y y n y n y n n y n n n n n n | 0.9 0.85 0.8 0.75 0.7 0.65 0.6 0.55 0.5 0.45 0.4 0.35 0.3 0.25 0.2

, , ,

**Threshold ≥ 0.9**. 所以**TP:** 1, **FP:** 0, **FN:** 4, **TN:** 10, **TPR:** = 0.2, **FPR**: , 所以点是（0，0.2）

**Threshold ≥ 0.85.** 所以**TP:** 2, **FP:** 0, **FN:** 3, **TN:** 10, **TPR**: = 0.4, **FPR:** , 所以点是（0，0.4）

**Threshold ≥ 0.75.** 所以**TP:** 3, **FP:** 1, **FN:** 2, **TN:** 9, **TPR:** = 0.6, **FPR**: , 所以点是（0.1，0.6）

**Threshold ≥ 0.65.** 所以**TP:** 4, **FP:** 2, **FN:** 1, **TN:** 8, **TPR:** = 0.8, **FPR**: , 所以点是（0.2，0.8）

**Threshold ≥ 0.55.** 所以**TP:** 4, **FP:** 4, **FN:** 1, **TN:** 6, **TPR:** = 0.8, **FPR**: , 所以点是（0.4，0.8）

**Threshold ≥ 0.5.** 所以**TP:** 5, **FP:** 4, **FN:** 0, **TN:** 6, **TPR:** = 1.0, **FPR**: , 所以点是（0.4，1.0）

*What is the accuracy at threshold 0.7?*

**Threshold ≥ 0.7.** 所以**TP:** 3, **FP:** 2, **FN:** 2, **TN:** 8, = = 0.733

What is the value of accuracy, Precision and Recall for dummy classifier that always predicts 'positive' on test set containing 30 positive and 70 negative instances?

**真正例（TP）**：实际为正类，预测为正类。TP = 30

**假正例（FP）**：实际为负类，预测为正类。FP = 70

**假负例（FN）**：实际为正类，预测为负类。FN = 0（因为始终预测正类）

**真负例（TN）**：实际为负类，预测为负类。TN = 0

带入公式计算Accuracy, Precision and Recall, 结果是0.3, 0.3, 1.0

Consider the following dummy classifiers: DS with its strategy set to "stratified", DF with its strategy set to "most frequent", DC with its strategy set to "constant" with value A

We fit them to a training set of instances with class values A, B, C, and D. Their respective proportions are 10%, 20%, 30%, and 40%. We also have a test set of instances with the same proportions.

What will the average accuracies of DS and DF be on our test set?

DS 的准确率=(0.1)2+(0.2)2+(0.3)2+(0.4)2=0.01+0.04+0.09+0.16=0.30

**DF（most frequent）** 始终预测**最频繁**的类别 D。

因为测试集中类别 D 占 40%。所以DF 的准确率=40%=0.4DF 的准确率=40%=0.4

What will the average precisions of DS and DC with respect to class A be on our test set?

**DS** 预测为 A 的概率为 10%。实际为 A 的概率也是 10%。因此，预测为 A 的实例中，实际为 A 的比例：

**DC** 始终预测 A。预测为 A 的实例全部为 DC 的预测，总数为 100%。实际为 A 的实例占 10%。

所以

What will the average recalls of DS and DC with respect to class A be on our test set?

**计算 DS和DC 对类别 A 的召回率**

1. **DS** 对实际为 A 的实例预测为 A 的概率为 10%。
2. **DC** 始终预测 A，因此所有实际为 A 的实例都被正确预测。
3. 所以是(0.1, 1.0)

|  |  |  |
| --- | --- | --- |
| Age | Fitness | Hiking |
| 20 | H | Y |
| 30 | L | Y |
| 18 | H | Y |
| 25 | L | N |
| 40 | H | N |
| 15 | H | Y |

1. What is the Gini index for a split on Age < 25 and Age >= 25?

Age <25:( 20 18 15| Y Y Y), Age >= 25: (30, 25, 40 | Y N N)

;

所以是0.22

1. What is the Gini index for a split on Fitness

**Fitness = High**:20,18,40,15 - Hiking: Yes, Yes, No, Yes - 类别比例：Yes =3/4, No =1/4

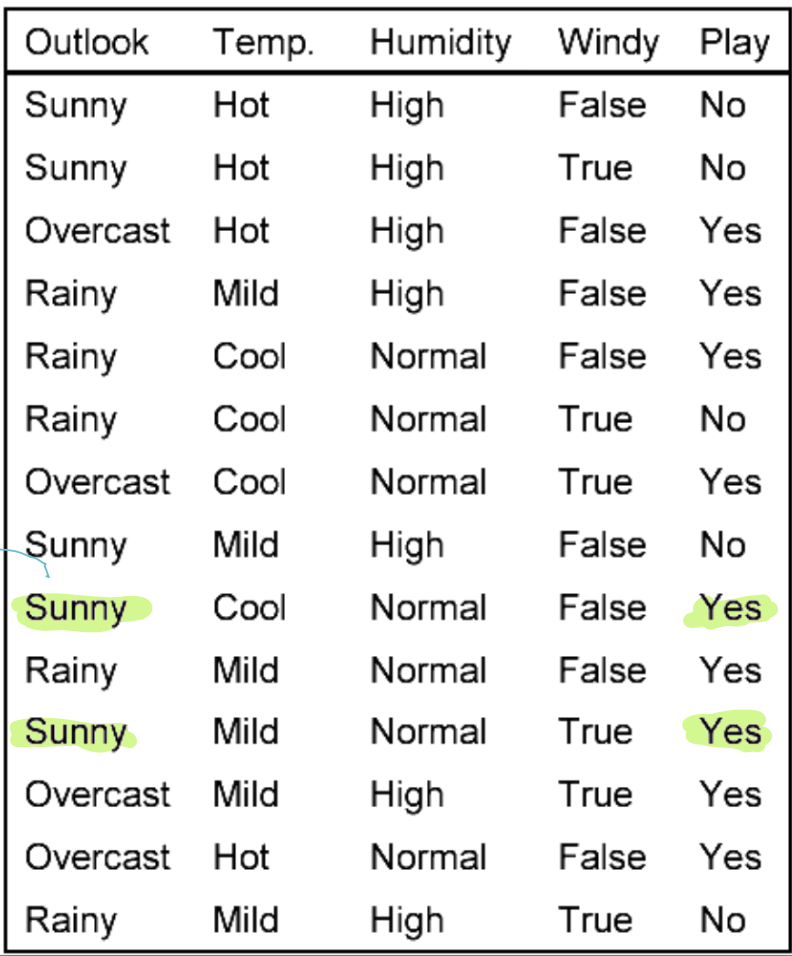
**Fitness = Low**:30,25 - Hiking: Yes, No - 类别比例：Yes =1/2, No =1/2.

所以是0.4167

然后Age < 25 and Age >= 25更适合因为他的Gini index小

这是Bayes

文本

描述已自动生成

文本

描述已自动生成文本

描述已自动生成

这是接着算具体的概率：

1. P(play=yes | E) + P(play=no | E) = 1

0.0053 / P(E) + 0.0206 / P(E) = 1

P(E) = 0.0053 + 0.0206

So,

1. P(play=yes | E) = 0.0053 / (0.0053 + 0.0206) = 20.5%

P(play=no | E) = 0.0206 / (0.0053 + 0.0206) = 79.5%

文本

中度可信度描述已自动生成

这是带数值的（只有temperature和humidity有数值）

例子中只有Yes的，No的也要计算

图片包含 表格

描述已自动生成

图形用户界面, 文本, 应用程序

描述已自动生成

Naïve Bayes

这类只看Training set是否包含关键字, 公式是直接套 +1和+2

图标

中度可信度描述已自动生成

公式中

* 是类别c中包含词t的文档数
* 是类别c中的文档总数
* +1 和 +2 是拉普拉斯平滑

举例

* 图片包含 文本

  描述已自动生成在的3个文档中，3个都包含"Chinese"
* = 3, = 3
* P(Chinese|c) = (3 + 1) / (3 + 2) = 4/5 = 0.8

这类是要看Training Set中的有多少关键字，需要统计得出N

图片包含 文本

描述已自动生成

公式解释公式为：

图标

描述已自动生成

* 是类别 *c* 中单词 t的出现次数。
* V 是词汇表的大小。

**具体计算**

1. **P(Chinese|c)**:

（在类别为“yes”的文档中，“Chinese”出现了5次）  
-词汇表大小 ∣V∣=6（总共有6个不同的单词）

分母：（类别为“yes”的文档中所有单词的总数）

-计算：

文本

描述已自动生成

线性回归

X = [1, 2, 3], y = [2,3,4], 学习率 α = 0.1

线性回归模型：

代价函数是：

计算初始代价函数值：  
m = 3 (数据点数量)

1. 对于 (1,2): ，误差为 1−2=−1，平方误差为 1。

2. 对于 (2,3): ，误差为 2−3=−1，平方误差为 1。

3. 对于 (3,4): ，误差为 3−4=−1，平方误差为 1。

*计算梯度：*

*更新参数：*

*计算更新后的代价函数值：*