1 DecisionTree

通过找到最合适的 point 去把数据分为两个子数据
属性类集 **C**, 第 i 个属性 $\mathbf{C}_i \in \mathbf{C}$, 第 i 个属性取值为 j, 记作 $\mathbf{C}_i = j$
离散属性先转为多列的 0 和 1 看做多个连续属性

1.1 Classifier

数据 **D**, 标签集 S 遍历所有的 point
数据集被 point 分为两份 **A** 和 **B**, 得到占比 $P(\mathbf{A} \mid \mathbf{D})$,$P(\mathbf{B} \mid \mathbf{D})$ 得到占比

A	$P_A = [P(S = S_1 \mid \mathbf{A}), P(S = S_2 \mid \mathbf{A}) \dots P(S = S_n \mid \mathbf{A})]$
B	$P_B = [P(S = S_1 \mid \mathbf{B}), P(S = S_2 \mid \mathbf{B}) \dots P(S = S_n \mid \mathbf{B})]$

得到信息熵

$$En_{poitn} = P(\mathbf{A} \mid \mathbf{D}) P_A^\top \log_2(P_A) + P(\mathbf{B} \mid \mathbf{D}) P_B^\top \log_2(P_B)$$

找出信息熵最小即使最佳的 point

1.2 Regression

标签值 S 数据集被 point 分为两份 **A** 和 **B**,$S_\mathbf{A}$ $S_\mathbf{B}$
计算 $S_\mathbf{A}$ $S_\mathbf{B}$ 平方误差。再相加得 E_{point}
取 E_{point} 最小时的 point