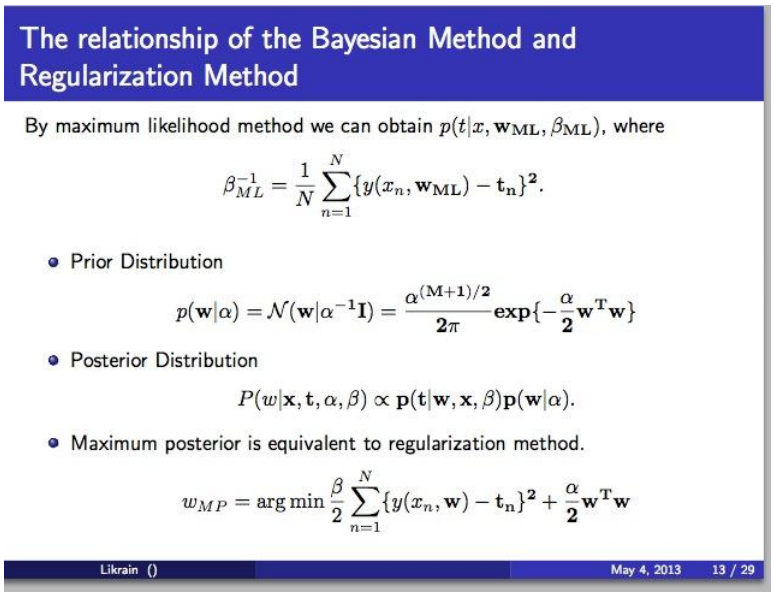
$$w_{ML} = \arg\max \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \arg\max\{-\frac{\beta}{2}\sum_{\mathbf{n}=1}^{\mathbf{N}}\{\mathbf{y}(\mathbf{x_n}, \mathbf{w}) - \mathbf{t_n}\}^2\}.$$

Maximum Likelihood method under the assumption that error term has the Gaussian distribution equal to minimizing the sum of squares error function.

网神(66707180) 20:10:31

欢迎 likrain 继续

likrain(261146056) 20:10:39




likrain(261146056) 20:10:44

两个重要观点：

最小二乘数学建模等价于高斯噪声最大释然估计统计建模，正则化最小二成等价于基于高斯噪声的最大化后验概率统计建模。

这里就像我说的，几乎所有的机器学习方法也许建立之初没有什么统计解释，最后大家发现，都可以通过统计的原理解释 。

所以这里只是两个简单的例子，所以因为这些观点，所以也就只需要从统计建模，建立机器学习模型就好了，所以就是统计机器学习。

planktonli(1027753147) 20:12:25

我感觉这样是不是有些牵强啊，先有了方法,然后用统计的东西给个模型解释。

likrain(261146056) 20:12:41

针对数据的建模过程，基于概率分布的建模过程，都是最后解释成了统计机器学习，发挥的淋漓尽致的就是 graphic model。

@planktonli：所以既然都有统计解释，为什么不直接从统计上直接建模呢，这本书基本就是这个观点，所以才要说 generative model，所谓生成模型。

planktonli(1027753147) 20:14:36

@likrain,明白

范涛@推荐系统(289765648) 20:14:40

回到最小二乘和最大似然的关系吧

likrain(261146056) 20:14:41

我刚才发的 ppt 有问题吗？

HX(458728037) 20:14:53

generative model 和 discriminative model

范涛@推荐系统(289765648) 20:14:56

都是 loss fucntion 问题，是吧？

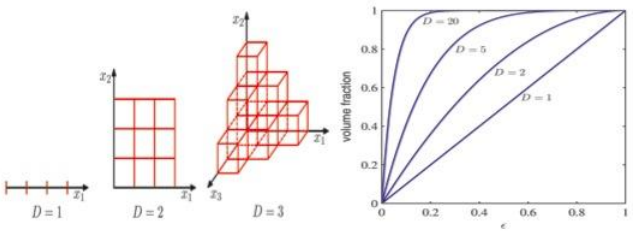
likrain(261146056) 20:15:23
loss function?
planktonli(1027753147) 20:15:35

likrain(261146056) 20:15:37
为什么突然提到这里？
planktonli(1027753147) 20:15:55
都是 objective function
likrain(261146056) 20:16:06
你可以理解为：l_2 loss function = gaussian noise
所以就是统计解释
范涛@推荐系统( (289765648) 20:16:44
我理解的最小二乘，无非就是求解模型参数的方法
likrain(261146056) 20:16:53
是的，这个就是数学建模 ，你可以想想，牛顿和你的理解是一样的，所以发明了牛顿第二定律。
而统计学家说 ok：我给你个统计解释，只要是高斯噪声，对应的从最大释然估计就是最小二乘，所以这是统计建模。
lost(549294286) 20:18:09

likrain(261146056) 20:18:29
所以如果你的模型是个线性回归，你的 noise 是拉普拉斯
范涛@推荐系统(289765648) 20:18:38
我们做回归分析，一个好的模型出来的，残差分布最好是个正态分布？？
likrain(261146056) 20:18:45
你如果用最小二乘，你就完了，正确的应该用最小一乘，叫做 LAD， 这是一个非常大的领域。
机器学习上面叫做=误差建模，统计上面=稳健估计
我回答这个问题结束。。。。。
维数灾难：

## The Curse of Dimensionality

- How to select the current model? There are many models can be selected. For example:$x_1, x_2, x_3, 4^3$.

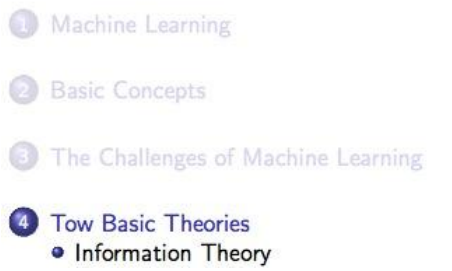



- In hight dimension vector space, we can see anything that is not clear.

两个方面：
第一， 模型的复杂性。
由于维数很大，简单的例子就是如果我们有 n 个变量那么我们如果回归也有 2^n 个模型
使用多项式函数空间 ，所以这几乎很难做到，如果我们函数空间选的更大，那么几乎是无法完成的。
第二， 几何体的难以想象的各种突变
这个地方没有很多经验我个人觉得比较难理解，你如何想象高维空间中的球体的数据，其实都集中在球壳附近。。。。。
如何想象高维空间的各种几何体，其实和三维空间中的完全不一样。我没有什么好的建议，如果大家真的想看看，就去学学 Functional Geometrical Analysis. 至于书上的例子，我不知道大家有什么问题没有。
对于最后一个高斯分布的例子，需要自己推导一下。这里我没时间，还没有具体完成呵呵呵
GFA 我学过一个学期，直接崩溃，颠覆我对世界的认识，有时间大家可以尝试读读，它会告诉你高维空间的数据分布的一些惊人的例子

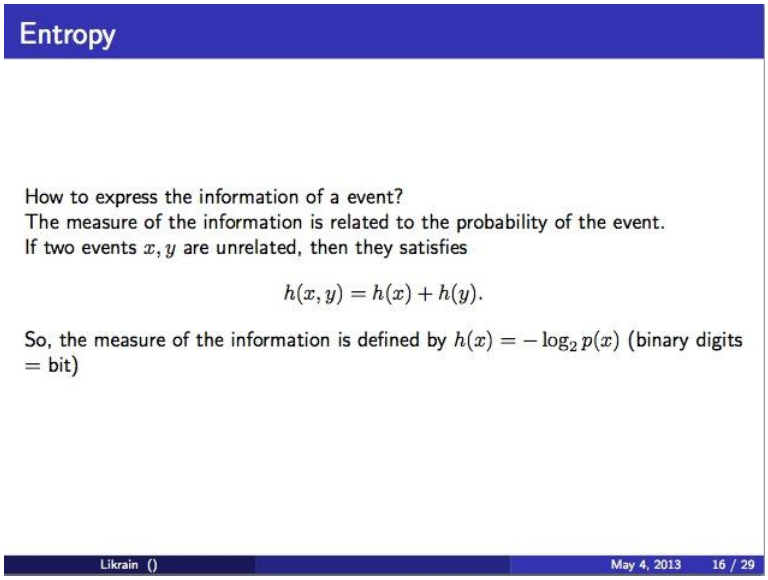
ok，这里暂停一下哈，讨论

planktonli(1027753147) 20:25:15
恩,Functional Geometrical Analysis 是比较难，和直观的现实世界完全不同，尤其 high dimension 的时候。
likrain(261146056) 20:25:55
如果不去读是没法理解的。。。。。所以很多高维空间的数据处理的方式都是从数据本身高维空间的“样子”给出的，当然所谓高维空间说的有些含糊了，严格的要给出各种度量，各种测度等。
planktonli(1027753147) 20:27:39
恩
likrain(261146056) 20:28:04
ok 那我继续哈哈
lost(549294286) 20:28:11
2^n 个模型
模型的个数用变量数 n 衡量吗
likrain(261146056)
是的

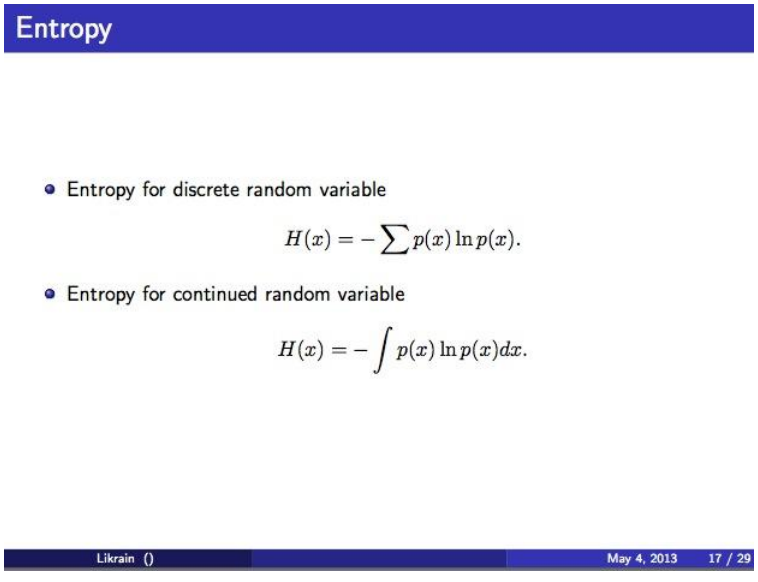



lost(549294286) 20:28:38
那非线性拟合
likrain(261146056) 20:28:53
那就更复杂了，度量函数空间，简单的例子
planktonli(1027753147) 20:29:16
范函
likrain(261146056) 20:29:20

你可以去用所谓的“数数”
lost(549294286) 20:29:28
额。。
likrain(261146056) 20:29:30
去数多项式空间，你怎么去“数数”，三角函数空间
所以 vapnik 在统计学习那本书里面写了那么多看似乱七八糟的定理，各种熵，VC 维才能度量函数空间然后再去数数。
lost(549294286) 20:30:39
还是继续把
likrain(261146056) 20:30:45
哦了




信息论，是机器学习的很多方面的另一种解释，可以用它去解释很多机器学习的模型，所以也可以想想很多人做信息论的就直接使用信息论的方法，重新建模机器学习方法 。
这里可以举个国内的吧，jun zhu，可以看看他的东西。



## Maximum Entropy

- Maximum Entropy for discrete random variable corresponded to $x$ is form the homogeneous distribution.
- Maximum Entropy for continued random variable corresponded to $x$ is form the Gaussian distribution.



对于信息论~~这个领域本身太大了，我就把我认为重要的概念贴上来了。。。。

## How to use the KL divergence.

How to use the KL divergence.

- Minimizing the KL divergence is equivalent to maximizing the likelihood function.

$$KL(p||q) \approx \sum_{n=1}^{N}\{-\ln q(x_n|\theta) + \ln p(x_n)\}$$

- Mutual information

$$I(x,y) = KL(p(x,y)||p(x)p(y)) = -\int\int p(x,y)\ln\{\frac{p(x)p(y)}{p(x,y)}\}dxdy.$$



## KL Divergence

Consider some unknown distribution $p(x)$, and suppose that we have modelled this using an approximating distribution $q(x)$. As this viewpoint we must give a measure of the dissimilarity of the two distribution $p(x)$ and $q(x)$.

- relative entropy or KL divergence

$$KL(p||q) = -\int p(x)\ln\{\frac{q(x)}{p(x)}\}dx.$$

- The property of the KL divergence

$$KL(p||q) \neq KL(q||p)$$

$$KL(p||q) \geq 0$$



好了我还真不知道这些概念需要说什么哈哈

就说这么多这次哈哈，多谢大家耐心听我唠叨

shrake-DM(965229647) 21:43:44

统计与机器学习 ikrain 已经解释的十分全面了 ,只是补充一下 ,最小二乘用的是 squre loss ;svm 是 hinge loss；所以你说前者是统计的，后者在这个意义下也应该是可以划入统计范畴的，而且 alex 及其追随者，把 loss 这里作了很多非常统一的 common sense，2000 年左右无数本书，可以看看，前面 ikrain 都提到了；GFA 有时间可以学下，cmu 有这个相关的课，很有启发，对于 random projection 启发大一些。我忘了很多了，但是高维空间的球的质量分布在球壳上或赤道上（记不清了），这个比较违反我们的直觉。一个统计的应用是高维高斯分布（维数真的要很高），随机产生点，球内是几乎找不到的，只有在球壳（或是赤道）这点忘了，出了球壳记得也是几乎没有点的。