比如兩組樣本:

第一組有以下三個樣本:3，4，5

第二組有一下三個樣本:2，4，6

這兩組的平均值都是 4，但是第一組的三個數值相對更靠近平均值，也就是離散程度小，均方差就是表示這個的。

同樣，方差、標準差(方差開根,因為單位不統一)都是表示資料的離散程度的。

因為誤差有正有負，如果用算術平均數計算，正負值抵消了好一大半，再除以誤差的個數，則算術平均數會很小。
所以將每個點的誤差值取平方值，再計算算術平均數，再開根號。則所有的值為正值，較能顯現出實際狀況。即均方根誤差法。
例如：以理想真圓為基準，測量某一圓 10 個位置的半徑誤差為：
+0.12，+0.05，-0.08，+0.03，-0.02，-0.11，-0.06，+0.01，-0.04，+0.05
若用算術平均數計算
=(0.12+0.05-0.08+0.03-0.02-0.11-0.06+0.01+0.04+0.05)/10
=0.003→幾乎接近真圓，與數據顯現的情況不符
若用均方根誤差計算
=√{[0.122+0.052+(-0.08)2+0.032+(-0.02)2+(-0.11)2+(-0.06)2+0.012+(-0.04)2+0.052]/10}
=0.0667→較接近數據顯現的情況
所取的測量點越多，越接近真實狀況。

機器學習---樸素貝葉斯分類器

朴素贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

通俗来说，就好比这么个道理，你在街上看到一个黑人，我问你你猜这哥们哪里来的，你十有八九猜非洲。为什么呢？因为黑人中非洲人的比率最高，当然人家也可能是美洲人或亚洲人，但在没有其它可用信息下，我们会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础。

樸素貝葉斯分類器是一組簡單快速的分類算法，在機器學習中，我們有時需要解決分類問題。也就是說，給定一個樣本的特征值（feature1,feature2,...feauren），我們想知道該樣本屬於哪個分類標簽（label1,label2,...labeln）。即：我們想要知道該樣本各個標簽的條件概率 P(label|features)是多少，這樣我們就可以知道該樣本屬於哪個分類。例如：假設數據集一共有 2 個分類（標簽），如果現在出現一個新的樣本，其 P(label1|features)>P(label2|features)，那麼我們就可以判定該樣本的標簽為 label1。

樸素貝葉斯算法的分類流程

讓我舉一個例子。下面我設計了一個天氣和響應目標變量「玩」的訓練數據集（計算「玩」的可能性）。我們需要根據天氣條件進行分類，判斷這個人能不能出去玩，以下是步驟：

步驟 1：將數據集轉換成頻率表；

步驟 2：計算不同天氣出去玩的機率，並創建似然表，如陰天的機率是 0.29；

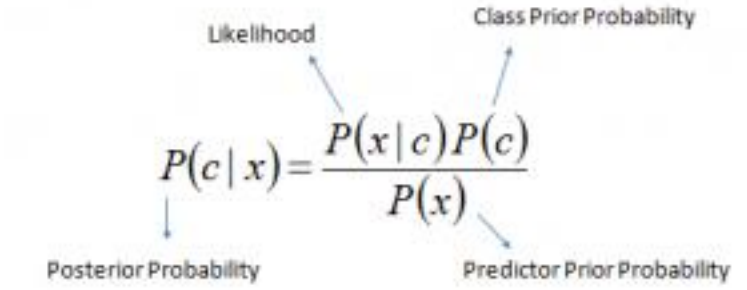
Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

步驟 3：使用貝葉斯公式計算每一類的後驗機率，數據最高那欄就是預測的結果。

問題：如果是晴天，這個人就能出去玩。這個說法是不是正確的？



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

- P(clx)是已知某樣本(c，目標)，(x，屬性)的機率。稱後驗機率。
- P(c)是該樣本「c」的機率。稱先驗機率。
- P(xlc)是已知該樣本「x」，該樣本「c」的機率。
- P(x)是該樣本「x」的機率。

P(是|晴朗)=P(晴朗|是)×P(是)/P(晴朗)

在這裡，P(晴朗|是)= 3/9 = 0.33，P(晴朗)= 5/14 = 0.36，P(是)= 9/14 = 0.64

現在，P(是|晴朗)=0.33×0.64/0.36=0.60，具有較高的機率。

樸素貝葉斯的 4 種應用

實時預測：毫無疑問，樸素貝葉斯很快。

多類預測：這個算法以多類別預測功能聞名，因此可以用來預測多類目標變量的機率。

文本分類/垃圾郵件過濾/情感分析：相比較其他算法，樸素貝葉斯的應用主要集中在文本分類（變量類型多，且更獨立），具有較高的成功率。因此被廣泛應用於垃圾郵件過濾（識別垃圾郵件）和情感分析（在社交媒體平台分辨積極情緒和消極情緒的用戶）。

推薦系統：樸素貝葉斯分類器和協同過濾結合使用可以過濾出用戶想看到的和不想看到的東西。