管窥机器学习

七月算法　邹博

2015年5月7日

机器学习

- 在具体学习机器学习的过程中，往往是因为推导造成的障碍
  - 了解基本的高等数学知识是必要的
- 机器学习比想象中要简单的多
  - 举例：kNN用于分类、基本的聚类过程
- 我无意中捡到一个草绳，拽啊拽，竟然在末端发现有头健牛。可悲的是，这头牛，我现在很难驾驭。而隐约看到在牛的后面，还跟着各种动物。

若干概念

- 交叉验证
- 泛化能力
- 监督学习
- 无监督学习
- 强化学习

机器学习算法的分类

- 监督
  - K近邻
  - 回归
  - SVM
  - 决策树
  - 朴素贝叶斯
  - BP神经网络
- 非监督
  - 聚类
  - Apriori
  - FP-growth

交叉验证

- 交叉验证(Cross-validation)也称为交叉比对，主要用于建模应用中。在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，称为PRESS(predicted Error Sum of Squares)。
- 交叉验证是常用的精度测试方法，其目的是为了得到可靠稳定的模型。例如10折交叉验证(10-fold cross validation)，将数据集分成十份，轮流将其中9份做训练1份做测试，10次的结果的均值作为对算法精度的估计，一般还需要进行多次10折交叉验证求均值，例如：10次10折交叉验证，以求更精确一点。

交叉验证的形式

- Holdout 验证
  - 通常来说，Holdout 验证并非一种交叉验证，因为数据并没有交叉使用。 随机从最初的样本中选出部分，形成交叉验证数据，而剩余的就当做训练数据。 一般来说，少于原本样本三分之一的数据被选做验证数据。
- K-fold cross-validation
  - K折交叉验证，初始采样分割成K个子样本，一个单独的子样本被保留作为验证模型的数据，其他K-1个样本用来训练。交叉验证重复K次，每个子样本验证一次，平均K次的结果或者使用其它结合方式，最终得到一个单一估测。这个方法的优势在于，同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次，10折交叉验证是最常用的。
- 留一验证
  - 意指只使用原本样本中的一项来当做验证资料， 而剩余的则留下来当做训练资料。 这个步骤一直持续到每个样本都被当做一次验证资料。 事实上，这等同于 K-fold 交叉验证是一样的，其中K为原本样本个数。

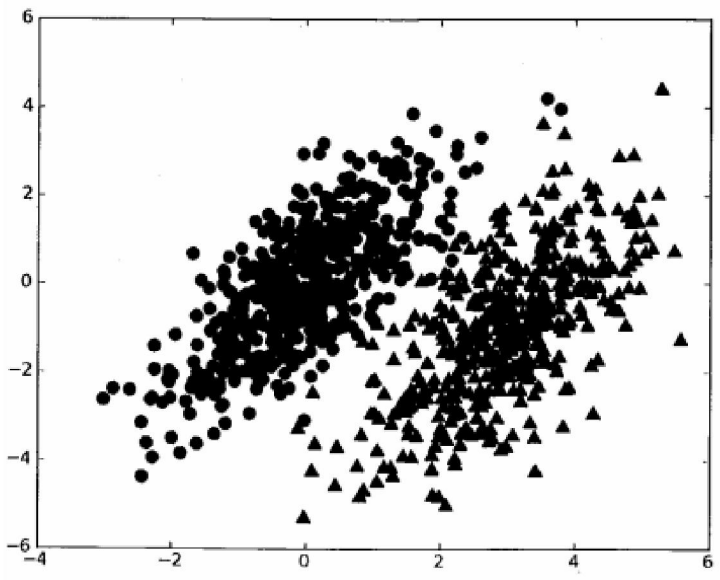
泛化能力

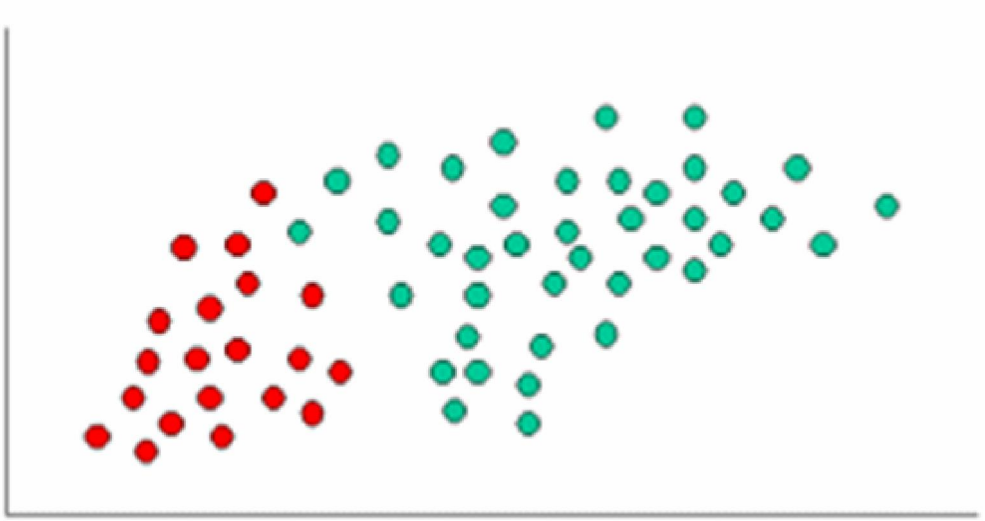
- 概括地说，所谓泛化能力（generalization ability）是指机器学习算法对新鲜样本的适应能力。学习的目的是学到隐含在数据对背后的规律，对具有同一规律的学习集以外的数据，经过训练的算法也能给出合适的输出，该能力称为泛化能力。
- 通常期望经训练样本训练的算法具有较强的泛化能力，也就是对新输入给出合理响应的能力。应当指出并非训练的次数越多越能得到正确的输入输出映射关系。算法的性能主要用它的泛化能力来衡量。

从下面几个问题入手机器学习

- k近邻
- 向量距离
- 聚类
- 线性回归
- 朴素贝叶斯

k近邻分类(属于有监督学习)





向量间相似度计算的方法

- 欧式距离
- Pearson相关系数(Pearson correlation)
- 余弦相似度(cosine similarity)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

k-均值聚类(属于无监督学习)

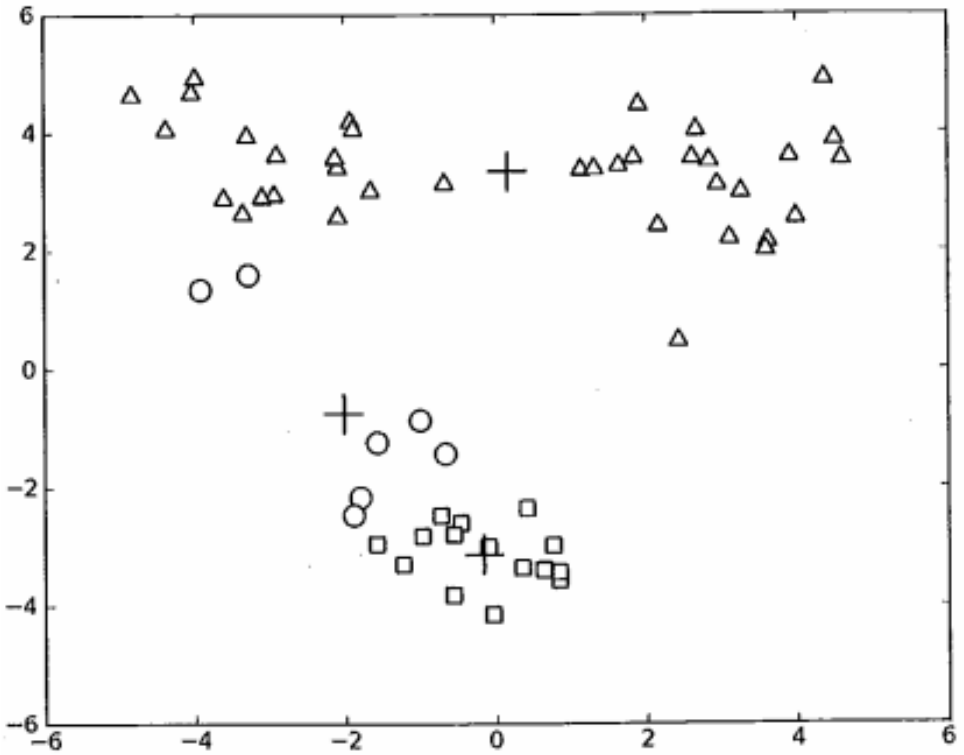
- 创建k个点作为起始质心(如：随机选择起始质心)
- 当任意一个点的簇分配结果发生改变时
  - 对数据集中的每个数据点
    - 对每个质心
      - 计算质心与数据点之间的距离
    - 将数据点分配到距其最近的簇
  - 对每个簇，计算簇中所有点的均值并作为质心

对K-Means的思考

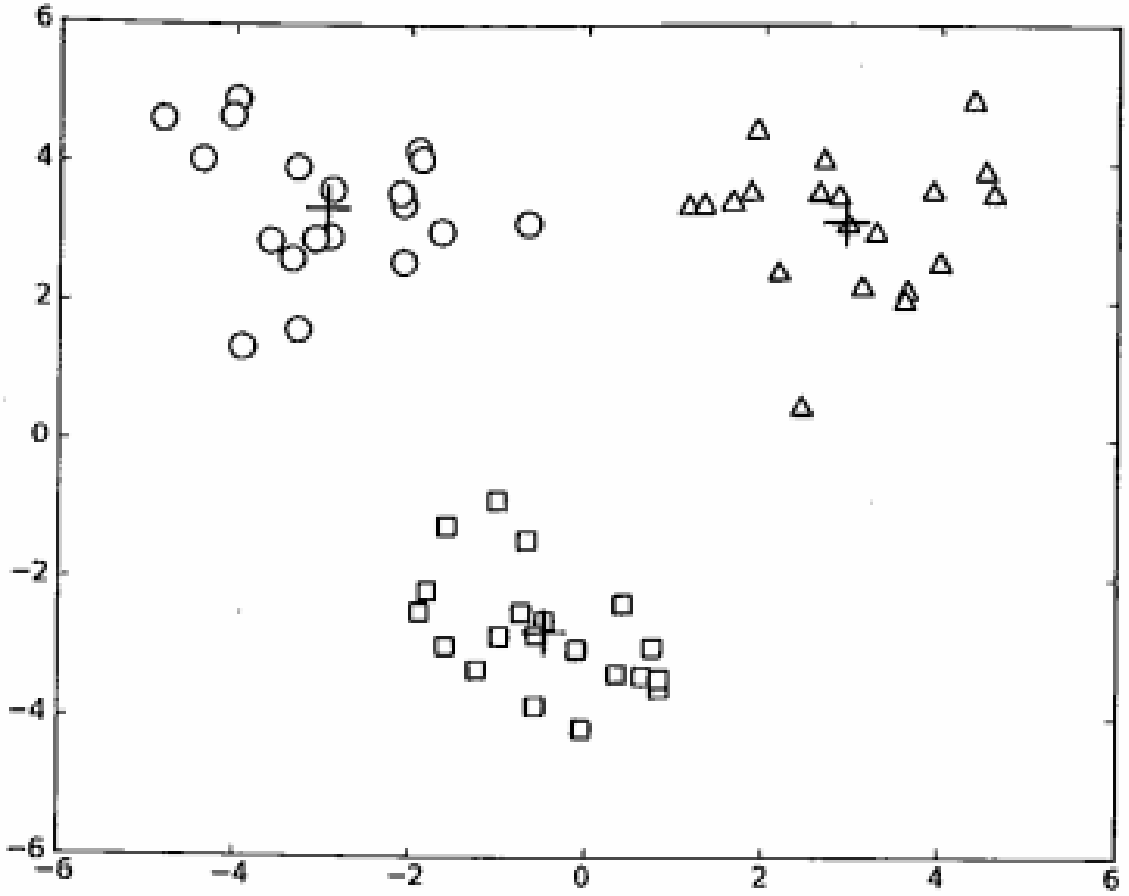
- K-Means将簇中所有点的均值作为新质心，若簇中含有异常点，将导致均值偏离严重。以一维数据为例：
  - 数组1、2、3、4、100的均值为22，显然距离“大多数”数据1、2、3、4比较远
  - 改成求数组的中位数3，在该实例中更为稳妥。
  - 这种聚类方式即K-Mediods聚类
- 点的簇分配结果发生改变的标准如何判断？
  - 实践中可以选择误差的平方和最小
- 初值的选择，对聚类结果有影响吗？
  - 如何避免？

利用SSE进行聚类后处理

- SSE: Sum of Squared Error 误差平方和

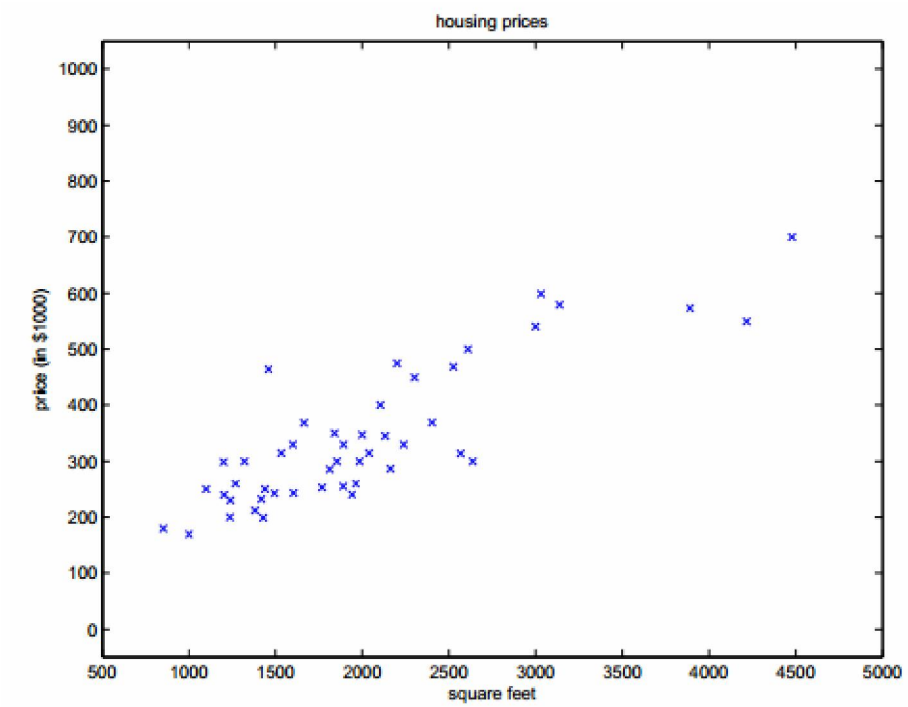


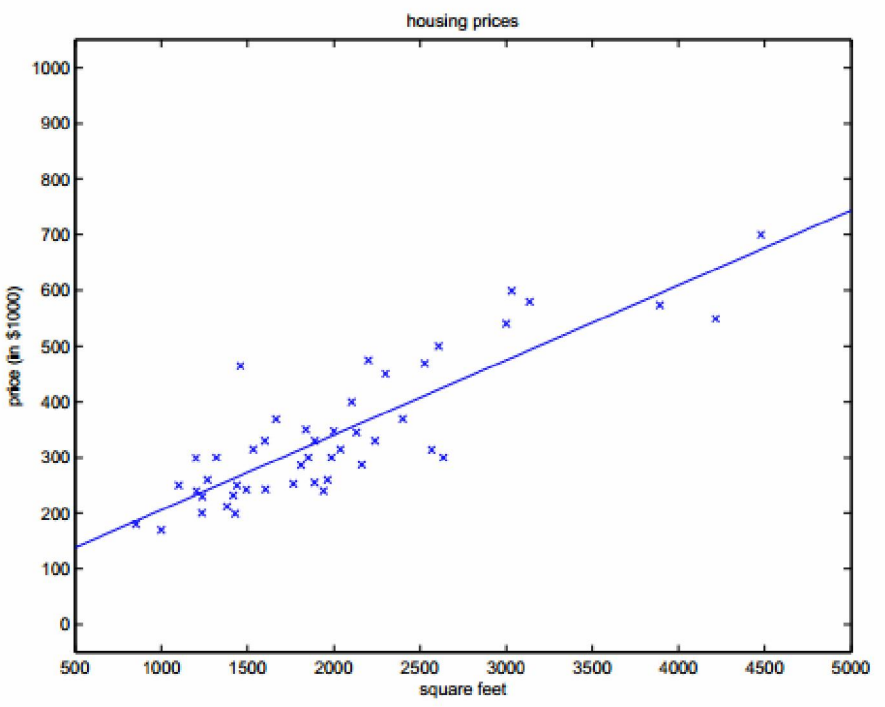
二分k-均值聚类后的结果



线性回归

- y=ax+b


housing prices
price (in $1000)
square feet
0
100
200
300
400
500
600
700
800
900
1000
500
1000
1500
2000
2500
3000
3500
4000
4500
5000


housing prices
price (in $1000)
square feet
0
100
200
300
400
500
600
700
800
900
1000
500
1000
1500
2000
2500
3000
3500
4000
4500
5000

多个变量的情形

- 考虑两个变量

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

最小二乘的目标函数

- m为样本个数，则一个比较“符合常理”的误差函数为：

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

- 思考：如何解释和定义“符合常理”？

使用极大似然估计解释最小二乘

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

the $\epsilon^{(i)}$ are distributed IID (independently and identically distributed) according to a Gaussian distribution (also called a Normal distribution) with mean zero and some variance σ^2

似然函数

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

对数似然

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2
\end{aligned}
$$

计算极大似然函数的最优解

$$\nabla_\theta(X\theta) = X^T$$

$$\begin{aligned}\frac{1}{2}(X\theta-\vec{y})^T(X\theta-\vec{y}) &= \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2 \\ &= J(\theta)\end{aligned}$$

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta\frac{1}{2}(X\theta-\vec{y})^T(X\theta-\vec{y}) \\ &= \frac{1}{2}\nabla_\theta\left(\theta^TX^TX\theta-\theta^TX^T\vec{y}-\vec{y}^TX\theta+\vec{y}^T\vec{y}\right) \\ &= \frac{1}{2}\left(X^TX\theta+X^TX\theta-2X^T\vec{y}\right) \\ &= X^TX\theta-X^T\vec{y}\end{aligned}$$

最小二乘意义下的参数最优解

$$X\theta = \vec{y}$$

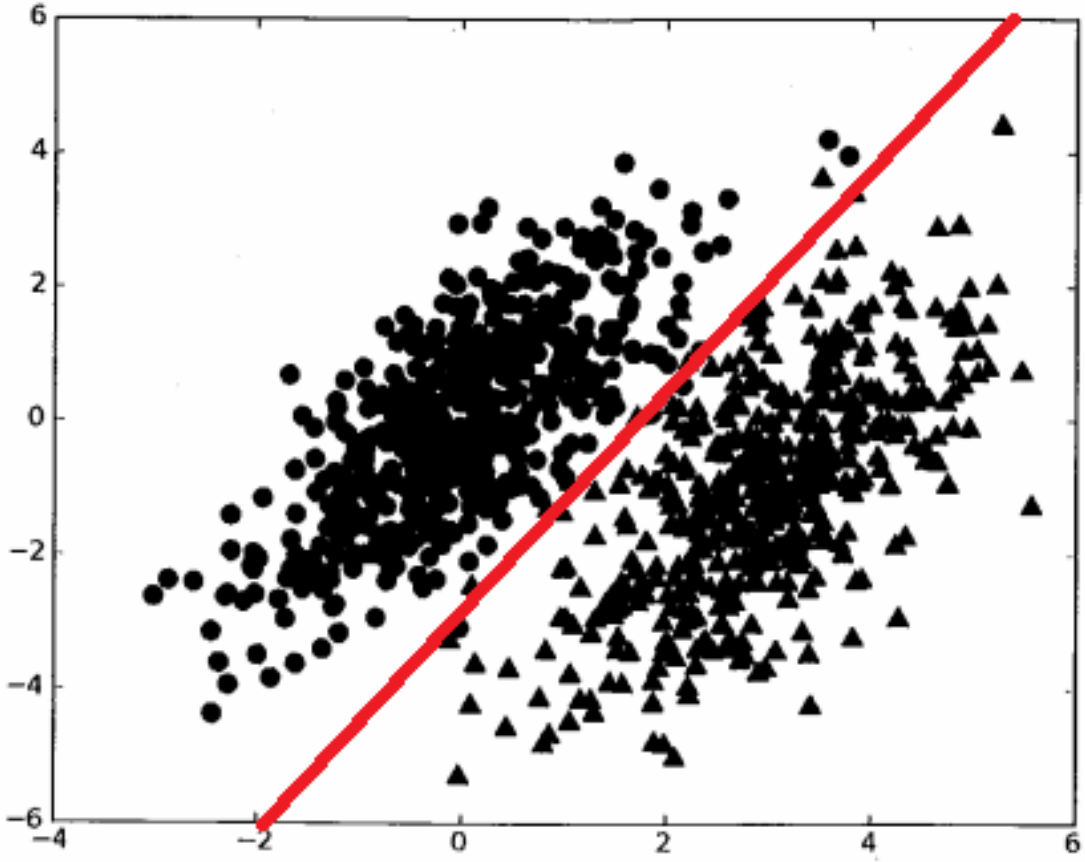
$$X^T X\theta = X^T \vec{y}$$

$$\theta = (X^T X)^{-1} X^T \vec{y}$$

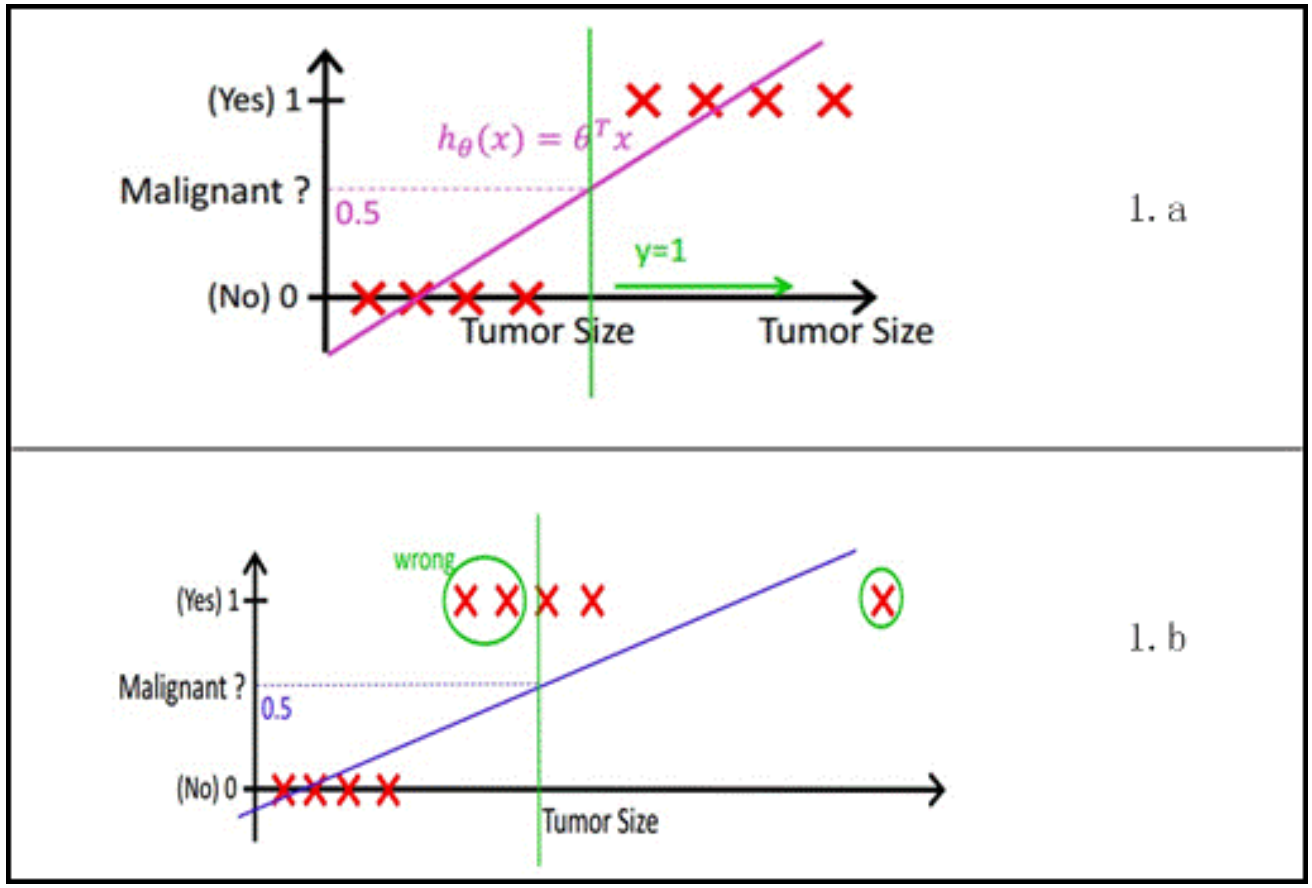
广义逆矩阵（伪逆）$A^+ = \left(A^T A\right)^{-1} A^T$

- 若A为非奇异矩阵,则线性方程组Ax=b的解为$x = A^{-1}b$ 其中A的A的逆矩阵A^{-1}满足$A^{-1}A = AA^{-1} = I$ (I为单位矩阵)。若A是奇异阵或长方阵, $x=A^+ b$。A^+叫做A的伪逆阵。
- 1955年R.彭罗斯证明了对每个m×n阶矩阵A,都存在惟一的n×m阶矩阵X，满足：①AXA=A；②XAX=X；③$(AX)^* = I$；④$(XA)^* = I$。通常称X为A的穆尔-彭罗斯广义逆矩阵,简称M-P逆，记作A^+。
- 在矛盾线性方程组Ax = b的最小二乘解中, $x=A^+b$是范数最小的一个解。
  - 在奇异值分解SVD的问题中，将继续该话题的讨论。

用回归解决分类问题，如何？



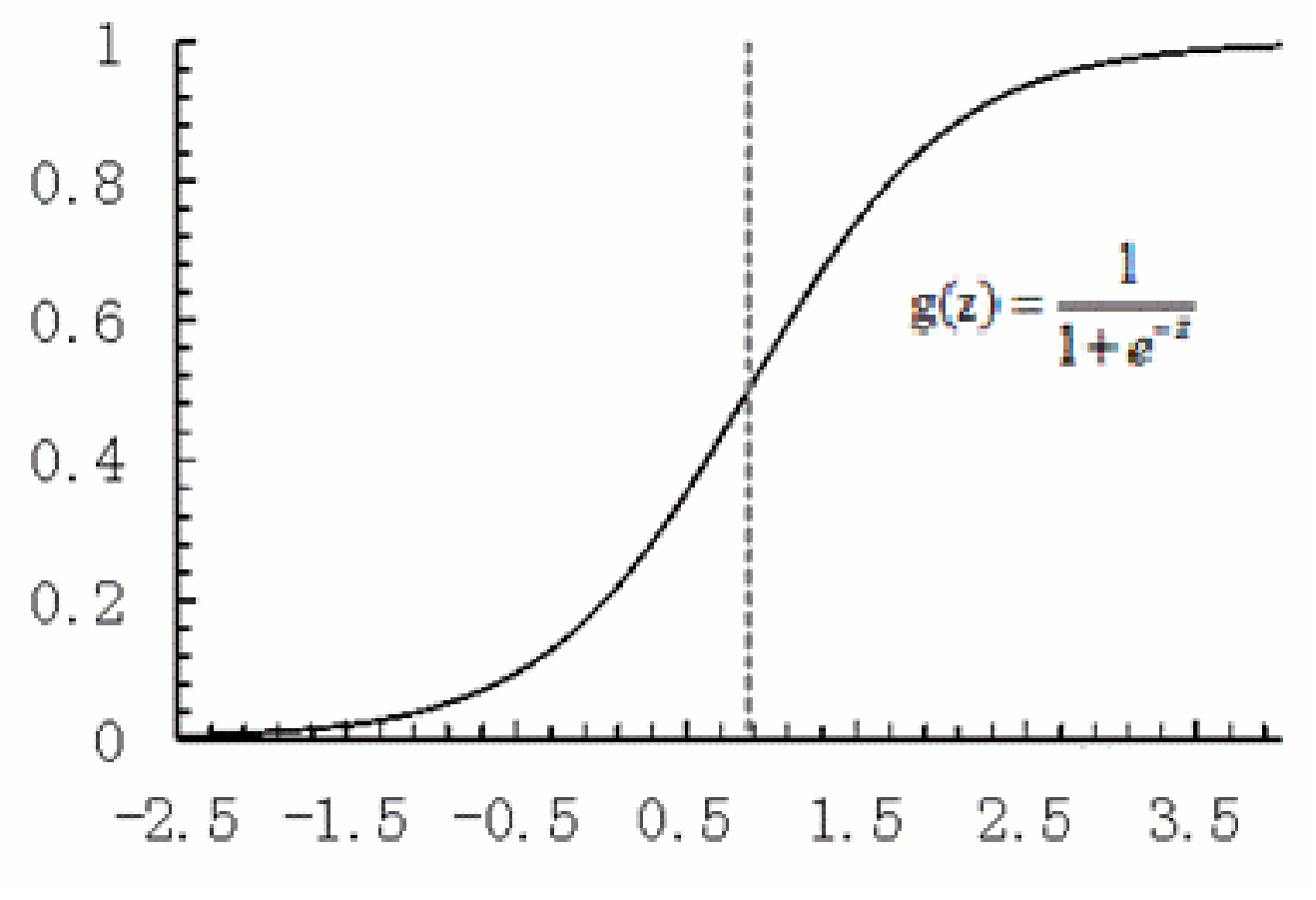
最简单的例子：一维回归


(Yes) 1
Malignant ?
0.5
(No) 0
$h_\theta(x) = \theta^T x$
y=1
Tumor Size
Tumor Size
1. a
wrong
(Yes) 1
Malignant ?
0.5
(No) 0
Tumor Size
1. b

对线性回归的思考

- 若目标y与观测向量X不是线性关系，怎么处理？
- 局部线性回归
  - 非参数方法
- 广义线性回归
  - 对数线性回归
  - Logistic回归

Logistic函数

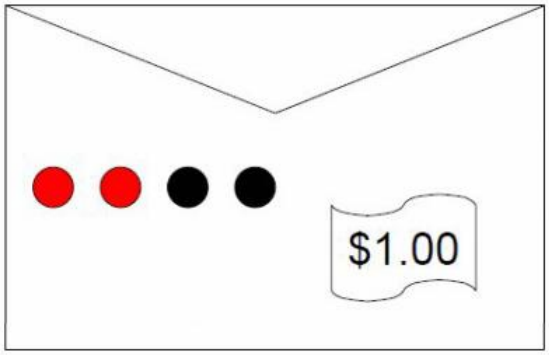


贝叶斯准则

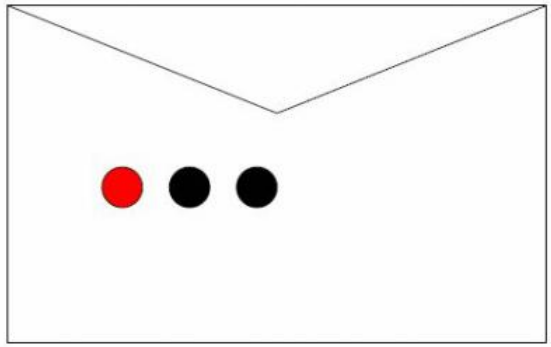
- 条件概率公式
  - $P(x|y) = P(x,y) / P(y) \rightarrow P(x,y) = P(x|y) * P(y)$
  - $P(y|x) = P(x,y) / P(x) \rightarrow P(x,y) = P(y|x) * P(x)$
  - 则$P(x|y) * P(y) = P(y|x) * P(x)$
- 从而： $P(x|y) = P(y|x) * P(x)/ P(y)$

- 分类原则：在给定的条件下，哪种分类发生的概率大，则属于那种分类。

Bayes的实例



The "Win" envelope has a dollar and four beads in it



The "Lose" envelope has three beads and no money

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

后验概率

- c1、c2表示左右两个信封。
- P(R)，P(B)表示摸到红球、黑球的概率。
- P(R)=P(R|c1)*P(c1) + P(R|c2)*P(c2)：全概率公式
- P(c1|R)=P(R|c1)*P(c1)/P(R)
  - P(R|c1)=2/4
  - P(R|c2)=1/3
  - P(c1)=P(c2)=1/2
- 如果摸到一个红球，那么，这个信封有1美元的概率是0.6
- 如果摸到一个黑球，那么，这个信封有1美元的概率是3/7

朴素贝叶斯的假设

- 一个特征出现的概率，与它相邻的特征没有关系（特征独立性）
- 每个特征同等重要（特征均衡性）

以文本分类为例

- 样本：1000封邮件，每个邮件被标记为垃圾邮件或者非垃圾邮件
- 分类目标：给定第1001封邮件，确定它是垃圾邮件还是非垃圾邮件
- 方法：朴素贝叶斯

分析

- 类别c：垃圾邮件c1，非垃圾邮件c2
- 词汇表：统计1000封邮件中出现的所有单词，记单词数目为N，即形成词汇表。
- 将每个样本si向量化：初始化N维向量xi，若词wj在si中出现，则xij=1，否则，为0。从而得到1000个N维向量x。
- 使用：P(c|x)=P(x|c)*P(c) / P(x)

分解

- P(c|x)=P(x|c)*P(c) / P(x)
- P(x|c)=P(x1,x2…xN|c)=P(x1|c)*P(x2|c)…P(xN|c)
- P(x)=P(x1,x2…xN)=P(x1)*P(x2)…P(xN)
- 带入公式： P(c|x)=P(x|c)*P(c) / P(x)

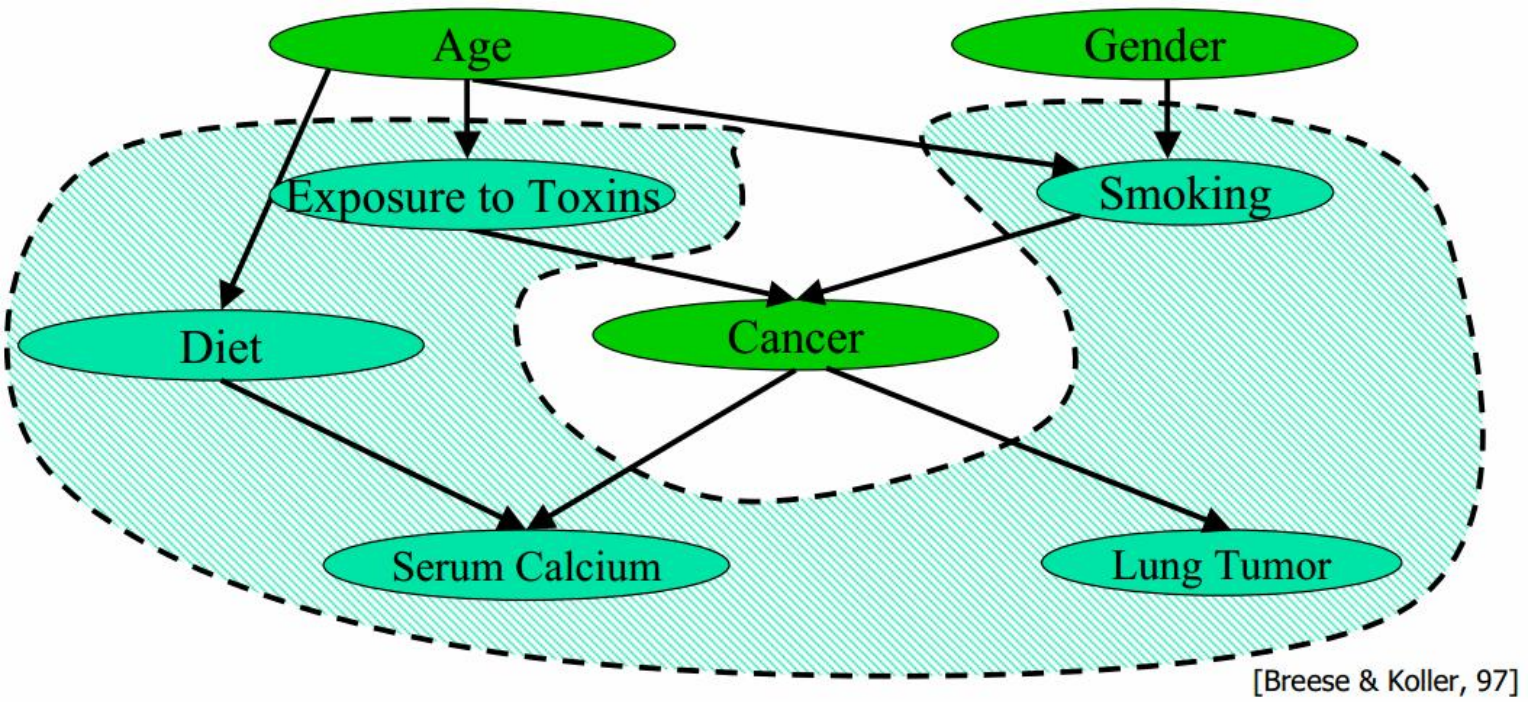
- 等式右侧各项的含义：
  - P(xi|cj)：在cj(此题目，cj要么为垃圾邮件1，要么为非垃圾邮件0)的前提下，第i个单词xi出现的概率
  - P(xi)：在所有样本中，单词xi出现的概率
  - P(cj)：(垃圾邮件)cj出现的概率

对朴素贝叶斯的思考

- 遇到生词怎么办？
  - 拉普拉斯平滑
- 编程的限制：小数乘积怎么办？
- 问题：一个词在样本中出现多次，和一个词在样本中出现一次，形成的词向量相同
  - 由0/1改成计数
- 如何判断两个文档的距离
  - 夹角余弦
- 如何判定该分类器的正确率
  - 样本中：K个生成分类器，1000-K个作为测试集
  - 交叉验证
- 若对象特征之间不独立，会演化成何种形式？

贝叶斯网络

背景知识：Serum Calcium(血清钙浓度)高于2.75mmol/L即为高钙血症。许多恶性肿瘤可并发高钙血症。



事实上，阴影部分的结点集合，是Cancer的“马尔科夫毯”(Markov Blanket)，这将在有向图模型(贝叶斯网络)中继续阐述。

若随机变量无法直接(完全)观察到

- 在西单商场随机挑选100位顾客，测量这100位顾客的身高，假定这100个样本服从正态分布$N(\mu, \sigma)$，试估计参数μ和σ。
- 若样本中存在男性和女性顾客，它们服从$N(\mu_1, \sigma_1)$和$N(\mu_2, \sigma_2)$的分布，试估计$\mu_1, \sigma_1, \mu_2, \sigma_2$。
- 矩估计
- 极大似然估计
- EM算法

参考文献

- Prof. Andrew Ng, Machine Learning, Stanford University
- Pattern Recognition and Machine Learning Chapter 8, M. Jordan, J. Kleinberg, ect, 2006
- 统计学习方法，李航著，清华大学出版社，2012年
- A tutorial on spectral clustering，Ulrike von Luxburg, 2007
- A Tutorial on Inference and Learning in Bayesian Networks, Irina Rish
- 高等数学，高等教育出版社，同济大学数学教研室 主编, 1996

我们在这里

- 更多算法面试题在 7 七月算法
  - http://www.julyedu.com/
    - 免费视频
    - 直播课程
    - 问答社区

- contact us：微博
  - @研究者July
  - @七月问答
  - @邹博_机器学习

感谢大家！

恳请大家批评指正！