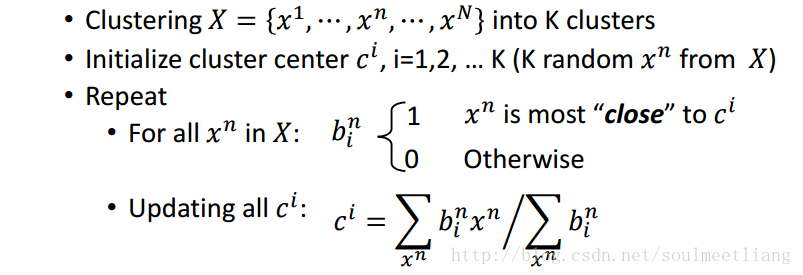
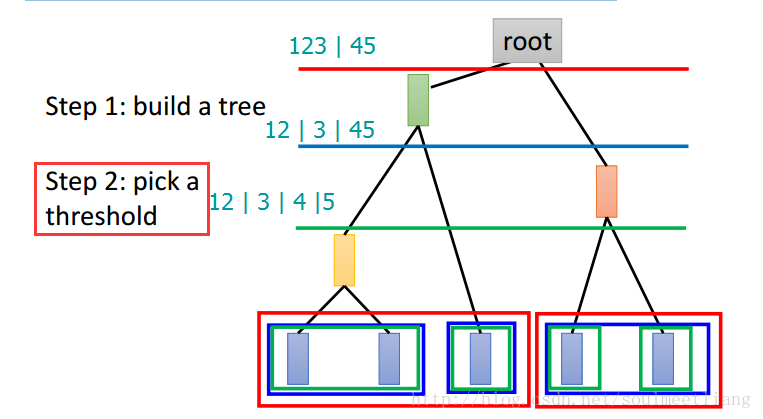
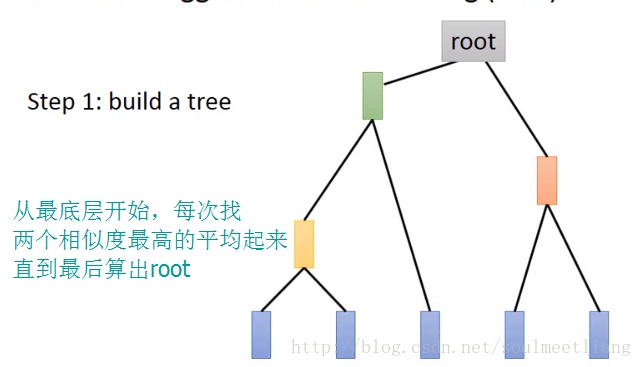
## Clustering

1. **K-Means（K均值）**



1. **Hierarchical Agglomerative Clustering (HAC阶层式汇聚分群法)**



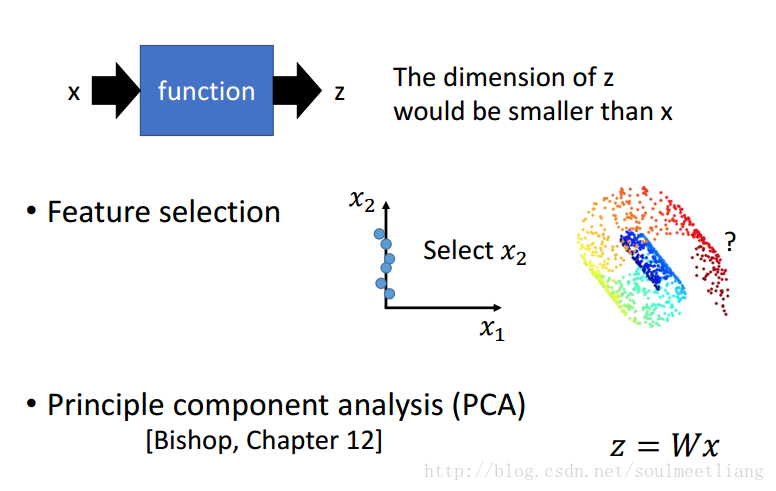
当不知道要分多少类的时候，用HAC会自动确定能分多少种cluster。

## Distributed Representation（分布式表征）

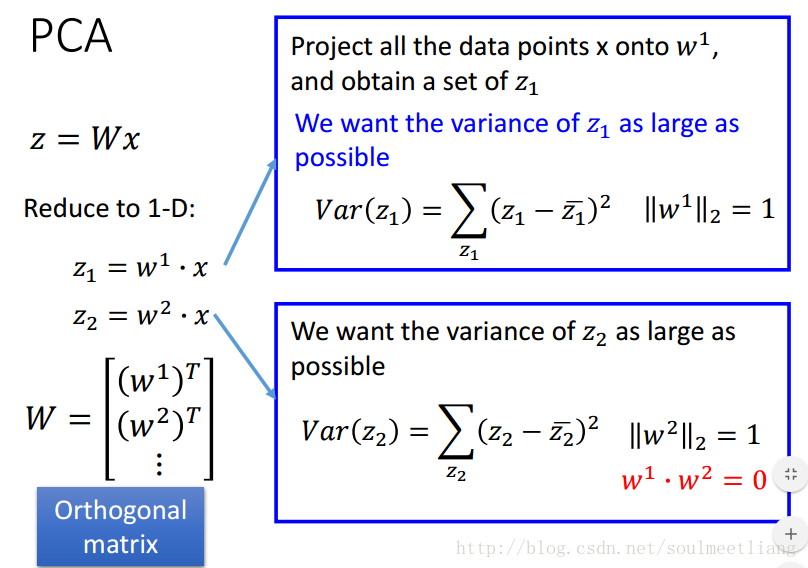
“光做clustering是很卡的，有的个体并不只属于一个大类，所以需要一个vector来表示在各个类中的概率。这样，从一个（高维）图片到一个各属性概率（低维）就是一个Dimension Reduction。”

**就是将一个object用一个vector来表示在各个类中的概率（特征）**

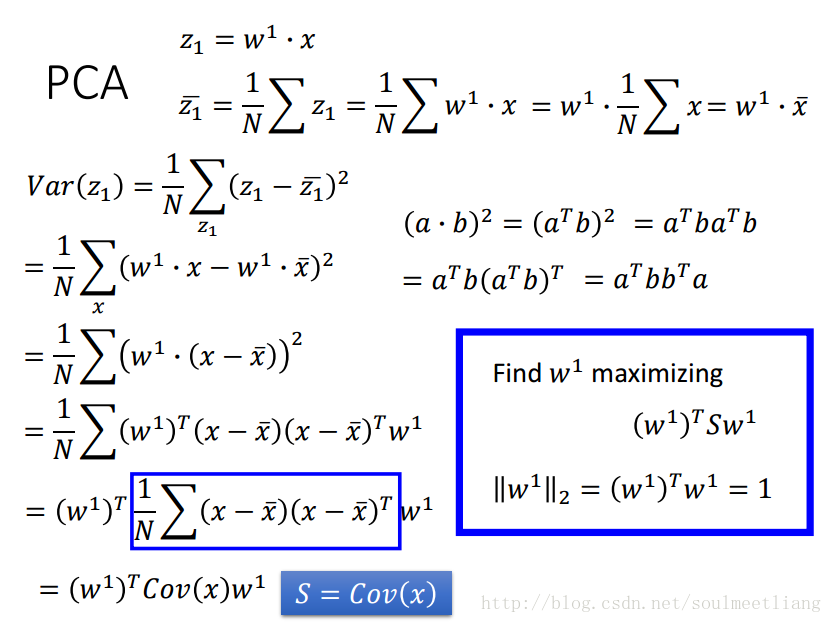
### Dimension Reduction



#### Principle Component Analysis (PCA) 主成分分析



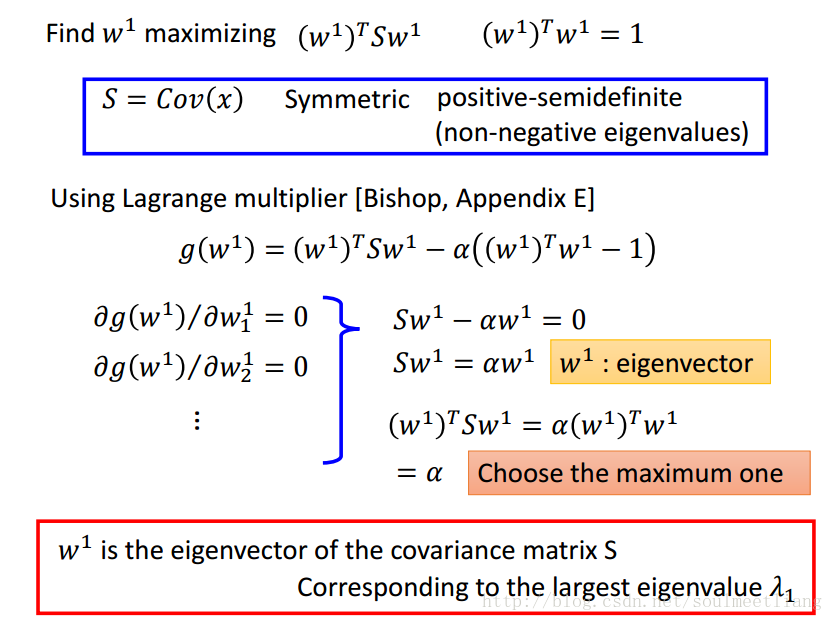
从一维的投影再到二维，2->3…线代中求标准正交基时的步骤



为了尽量减少信息在降维过程中的损失，需要使协方差最大化，协方差越大，说明散布越开，信息量更大。

如何找到使方差最大的w1？

线性代数复习



结论：

g（w1）对w1做偏微分，w1是矩阵

协方差知识详解，点击链接

**流形学习（manifold learning）**

http://blog.csdn.net/chl033/article/details/6107042

假设数据是均匀采样于一个高维欧氏空间中的低维流形，流形学习就是从高维采样数据中恢复低维流形结构，即找到高维空间中的低维流形，并求出相应的嵌入映射，以实现维数约简或者数据可视化。它是从观测到的现象中去寻找事物的本质，找到产生数据的内在规律。流形学习方法是模式识别中的基本方法，分为线性流形学习算法和非线性流形学习算法，线性方法就是传统的方法如主成分分析（PCA）和线性判别分析（LDA），非线行流形学习算法包括等距映射（Isomap），拉普拉斯特征映射（LE）等

1. 流形学习的基本概念   
  
所谓流形（manifold）就是一般的几何对象的总称。流形包括各种维数的曲线曲面等。和一般的降维分析一样，流形学习把一组在高维空间中的数据在低维空间中重新表示。和以往方法不同的是，在流形学习中有一个假设，就是所处理的数据采样于一个潜在的流形上，或是说对于这组数据存在一个潜在的流形。对于不同的方法，对于流形性质的要求各不相同，这也就产生了在流形假设下的各种不同性质的假设，比如在Laplacian Eigenmaps中要假设这个流形是紧致黎曼流形等。对于描述流形上的点，我们要用坐标，而流形上本身是没有坐标的，所以为了表示流形上的点，必须把流形放入外围空间（ambient space）中，那末流形上的点就可以用外围空间的坐标来表示。比如R^3中的球面是个2维的曲面，因为球面上只有两个自由度，但是球面上的点一般是用外围R^3空间中的坐标表示的，所以我们看到的R^3中球面上的点有3个数来表示的。当然球面还有柱坐标球坐标等表示。对于R^3中的球面来说，那末流形学习可以粗略的概括为给出R^3中的表示，在保持球面上点某些几何性质的条件下，找出找到一组对应的内蕴坐标（intrinsic coordinate）表示，显然这个表示应该是两维的，因为球面的维数是两维的。这个过程也叫参数化（parameterization）。直观上来说，就是把这个球面尽量好的展开在通过原点的平面上。在PAMI中，这样的低维表示也叫内蕴特征（intrinsic feature）。一般外围空间的维数也叫观察维数，其表示也叫自然坐标（外围空间是欧式空间）表示,在统计中一般叫observation。   
  
了解了流形学习的这个基础，那末流形学习中的一些是非也就很自然了，这个下面穿插来说。由此，如果你想学好流形学习里的方法，你至少要了解一些微分流形和黎曼几何的基本知识。

 LLE (Locally linear Embedding)