

模式识别

度量Metric

信息论Information Theory简介

决策树Decision tree

吴建鑫

南京大学计算机系 & 人工智能学院，2020

目标

- ✓ 理解和掌握度量的基本知识
- ✓ 了解常见的度量
- ✓ 掌握信息论的基本概念
- ✓ 掌握决策树的基本知识
- ✓ 提高目标
 - 进一步能通过独立阅读、了解distance metric learning
 - 进一步能通过独立阅读、了解ensemble of decision tree和random forest

度量

Metric

特征的表示和比较

✓ 两个重要的任务：

- 特征的表示：特征抽取后，如何表示为数学化或者计算机可以理解的数据形式？
 - 到目前为止：所有数据均表示为一个连续的实数值的向量 $x \in \mathbb{R}^d$
- 特征的比较：比较两个点的相似性
 - 在NN、线性分类器、SVM中到目前为止是用欧式距离
 - 在概率方法中，如高斯分布和KDE，也是欧式距离

✓ 对这些数据（实数向量、可以计算距离或相似程度），称为metric data

✓ 但是：还有很多其他类型的数据

更多的数据类型

- ✓ 标记数据Nominal data
 - 如数据1,2,3分别表示苹果、梨和香蕉
 - 不是连续的实数值、也不可以比较大小（ $1 < 2$ 代表苹果不如梨吗？）、不可以比较相似性
- ✓ 时间序列数据time series data
 - 如一个序列(63,64,62)是单个样例，表示某人今天早中晚测量的体重；(61,65)是第二天早晚的体重
 - 不是向量，测量次数不等，如何比较？
- ✓ ...
- ✓ 后续章节将针对不同数据的模式识别进行介绍

更多的度量

- ✓ 目前已用
 - 不相似程度或距离：欧式距离
 - 相似程度：内积或者RBF核
 - 两种紧密关联
- ✓ 但是，数据的不同特点要求使用不同的度量
- ✓ 那么，什么是度量metric？

Metric

- ✓ 一个度量 d 必须满足：对任意向量 $\mathbf{x}, \mathbf{y}, \mathbf{z}$
 - 非负nonnegative: $d(\mathbf{x}, \mathbf{y}) \geq 0$
 - 自反: $d(\mathbf{x}, \mathbf{y}) = 0$ 当且仅当 $\mathbf{x} = \mathbf{y}$
 - 对称symmetric: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
 - 三角不等式triangle inequality: $d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$
- ✓ 欧式距离满足这些条件吗？

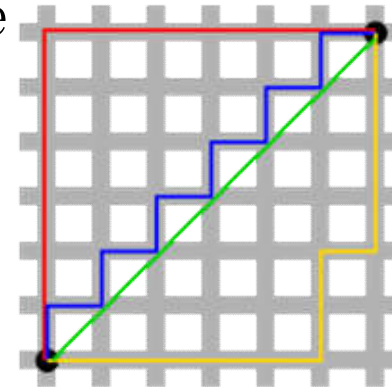
从欧式距离到度量学习

- ✓ Euclidean distance: $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})$
- ✓ Mahalanobis distance: $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})$
 - Σ 是数据的协方差矩阵
 - 练习：若对数据进行白化操作，则原空间中的马氏距离等价于白化变化以后新空间的欧式距离
- ✓ 进一步推广：可以用一个半正定的矩阵 A 代替 Σ^{-1}
 - $d_A^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})$
 - A 半正定，存在 G ，使得 $A = G^T G$
 - 因此， $d_A^2(\mathbf{x}, \mathbf{y}) = \|G\mathbf{x} - G\mathbf{y}\|_2^2$ （如 A 不正定，？）
 - 那么，如何设置 A 的值？可以使用标记信息！
 - 进一步阅读：distance metric learning度量学习

固定形式的distance

✓ Minkowski distance: $d_p(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^d |x_i - y_i|^p)^{\frac{1}{p}}$

- $p \geq 1$ 时是 metric
- $p = 2$ 时是欧式距离
- $p = 1$: $\sum_{i=1}^d |x_i - y_i|$, 称为 Manhattan distance 曼哈顿距离, 或者 city block distance
- 若 $p < 1$, 不是 metric (举例?)
 - 但是有时仍然可以用来比较两样例



图片来自英文Wiki



Norm、distance、similarity

✓ 一个向量 \mathbf{x} 的 p norm (或者 L_p norm) :

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p \right)^{\frac{1}{p}}$$

- 限制条件: $p \geq 1$

- $\|\mathbf{x}\|_\infty = \max(|x_1|, \dots, |x_d|)$

✓ 距离和长度的关系: $d_p(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p$

✓ 从距离 (不相似度) 到相似度, 例如
 $\exp(-\gamma \|\mathbf{x} - \mathbf{y}\|_p)$

幂平均函数

✓ 幂平均 power mean (generalized mean) function

- $M_p(x_1, \dots, x_n) = \left(\frac{1}{n} \sum_{i=1}^n x_i^p \right)^{\frac{1}{p}}$ 要求

$x_i > 0$

- 对 p 在整个实数轴上都有定义 (有些通过极限定义)

- $M_{-\infty} = \min(x_1, \dots, x_n)$

- M_{-1} -- 调和平均 harmonic mean 已经用过

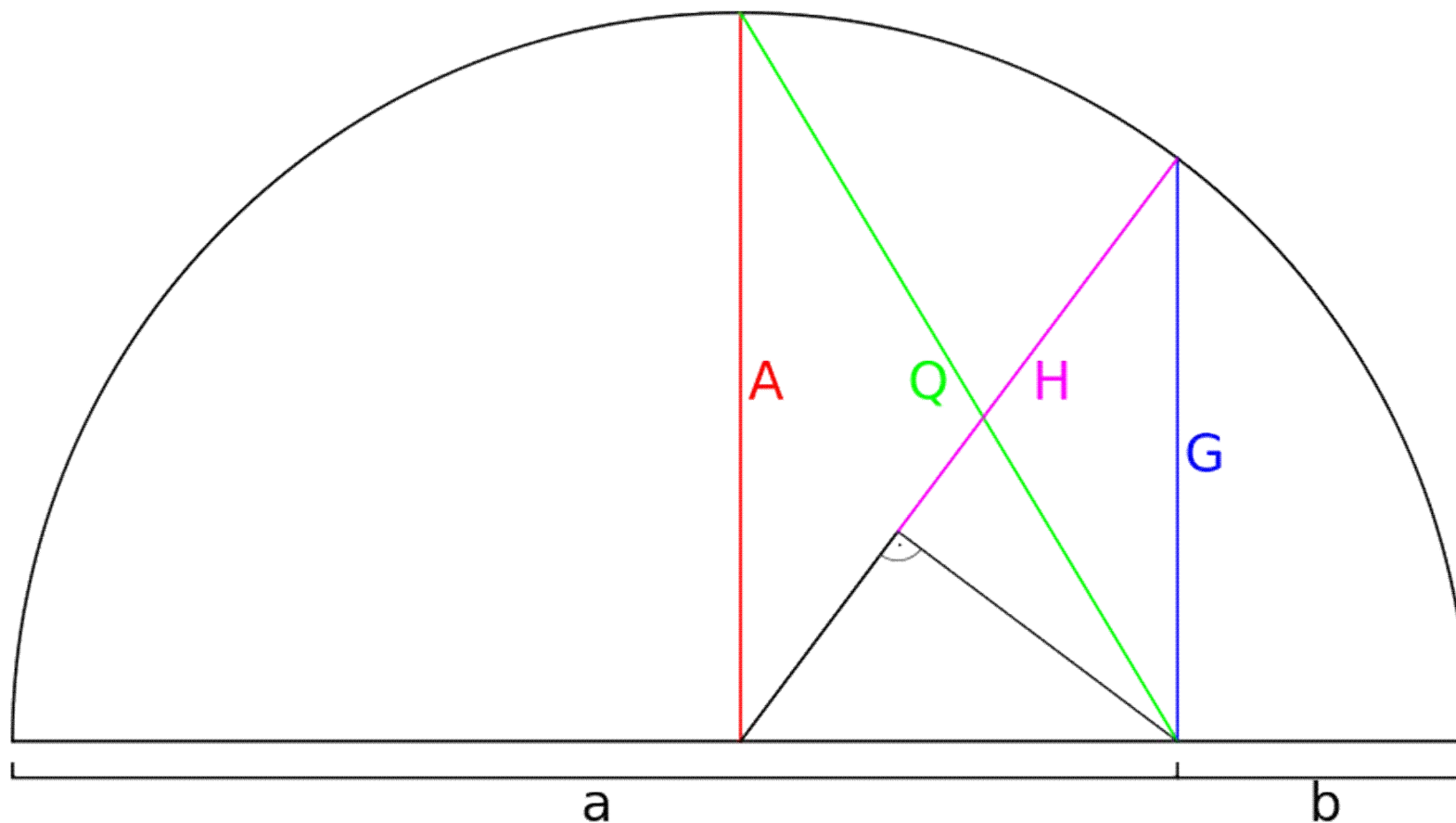
- $M_0 = \sqrt[n]{x_1 x_2 \cdots x_n}$ -- 几何平均

- M_1 -- 算术平均

- M_2 -- root mean square

- $M_{\infty} = \max(x_1, \dots, x_n)$

- 若 $p < q$, 则 $M_p(x_1, \dots, x_n) \leq M_q(x_1, \dots, x_n)$



若考虑两个实数 a 和 b ,
则 $M_p(a, b)$ 可以视为比较他们的相似程度

幂平均核power mean kernel

$$M_p(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^d M_p(x_i, y_i)$$

- ✓ 当 $p \leq 0$ 时，以上函数为Mercer核
- ✓ 属于加性核additive kernel
 - $p = 0$, $\sqrt{x_i y_i}$ -- Hellinger's Kernel
 - $p = -1$, $\frac{2x_i y_i}{x_i + y_i}$ -- χ^2 核
 - $p = -\infty$, $\min(x_i, y_i)$ -- histogram intersection核
(直方图相交?)
- ✓ 当特征是直方图时，加性核效果极佳
- ✓ 进一步阅读：关于加性核

Nominal data

标记数据

标记数据的比较

- ✓ 标记数据Nominal data
 - 如数据1,2,3分别表示苹果、梨和香蕉，怎么比较？
- ✓ 基本思想：相同则为1，否则为0，即两个标记数据 x 和 y 的相似度为 $\mathbb{I}(x = y)$
- ✓ 度量化
 - 设标记数据可以取 m 个不同的值，标记为 $\{1, 2, \dots, m\}$
 - 将标记数据 $x = i$ 转换成一个向量 $\begin{pmatrix} \mathbf{0}_{i-1} & 1 & \mathbf{0}_{d-i} \\ < i & i & > i \end{pmatrix}$
 - 假设 x, y 转换为 \mathbf{x}, \mathbf{y} ，那么 $\mathbb{I}(x = y) = ?$
 - SVM即可用该方法处理标记数据

从度量化到直方图

- ✓ 可以看成，度量化的过程是将一个标记数据转化为一个所有可能取值的直方图
 - 一个直方图histogram是对一个集合中元素的计数
 - 若 $x = i$ ，其度量化的结果 \mathbf{x} 为 m 个bin的直方图
 - 第 i 个bin值为1，表示有一个样例取值为 i
 - 其余所有bin为0，表示没有任何样例取这些值
 - 是一个有效的对集合 $\{\mathbf{x}\}$ 的直方图吗？
- ✓ 那么，假设有两组数据，直方图分别为 \mathbf{x} 和 \mathbf{y}
 - 应该怎么计算其相似性？
 - $\min(x_i, y_i)!$

信息论（极）简介

A (very) brief introduction to the information theory

从直方图到概率分布

- ✓ 在非参数估计中，我们怎么估计一个分布？
 - 最早从直方图开始
- ✓ 那么我们怎么比较两个分布呢？
 - 假设 p 和 q 是两个离散分布，那么HIK可以用吗？怎么用？
 - 如果是连续分布呢？有没有理论上完备的方法？
 - 信息论！ Information theory

信息information

- ✓ 描述一个随机变量需要多少信息？
 - 假设用bit来作为信息的单位
 - 若离散变量满足 $P(x = 2) = 1, P(x \neq 2) = 0$?
 - 若离散变量是 $\{1, 2, 3, 4\}$ 上的均匀uniform分布?
- ✓ 熵entropy
 - $H = -\sum_{i=1}^m P_i \log_2 P_i$ (m 个离散可能取值, 各为 P_i)
 - 如果 $P_i = 0$?
 - 定义 $0 \log_2 0 = 0!$, 因为 $\lim_{x \rightarrow 0} x \log_2 x = 0$
 - 什么时候最大? 什么时候最小?
 - 均匀分布的时候最大, $\log_2 m$
 - 单点分布最小, 0

Differential entropy

✓ 如果分布是连续的？

$$h(x) = -\int p(x) \ln(p(x)) dx$$

- 自然对数，单位是nat（奈特）

- 若 $X \sim N(\mu, \sigma^2)$ ，则 $h(X) = \frac{1}{2} \ln(2\pi e \sigma^2)$ nats

✓ 在所有均值和方差固定的连续分布中，高斯分布具有最大的熵

- 或者说，不确定性uncertainty最大

Joint, conditional entropy

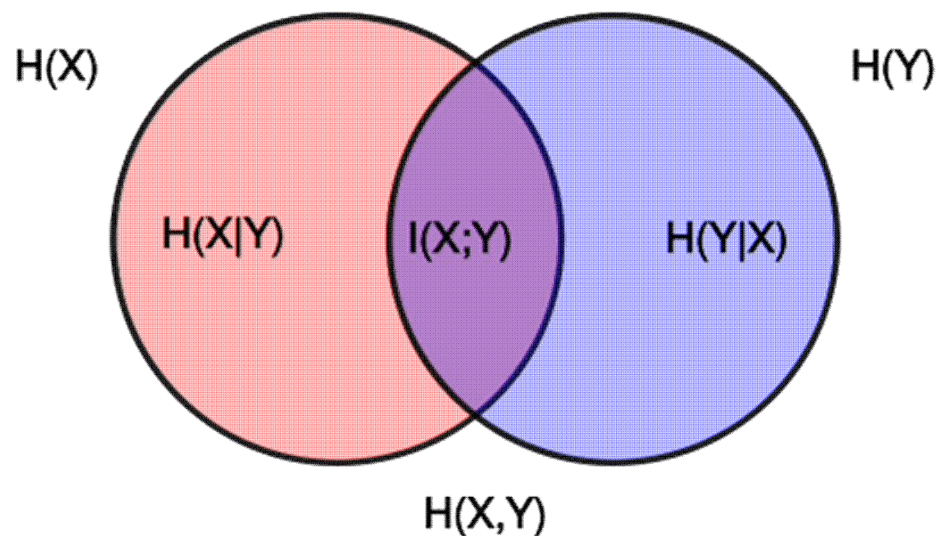
✓ $H(X, Y) = -\sum_x \sum_y P(x, y) \log_2 P(x, y)$

✓ $h(X, Y) = -\int p(x, y) \ln p(x, y) dx dy$

✓ $H(X|Y) = \sum_y p(y) H(X|Y = y) =$
 $\sum_{x,y} P(x, y) \log_2 \frac{P(y)}{P(x, y)} = -\sum_{x,y} P(x, y) \log_2 \frac{P(x, y)}{P(y)}$

✓ $h(X|Y) = -\int p(x, y) \ln p(x|y) dx dy =$
 $-\int p(x, y) \ln \frac{p(x, y)}{p(y)} dx dy$

各种熵之间的关系



用“描述长度”来记忆

- $H(X, Y) = H(Y) + H(X|Y)$
- $H(X, Y) = H(X) + H(Y|X)$
- $H(X|Y) \leq H(X)$
- $H(Y|X) \leq H(Y)$

问题

- $H(X|Y) = H(Y|X)$?
- 那么 $I(X;Y)$ 代表什么?

互信息 Mutual information

✓ 如果 X 和 Y 互相独立, 即 $p(x, y) = p(x)p(y)$, 或者 $P(x, y) = P(x)P(y)$

- 上面的图应该怎么画?
- $I(X; Y)$ 表示 X 和 Y 共同的那部分信息

$$I(X; Y) = H(X) - H(X|Y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$
$$= H(Y) - H(Y|X)$$

✓ $I(X; Y) = I(Y; X)$?

✓ 可以粗略的看成相似程度或者相关程度

KL散度

✓ Kullback-Leibler divergence: 两个离散分布 P 和 Q

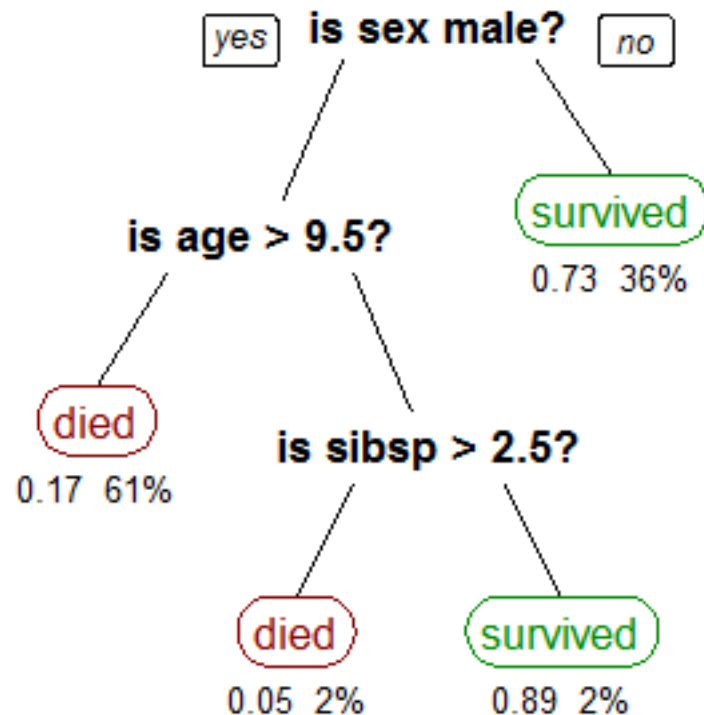
$$D_{KL}(P||Q) = \sum_i P_i \log_2 \frac{P_i}{Q_i}$$

- $D_{KL}(P||Q) \geq 0$, 等号当且仅当 $\forall i, P_i = Q_i$ 时成立
- $I(X; Y) = D_{KL}(p(x, y)||p(x)p(y))$
- 可以粗略看成“距离”
- 但是, KL散度对称吗?

决策树

Decision Tree

Titanic survivors

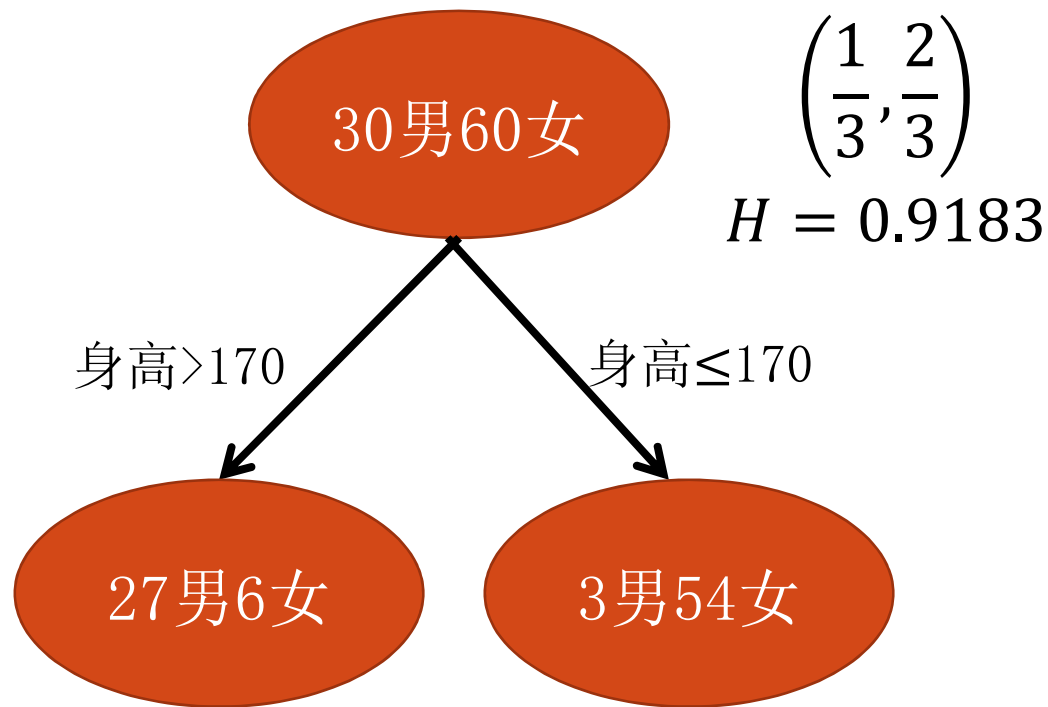


- 该判断模型是树tree
- 每次根据一个数据（称为属性）分成若干部分
- 当不可再分时（叶节点），给出一个决策decision
 - 通常输出的决策是标记数据
 - 可以输出一个概率分布

那么, 选哪个属性来分?

- ✓ 问题的输出是标记数据, 有 m 个可能的值
- ✓ 如果当前节点一共包含 n 个样例, 记为集合 T
- ✓ 其对应样例的grountruth输出是集合 y_T
- ✓ 计算 $H(y_T)$ - 当前节点的不纯度impurity
- ✓ 对每一个属性 j
 - 其不同值将当前节点数据分为若干子集 T_1, T_2, \dots
 - 计算每个子集的entropy: $H(y_{T_k})$ 和比例 w_k
 - 计算按此属性分开后的平均不纯度 $\sum_k w_k H(y_{T_k})$
 - Information gain信息增益: $H(y_T) - \sum_k w_k H(y_{T_k})$
- ✓ 选择信息增益最大的那个属性

示意图：判断性别



Information gain:
0.4791

$$H - (w_1 H_1 + w_2 H_2)$$

$$\left(\frac{9}{11}, \frac{2}{11}\right) \quad w_1 = \frac{33}{90} \quad \left(\frac{1}{19}, \frac{18}{19}\right) \quad w_2 = \frac{57}{90}$$
$$H_1 = 0.6840 \quad H_2 = 0.2975$$

其他问题

- ✓ 信息增益是一种选择的方法，其他方法很多
 - 在数据挖掘课程中讲述，这里不讲
- ✓ 但是，可以想象可能存在的其余问题？
 - 分到什么程度为止？即，什么时候不再分了？
 - 如果某属性有100个可能的取值，分100个嘛？
 - 其中有连续属性怎么办？
 - 计算和存储复杂度是多少？
 - ...

进一步的阅读

- ✓ 如果对本章的内容感兴趣，可以参考如下文献
 - Distance metric learning: http://www.cse.ohio-state.edu/~kulis/pubs/ftml_metric_learning.pdf
 - 加性核: [W5] in <http://cs.nju.edu.cn/wujx/publication.htm>
 - 信息论: Elements of Information Theory, 2nd edition, <http://www.amazon.com/Elements-Information-Theory-Telecommunications-Processing/dp/0471062596>
 - Information Theory, Inference, and Learning Algorithms, by David MacKay, <http://www.inference.phy.cam.ac.uk/itila/book.html>
- ✓ Random forest: http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm