**目录**

[第一章 、介绍 2](#_Toc47861640)

[第二章 、线性代数 3](#_Toc47861641)

[第三章 、概率统计 4](#_Toc47861642)

[第四章 、优化 7](#_Toc47861643)

[第五章 、贝叶斯决策理论 10](#_Toc47861644)

[第六章 、概率密度估计 12](#_Toc47861645)

[第七章 、聚类和评估 15](#_Toc47861646)

[第八章 、回归 17](#_Toc47861647)

[第九章 、分类 19](#_Toc47861648)

[第十章 、线性降维与统计学习理论 21](#_Toc47861649)

[第十一章 、支持向量机 24](#_Toc47861650)

[第十二章 、神经网络 25](#_Toc47861651)

[第十三章 、核回归与高斯过程 26](#_Toc47861652)

# 、介绍

1. **机器学习**
   1. **什么是机器学习**

Develop a machine or an algorithm that learns to perform a task from **past experience**.

* 1. **为什么发展机器学习**
* 这是所有只能/自动化系统的基础
* 发现数据中的规律和pattern
* 实现系统的自适应(automatic adaptation)
* 试图借此理解生物智能（人的大脑怎么运作的）
  1. **机器学习的数学描述**

机器学习是学习一个从输入到输出的映射。换个方式说，试图通过输入来预测输出，其中是需要学习的参数

1. **分类与回归的区别**

分类是映射到离散空间，回归是映射到连续空间

1. **机器学习的类型**

* **Supervised Learning：**带有标签label的数据（输入输出对）。比如图像识别
* **Unsupervised Learning：**没有标签的数据。比如聚类、降维、概率密度估计
* **Semi-Supervised Learning：**数据有/没有标签
* **Learn-by-Doing：**没有输入例子的学习。比如增强学习reinforcement learning

1. **挑战**
   1. **泛化能力Generalization**

* We can’t simply memorize the training set。学习结果还要对没见过的输入（测试集）也有好的输出
* 如何达到好的泛化能力？**Occam’s Razor**：不能让模型太复杂
  1. **特征Feature的选取**
  2. **维度的诅咒**

所以要对输入数据做降维处理

* 1. **性能评价**
* 如何正确评价/描述性能？比如：“语音识别正切率为99%”到底意味着什么？
* 评价时注意对比在训练集和测试集上的表现
  1. **误差度量Error Metric**
  2. **模型选取**
  3. **计算难度**
* **学习：**这通常是个优化问题：通过给定的训练集找到一个最优参数，使得预先定义的误差最小。训练集通常相当大，如何快速求解？
* **预测：**给定一个测试集输入，如何快速计算出prediction

# 、线性代数

1. **方阵的逆**



1. **非方阵的伪逆**

**左伪逆：**若矩阵**列满秩**，则定义其左伪逆

**右伪逆：**若矩阵**行满秩**，则定义其右伪逆

# 、概率统计

1. **离散分布**

|  |  |  |  |
| --- | --- | --- | --- |
| **名称** | **定义** | **期望** | **方差** |
| **Binomial** |  |  |  |
| **Multinoulli** |  |  |  |
| **Multinomial** |  |  |  |
| **Poisson** |  |  |  |

* 多努力、多次项(用于multi-class分类)分布分别是伯努利、二次项分布的推广
* 二次项分布是多次做伯努利随机，求特定次数成功的概率。多次向分布类似
* 泊松分布是二次向分布中，但的极限

1. **高斯分布**

* 原始形式

上式：一元变量x，其中µ是均值， σ2是方差。

下式：多维向量**x，**其中µ是一个D维均值向量， Σ是一个D × D的协方差矩阵， |Σ|是Σ的行列式。

* **为何重要**？因为无穷多的随机变量之和的极限就是高斯分布。中心极限定理（central limit theorem）告诉我们，对于某些温和的情况，一组随机变量之和（当然也是随机变量）的概率分布随着和式中项的数量的增加而逐渐趋向高斯分布
* 对角化



* 分块Partitioned高斯分布



， 

1. **期望、方差、矩**

* **期望**定义式：



当分布未知时，可以用**采样**来**估计**期望：



* **方差**是随机变量之dispersion的度量，描述the expected spread of the variable in relation to its mean:



* **协方差**是correlation的度量，描述how much two variables change together:



* 一个重要公式：
* **矩**：， **中心距**：

1. **指数家族**

* **通式：**



其中是natural parameter，是归一化因子。归一化条件

* 指数家族的成员都是单模**uni-modal**的

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** |  |  |  |  |
| **Bernoulli** | 1 |  |  |  |
| **Multinoulli** | 1 | 1 |  |  |
| **Gaussian** | 1 |  |  |  |

1. **信息论与熵**
   1. **信息论核心问题**

* 如何用最少的bits**表达**信息？
  + Compressing text like with GZIP（一种压缩文件格式）
  + Compressing pictures like in JPEG, movies like in MPEG
  + Compressing sound using MP3
* 如何可靠地**传输储存**信息？
  + ECC memory
  + Error Correction on CDs
  + Communication with space probes
  1. **什么是信息**
* What is the number of bits you need to represent 27 characters? ≈ = 5 bits
* 定义为**单个字母**携带的信息（以英语字母表为例）pi是每个字母的用到的概率
* **熵(entropy)**：平均每个字母携带的信息：
  1. **Kullback-Leibler Divergence**
* 描述两个概率分布的相似程度。Represents the average addition bits needed to specify *x*, given that its underlying probability distribution is the estimated *q*(*x*) and not the true one *p*(*x*):



* 重要**性质**：

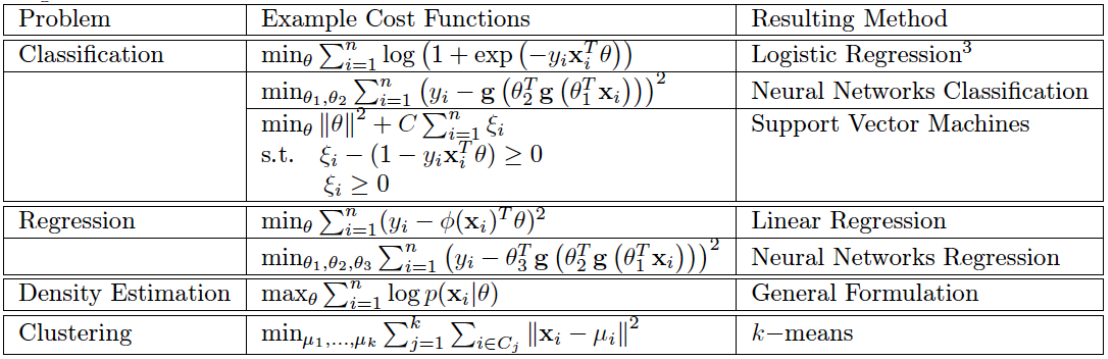


# 、优化

1. **动机**

*所有机器学习问题本质上都是优化问题*

* 1. **优化问题的组成部分**
* 要优化的参数
* 优化利用的数据库
* 评价性能measure our performance 的cost function 
* 约束条件

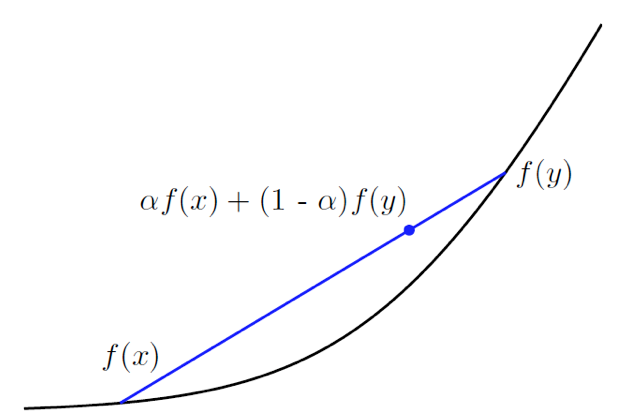


机器学习告诉我们，如何提出基于data的cost function，从而使优化能够解决这些问题!

**Good**机器学习告诉我们，如何提出基于data的cost function，从而使优化能够解决这些问题 **efficiently**!

1. **凸Convexity**
   1. **Convex Sets**

* **凸Sets：**A set C ⊆ Rn is convex if
* **例子：**Affine subspaces (linear manifold): Ax = b; Ay = b, then A (αx + (1 − α) y) = αAx + (1 − α) Ay = b
  1. **Convex functions**
* **凸函数：**不严格来说，弦在函数上面的函数



* **例子：**
  + Linear/affine functions: f (**x**) = **b**T **x** + c
  + Quadratic functions:
  + Log-sum-exp:
* **凸函数的一阶特征：**
  + 等号右边是对函数在x点的一阶近似（在该点的切线）
  + 一阶条件的意义是，对于函数在定义域的任意取值，函数的值都大于或者等于对函数在这点的一阶近似。
  + 这个很难算，通常使用二阶特征来判断一个函数是不是一个凸函数
* **Subgradient：** 。注意没要求函数可导，可以是折线
* **二阶可导函数凸的充分条件：**是半正定阵
* **理想的优化问题：**cost function凸；等式约束仿射/线性；不等式约束为凸集

1. **有约束的优化**
   1. **套路**

* 标准形式的原**始优化问题** 
* 写出对应的**拉格朗日函数** 
* 那么原优化问题的**偶问题** .对求解之（先对求里面，再求外面）
  1. **为什么引入偶问题**
* 偶问题的约束非常简单；
* 偶问题永远是**凸**的

1. **数值优化**
   1. **基本思路**

找一个使得；再迭代更新（p47）

**→→→**如何选择与？

* 1. **选**
* **激进派：**，一条路走到底。**特点：**下一次的方向一定和这一次的方向垂直，最终形成锯齿路径（如何理解：考虑梯度的意义）
* **保守派：**
  1. **选**
* **Axial Iteration (aka coordinate descent)**
* **Steepest descent**
  + , the update rule becomes
  + , the update rule becomes
  + With
* **牛顿法：**首先在current点泰勒展开



然后选取使函数值最小的



注意到只有是变量，是参数。对求梯度，令为零，得到：

最后进行迭代

**特点：**

**a).** 对quadratic的目标函数，只要一次迭代；

**b).** 迭代具有quadratic的收敛性质；

**c).** 若Hessian正定，则目标函数值一定单减；

**d).** 若Hessian是单位矩阵(**H**=**I**)，则退化为steepest descent

**缺点：**

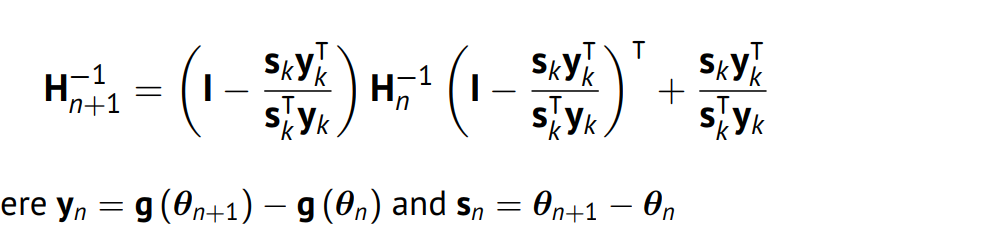
目标函数必须一阶二阶可导

每次迭代都要算Hessian及其逆矩阵，消耗大

目标函数不是完全的凸函数时，容易陷入鞍点，导致更新朝着错误的方向移动

* **准牛顿法BFGS：**注意到Hessian变化慢、对称、是Jacobian的梯度，因此引入下面优化问题来估计之：





* **共轭梯度法Conjugate Gradient：**This method chooses successive descent directions  such that it is guaranteed to reach the minimum in a finite number of steps.

**特点：**所有选区的方向两两由Hessian关联。方向两两线性无关

|  |  |  |  |
| --- | --- | --- | --- |
| **算法** | **单次迭代耗时** | **迭代数** | **陷入停滞** |
| **准牛顿BFGS** | 大 | 少 | 难 |
| **共轭梯度** | 小 | 多 | 易 |

https://zhuanlan.zhihu.com/p/27249004

# 、贝叶斯决策理论

1. **Bayesian决策论**
   1. **一些概念**

* **Class Conditional Probability：**Probability of making an observation *x* knowing that it comes from class *Ck*
* **Class Prior：**The prior probability of a data point belonging to a particular class
  1. **什么是贝叶斯决策论**

Use the previous knowledge to find the posterior probability of the class *Ck* given the observation *x*.



* class **prior**:
* class-conditional probability (**likelihood**):
* class **posterior**后验概率:
* normalization term:
  1. **为什么叫贝叶斯决策论**

这是基于**贝叶斯概率**的决策论：

* 概率不再仅仅解读为事件发生的频率，而是看作一个输出的置信度degree of belief
* 只有这样才能对“输入A对应的输出为B”这个assert一个先验prior belief
* 目标是减少错误分类的概率
  1. **如何决策**
* Decide if
* Decide for class k if

1. **Risk Minimization**
   1. **思路**

* 引入一个loss function ，描述将属于的归类为时受到的损失
* 将样本归类为时的expected loss (Overall Risk)

。

这也叫risk of making a decision of .

* P28例子：class有两种，对应decision也会有两种。来了一个样本，应该决策成哪个呢，这会儿就要计算26页的损失R，哪个损失小就决策成哪个。损失是用来衡量分类或者回归的正确性的，预测越靠近真实值，损失越小，相反，预测越偏离真实值，就需要惩罚，loss就越大，我们的目的是最小化这个Loss。
  1. **推导**
* **目的：**最小化overall risk。若，则做决定
* **二元分类：**

做决定，如果满足：

If ,

* **Class Prior：**The prior probability of a data point belonging to a particular class
  1. **特殊情况**
* 如果loss function选为0-1 loss，那么risk minimization退化为最小化错误分类率

1. **Bayesian Optimal Classifier**

永远选择使得最大的那个的classifier

1. **高维的处理**

整个的feature space。用mutivariate class-conditional density 找decision boundary。

# 、概率密度估计

1. **概率密度**

* Bayes optimal classification, 基于, 易求，我们需要估计（learn）

1. **有参模型——极大似然**
   1. **有参模型的概念**

有参密度模型写出是。“学习”指基于训练集估计的参数出来。

例子：高斯分布

}

* 1. **有参模型的极大似然**
* **Likelihood of** **：**数据集是由参数为的概率密度函数生成的概率，记作
* **IID假设：**

**i.i.d:** independent and identically distributed(独立且均匀分布)

独立：

均匀：

使得似然函数变成连乘形式

* Log-Likelihood
* **Gauss的ML：**

。**要能自己推！**

* 1. **极大似然的退化形式**

当*N=*1，即训练集只有一个元素时，极大似然得到的结果是delta函数。**如何避免**？引入prior，用贝叶斯估计

* 1. **Bayesian Estimation**
* **点子：**Assumes that the parameters are not fixed, but are random variables too.
* **描述：**Formalize this as a conditional probability

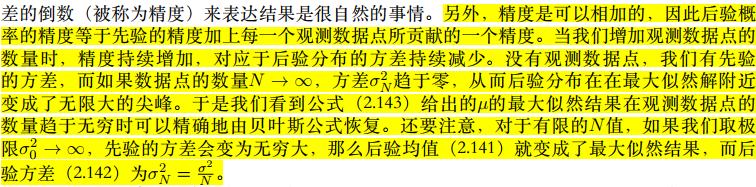


* **Bayes Point：**如果 只对某个特别大，则称作Bayes Point，且成立估计式
* ? What happens, when N grows to infinity

Bayesian estimation:

,

理解：

* 是后验概率，正比于
* 是先验均值和先验方差，是似然函数中的参数
* 是的最大似然解，由样本均值给出，
* 方差的倒数是精度，精度可相加。
* 当N=0，这个公式就变成了先验均值。
* 当N🡪, 后验均值由最大似然解给出。
* 
  1. **Conjugate Priors**

Conjugate Priors are prior distributions for the parameters that don’t “change” the type of the parametric model.

1. **无参模型**

*通常无法预知是哪种模型*

* 1. **Histograms**直方图
* **思路：**将feature space离散成bin，bin宽度为
* **性质：**(1) 数据集无穷大且无穷小时无限精确；(2)这是个暴力法brute-force
* **问题：**(1) 高维诅咒：bin数量几何增长、需要的数据量也几何增长；(2)、如何选择
  1. **一种一般考虑**
* 数据点*x*落在区域*R*内的概率：
* 若*R*足够小*R=V*，有，提出积分：
* 若*R*足够大，点落在*R*内的概率就是训练集中点落在*R*的频率：
  1. **Kernel density estimation ——定V求K——K怎么求？**
* **Parzen Window：**， ，
* **Gauss：**， ，
* **一般核：**，其中，，
* **Kernel的问题：**要选择核函数和bandwidth *h*
  1. **K-Nearest Neighbors (KNN)——定K求V**

**KNN用于分类：**，用领域内的类频率来表示该点类概率

1. **混合模型**

|  |  |  |  |
| --- | --- | --- | --- |
| **有参模型** | 好的解析性 | 小存储 | 快 |
| **无参模型** | 通常不解析 | 大存储 | 慢 |

* 1. **混合Gauss及其ML**
* ，其中。j是混合模型中第j个分量，一共有J个高斯模型混合在一起，是混和系数，Mixture parameters是
* **ML结果：** 怎么算的？？
* **ML问题：**Circular dependence，没有解析解，全是耦合的
* **Gradient Ascent：**不可行！梯度算出来很复杂(nonlinear, circular dependencies)
  1. **Latent Variable与ML（以混合高斯为例）**
* **隐藏变量：**ML的结果中，是观察不到的，叫latent variable
* **假设已知**：则可以算出的maximal likelihood solution，即得到distribution
* **假设已知分布：**通过贝叶斯决策论边际概率等方法得到latent variable
* **然而：**两个现在都不晓得
  1. **Expectation Maximization (EM)法（以混合高斯为例）**

*Method of performing maximum likelihood estimation, even we only have access to observed variables*

* **算法：**
  + **初始化参数:** ,, … ,
  + **E-step：**用当前参数值估计latent variables
    - 计算每个混合成分和所有数据点的后验分布
    - 算
  + **M-step**：按照ML的结果，用latent variables算参数
    - Compute the new means as the weighted means of all data points
    - 算
* **性质：**
  + Incomplete (observed) data:
  + Hidden (unobserved) data:
  + Complete data:
  + 联合密度: with parameters
  + : 混合模型的似然
  + : 混合成分的预测
  + Incomplete likelihood:
  + Complete likelihood:

# 、聚类和评估

1. **聚类**
   1. **混合模型在聚类中的应用**

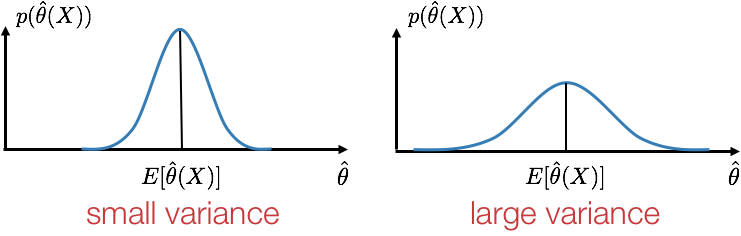
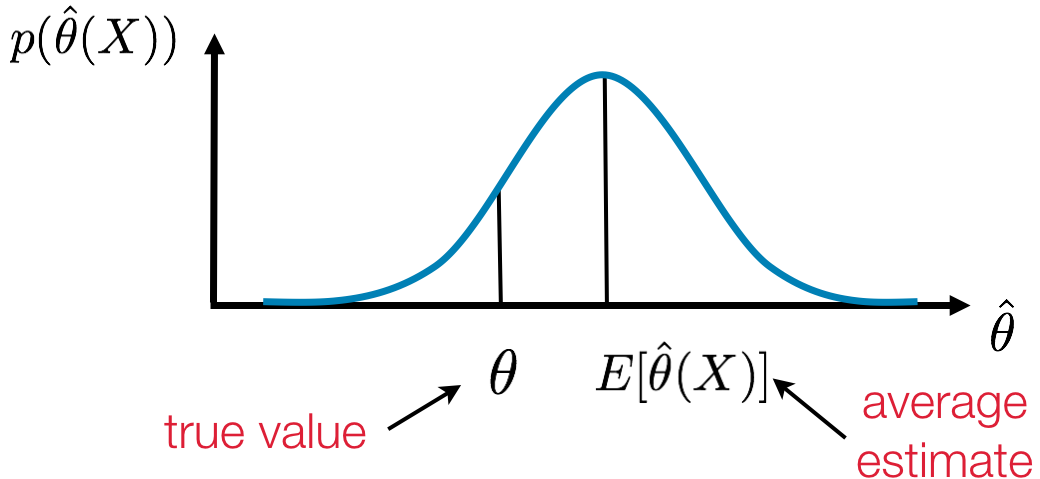
(1)、Divide the feature space into meaningful groups. (2)、Find the group assignment.

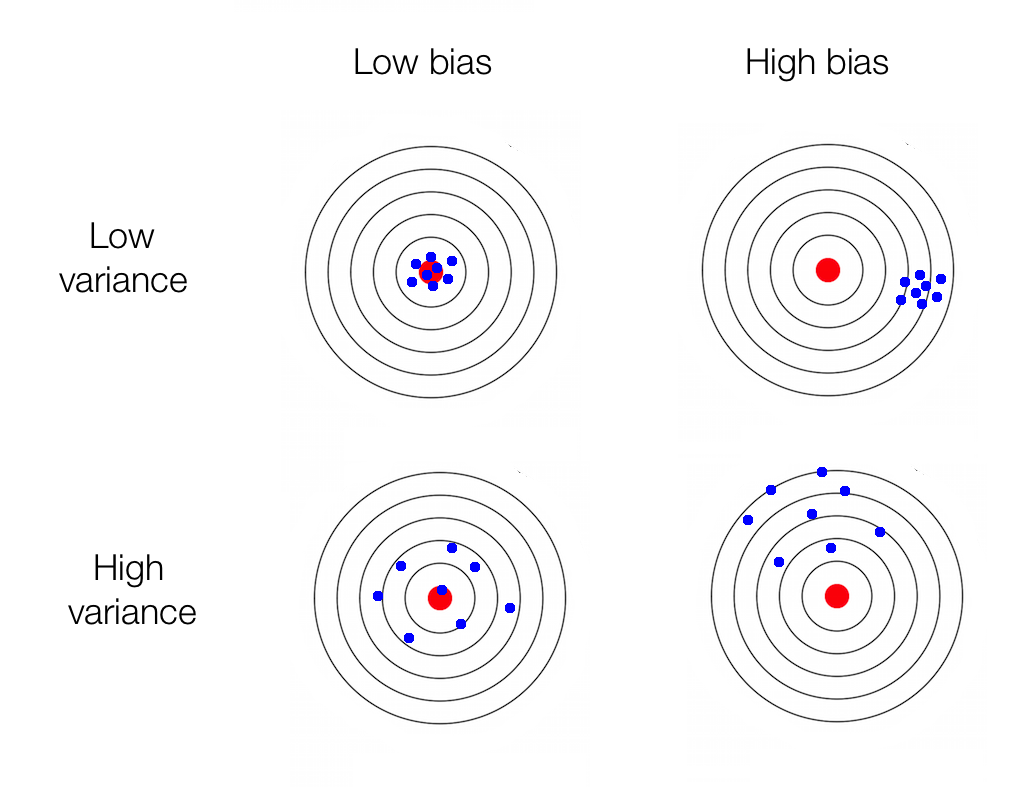
* 1. **聚类算法**
* **Agglomerative：**先将每个点单独整成一个聚类，再将有最小inter-class distance的两个融合成一个
* **Divisive：**先将所有点整成一个，再将某个聚类分成两个有最大inter-class distance的聚类
* **Mean Shift：**Mean shift is a method for finding modes in a cloud of data points where the points are most dense.利用Kernel密度估计和local search算法，从不同的数据点出发，不断地在领域内寻找目的最大的位置。收敛到同一局部最大的出发点归为同一类。

1. **评估**

***Occams’ Razor:*** *Always choose the simplest model that fits the data*

* 1. **估计器的Bias和Variance**
* 是基于数据集对参数的估计器。其中是a random sample from the same underlying distribution
* **Bias：** expected deviation from the true parameter
* **Variance：** expected squared error between the estimator and mean estimator





* 1. **MVUE和BLUE**
* **MVUE：**Minimum Variance Unbiased Estimator。无偏，小方差
* **BLUE：**Best Linear Unbiased Estimator。linear in the features的MVUE
* The MLE of the mean of a Gaussian is UNBIASED
* The MLE of the variance of a Gaussian is BIASED
  1. **Bias-Variance-Tradeoff（以回归为例）**
* 设数据点由这个式子产生，其中为噪音。设估计器估计
* **Expected Squared Error：**estimated from all possible data set：



* **Bias：**bias2{} structure error. 对模型 是固定的
* **Variance：** estimation error，可以通过增大样本而下降,有限的数据集一定会有error
* **Expected Total Error Bias2 + Variance**

1. **模型选择**

*按道理要在测试集上evaluate模型，但是没有测试集时怎么办？*

* Split the dataset into:
* **Training Set**: Fit Parameters
* **Validation Set**: Choose model class or single parameters
* **Test Set**: Estimate prediction error of trained model
* **Exhaustive Cross Validation：**尝试所有可能的partitioning，计算量大
* **Bootstrap：**Randomly sample non-overlapping training/validation sets。对数据集再采样
* **K-Fold：**计算量最小。分成*K*个，*K*-1个来训练，1个来validate, 循环
* **Leave-One-Out：***K*=*N*-1，validation集的容量为1，属于Exhaustive Cross Validation

# 、回归

1. **最小二乘回归**

* **Linear Regressor：**。引入化成
* **最小二乘：**。求梯度，令为零，结果为
* **问题：**算逆的时间消耗很大 → 梯度下降、降维
* **多项式回归：**整个feature transformation，然后套用最小二乘

1. **极大似然回归**
   1. **引入概率的极大似然回归**

* **假设：**(a)、样本带噪音。(b)、噪音为正态分布。

(c)、因此输出是个概率分布

* **Conditional Likelihood：** （设用多项式）



* **求梯度：**
* **结果：** ，和最小二乘法的结果一样！
* 带高斯噪音的极大似然结果虽然和最小二乘的结果一样，但是ML比最小二乘更为强大，因为ML可以算，能够测量uncertainty of estimation
  1. **Loss Function在回归中**
* **动机：**上面的ML回归，最后得到的是输出的一个概率分布。如何由这个概率分布得到一个具体的值？
* **Expected Loss：**，其中*y*是真实输出，*f*(*y*)是回归输出
* **标准：**选择最小化全局Expected Loss的作为输出

**一个简单case，Squared Loss。**

* ，求导，结果
* 因此，squared loss也叫mean prediction

1. **贝叶斯回归**

*ML回归总是倾向于过拟合*

* 1. **一般思路**
* 给参数*w*引入一个Gauss的先验, 其中Zero mean and spherical covariance球面协方差矩阵，也可以给别的值
* 那么后验
  1. **Maximum A-Posteriori MAP**
* **Log后验：**



相当于，控制数据相关的误差（第一项）和正则化项（第二项）的相对重要性。

* **求梯度：**
* **结果：**
* **先验的角色：**has the effect that it regularizes the pseudo-inverse可以规范伪逆，这个也叫作ridge regression
* **与最小二乘比：**

MAP和带regularizer最小二乘进行比较，

和结果一样。

因此当在最小二乘中添加regularizer时，暗中假定输出和参数都高斯分布。

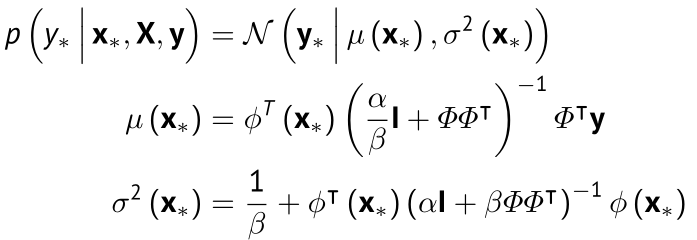
Regularizer 控制模型的复杂度complexity of the model 和过拟合程度degree of overfitting

* 1. **Full Bayesian Regression**
* **动机：**在这之前都是在估计参数，但我们本质上并不关心参数, all we want to do is to predict a function value based on the training data
* **求边缘概率密度**，消灭参数*w*：(边缘化marfinalizing方法是对其求积分)



predicted value; test input; X: training data points; y: training function values

* 这样整完得到的predictive distribution仍是一个高斯分布：

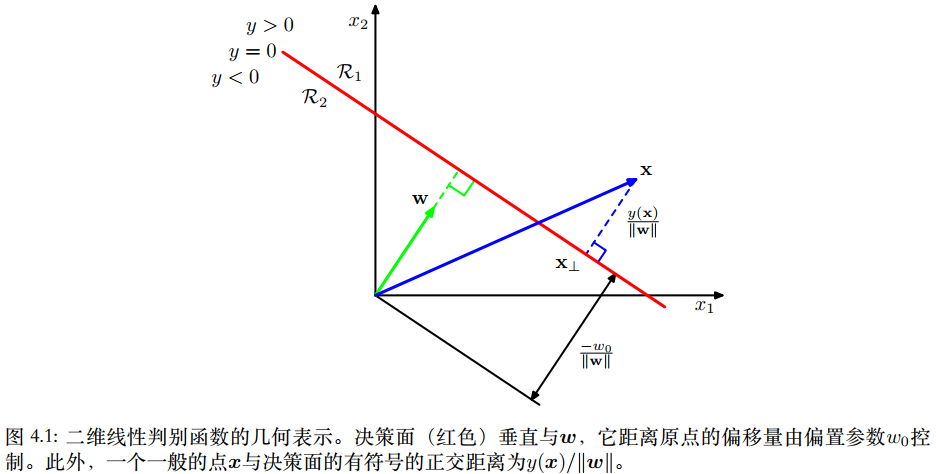


: predicted value; : test input; X, y: training data

注意方差和无关，称作state independent

# 、分类

1. **判别函数Discriminant Function**
   1. **新旧路线**

* **旧路线：**先整class-conditional density和class prior，再贝叶斯算class-posterior，并max之以min error。谁大选谁。
* **新路线：**直接求解decision boundary，放弃模拟density。最后仍然是min错误率
  1. **判别函数Discriminant Function**
* **思路：**判别函数。若，则判为*k*类
* **例子：**来由贝叶斯分类衍生出的判别函数：
* **两类时：**两个判别函数可以退化成一个2-class classifier：
* **线性判别：**，其中是决策边界法向量，是offset，决定原点到**边界（超平面）**的距离
* 
* **Linearly Seperable：**能用线性判别函数完全分开
  1. **多类情况**
* **问题：**如果用若干个2-class classifier，那么不管是one-versus-rest还是one-versus-one都会产生模糊区域
  + one-versus-rest 区分是否属于，数量为K-1
  + one-versus-one 为每一对类别都设置一个判别函数，数量为
* **解决方案：**有几个类就整几个判别函数。若，则判为*k*类。决策面为，D-1维平面。If the discriminant functions are linear, the decision regions are connected and convex

1. **Fisher判别分析**（为了简单，只整两类）
   1. **先看个失败方法：最小二乘法**

* **引入：**目标是得到这样一个判别函数。输入，输出
* **线性判别方程：**
* **转化：**。这个方程组是overdetermined的，一般样本数远大于类数, n equations and d+1 unknows
* **最小二乘：**
* **问题1：**当样本存在outlier时，最小二乘非常不好
* **问题2：**事实上最小二乘处理分类天生不足。因为最小二乘对应Gauss分布的极大似然，而分类是0-1分布
  1. **Fisher判别分析**
* **思路：降维**Find a linear projection of the data and classify the projected values

1. **最大化投影后各类mean的距离**

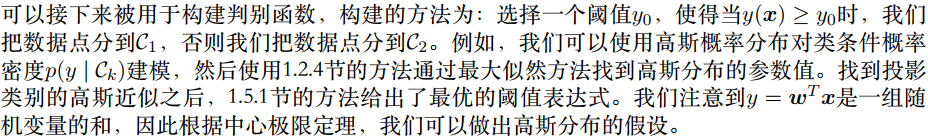
* **投影前的mean：** **投影后的mean：**
* **数学表达：** **结果：**
* **问题：**得到的投影方向不见得就是最好的方向

1. **在尽可能最大化投影后各类mean之间的距离的同时最小化各类自己的variance**

* **定义within-class variance：**
* **目标（Fisher criterion）：**
* **化简分子：**
* **化简分母：**
* **求解：**求梯度，令为零：。由于括号内两个都是标量：。

又 ，所以

我们只关心的方向，不关心它的大小，这个式子也只给了我们方向



**贝叶斯最优：**若class-conditional分布相同，且variance对角，则Fisher判别为贝叶斯最优

**问题：**这种方法对噪音很敏感

1. **Perceptron算法**

*对于线性可分的问题，如何线性分别之？*

* **Perceptron Discriminative Function：**
* **算法：**(a)、初始化一个***w***, *b*。(b)、对于每个点，若感知器对其分类不正确，若本来是1(0)分成0(1)，则



* **线性不可分？：**Transform the input data nonlinearly so that the problem becomes linearly separable

1. **Logistic回归**
   1. **分类问题两种求解思路**

* **Generative：**先class-conditional density和class prior，再class-posterior 。如naive Bayes
* **Discriminative：**直接上class-posterior。如Logistic Regression
  1. **基于概率的判别模型Probabilistic Discriminative Model**
* **后验：**，其中
* **设*a*由线性判别函数给出：**。这在类条件密度高斯且equal covariance适用
* **更新：极大似然：**设iid且。有似然函数



# 、线性降维与统计学习理论

1. **线性降维**
   1. **数学描述**

对于原始数据，找一个线性映射将其映射到低维的空间上，其中可看作是数据在低维空间的坐标。 选取高维正交基，可将写成，选取前面*D*个成分，忽略其他的，就相当于完成了投影

* **基的选取：**选取使**Mean Squared Error最小**：
* **若投影后就单基，有**
* **零mean数据：**最小化MSE和最大化投影后方差是等价的
  1. **主成分分析PCA**
* **背后的直觉：**对于高斯分布，沿着主轴投影，投影后的variance最大。且沿主轴没有协方差
* **目标：**找到principal directions和沿着各主方向的variance
* **投影后的方差：**
* **最大化方差的结果：**利用拉格朗日乘子法，得到
* **注意：前处理：**做PCA之前通常要normalize the mean and variance for each dimensions
  1. **选取D**
* D选大了，误差小，但是没降到维；选小了，误差大
* **选取方法：**(1)、选取满足应用要求的最小D。(2)、选取捕捉到一定variance的那么多D
  1. **PCA总结**
* PCA将高维数据变为低维数据，同时capturing the essence of the data
* PCA finds a more natural coordinate system for the data

1. **统计学习理论Statistical Learning Theory**
   1. **至今干的事**

* 假定模型正确，然后用训练集对模型学习一组最有的参数
* **然而：**(1)、对数据选取的feature可能不合适。(2)、复杂的模型会带来过拟合
* **泛化能力：**训练出来的模型对新的数据集的表现的能力
  1. **统计学习理论**
* 如何控制generalization ability
* 怎样正式地表述泛化能力
  1. **以有监督学习为例讲解统计学习理论**
* **Environment**是稳定的，即数据点是由固定的概率密度产生。**Supervisor**对每个数据点返回个可能带有噪音的标签。**Learning Machine**通过参数估计个标签
* **选取：**从众多的学习到的模型中选取一个出来，使得desired output be approximated optimally（用**risk**判定）

|  |  |  |  |
| --- | --- | --- | --- |
|  | 训练集的loss的算术平均 | 学习算法一般min的是他 |  |
|  | 所有可能按照概率的加权平均 | 泛化误差的期望值 | 然而不知道 |

* 1. **Risk的收敛性**
* **定义：**，
* **假设：**数据足够多时，经验风险收敛于实际风险：
* **假设：**收敛是uniform的是稳定的：
* **一个结论：**有上面两个假设可证，若，有
* **收敛定理：**uniform收敛  当数据无穷多时min经验风险的同时实际风险也得到了min
  1. **Risk Bound**

上面收敛定理虽得到了a formal criterion as to what we can expect in terms of generalization，但是样本并非无穷多

* **思路：**通过经验风险得到实际风险的一个上界。

**其中：***p*\*: probability that the bound is met. *h*: VC维度

* 1. **VC维度**
* **非正式定义：**The VC维度of a family of functions is the 最大数量data points that can be 正确分类by 族类一函数，无论label configuration
* **又名：**a measure for capacity, learning power
* *N*维空间中，线性分类器的VC维度为*n*+1
* **与参数的关系：**一般直接与参数数量相关。反例：的VC维度无穷大
  1. **Structural Risk Minimization**
* **Given** a family of n models with non-decreasing VC-dimension
* 对每个model最小化经验风险
* 选择min风险边界的model。通常这不是min经验风险的model
* **然而：**这个结果只有当bound tight时才sensible

# 、支持向量机

# 、神经网络

1. **学习的表达、及shift到神经网络**

* **学习的表达：**至今，“学习”指的是选取好的feature来解决问题。但是如何选择好的feature？并且对数据的representation也会影响学习的方法及效果（如加法问题，进制不同，学习也不同）
* **当前趋势：**数据变多；函数变非线不凸；feature难提取；参数变多；optimization becomes easier by being deep
* **神经网络：**可以learn complex data representations by combination of simpler ones
* **神经元的生物抽象：**，其中activation function；weight vector；bias/offset/threshold
* **神经网络的生物抽象：**大脑神经网络are determined by the sheets of tissue。其中sheet = vector of neurons

1. **单层神经网络**（以多类分类为例）

* **输出层：**有几类就整几个输出节点。假定模型正确，然后用训练集对模型学习一组最有的参数

# 、核回归与高斯过程

# 、易忘知识点