# 用于统计分类的贝叶斯学习 用于统计分类的贝叶斯学习 原文链接： 贝叶斯统计分类问题的介绍 概率论是我们描述宇宙最基本的工具之一。它与统计分类特别相关，可用于推导出大量重要结果并为我们的理解提供信息。 团队要求Peter Mills帮助您理解概率论和贝叶斯定理以及它们如何应用于统计分类。它将允许您导出非显而易见的结果，这些结果可以极大地改善和简化您的分类模型。 img img 这种对贝叶斯学习进行统计分类的介绍将提供贝叶斯定理和概率在统计分类中的应用的几个例子。它还将超出一般概率知识，涵盖该领域的其他重要领域，包括校准和验证。

原文链接：[Bayesian Learning for Statistical Classification](https://blog.statsbot.co/bayesian-learning-for-statistical-classification-f2362d620428?from=hackcv&hmsr=hackcv.com&utm_medium=hackcv.com&utm_source=hackcv.com)

## 贝叶斯统计分类问题的介绍

*概率论是我们描述宇宙最基本的工具之一。它与统计分类特别相关，可用于推导出大量重要结果并为我们的理解提供信息。*

[*Statsbot*](http://statsbot.co/?utm_source=blog&utm_medium=article&utm_campaign=bayesian_learning) *团队要求Peter Mills帮助您理解概率论和贝叶斯定理以及它们如何应用于统计分类。它将允许您导出非显而易见的结果，这些结果可以极大地改善和简化您的分类模型。*

img

img

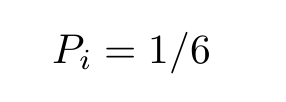
这种对贝叶斯学习进行统计分类的介绍将提供贝叶斯定理和概率在统计分类中的应用的几个例子。它还将超出一般概率知识，涵盖该领域的其他重要领域，包括校准和验证。

请注意，本文虽然是针对初学者的，但仍假设了大学一二年级水平数学知识，特别是线性代数，还有一些单变量和多变量微积分。如果方程式较难理解，请尝试着重于解决实际问题。  
请注意，本文虽然是针对初学者的，但仍假设了大学一二年级水平数学知识，特别是线性代数，还有一些单变量和多变量微积分。如果方程式较难理解，请尝试着重于解决实际问题。  
通过一些例子，您将学习更多关于概率和统计分类的知识，而不仅仅是阅读它或阅读方程式。出于这个原因，我们在文章的最后准备了一系列问题。  
回顾基本概率知识  
假设我们掷骰子。将有六种可能性，每种可能性（在一样重的骰子中）有1/6的概率。我们可以写成这样  
img  
其中i是骰子顶部的数字。由于至少有一面一定在顶部，我们也可以写：

通过一些例子，您将学习更多关于概率和统计分类的知识，而不仅仅是阅读它或阅读方程式。出于这个原因，我们在文章的最后准备了一系列问题。

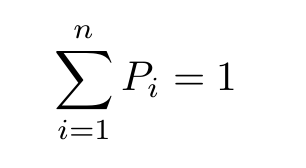
### 回顾基本概率知识

假设我们掷骰子。将有六种可能性，每种可能性（在一样重的骰子中）有1/6的概率。我们可以写成这样



img

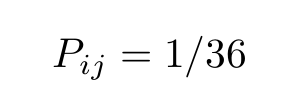
其中*i*是骰子顶部的数字。由于至少有一面一定在顶部，我们也可以写：



img  
img  
其中n = 6是可能性的总数。  
现在假设我们掷两个骰子。获得36对数字之一的联合概率：  
img  
其中i是第一个骰子上的数字，j是第二个骰子上的数字。  
如果我们忽略第二个骰子上的数字，则给出在第一个骰子上获得一定数量（比如6）的概率：  
img  
这被称为先验概率。  
事情变得越来越复杂。鉴于另一个骰子上的某个数字已经出现，在一个骰子上获得一定数量的概率是多少？在这种情况下，这两个事件是不相关的，因此1/6的值将始终相同，但这种情况确是不一定的。

其中*n* = 6是可能性的总数。

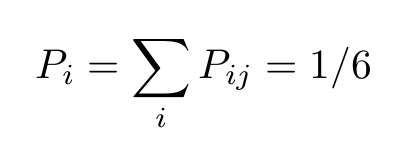
现在假设我们掷两个骰子。获得36对数字之一的联合概率：



img

其中*i*是第一个骰子上的数字，*j*是第二个骰子上的数字。

如果我们忽略第二个骰子上的数字，则给出在第一个骰子上获得一定数量（比如6）的概率：



img

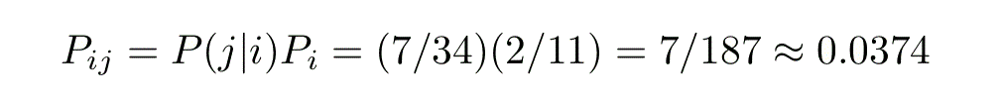
这被称为*先验概率*。

事情变得越来越复杂。鉴于另一个骰子上的某个数字已经出现，在一个骰子上获得一定数量的概率是多少？在这种情况下，这两个事件是不相关的，因此1/6的值将始终相同，但这种情况确是不一定的。

考虑二十一点游戏。考虑到前一张牌十点，下一张牌的概率是多少（十点或一张面牌）的概率是多少？  
考虑二十一点游戏。考虑到前一张牌十点，下一张牌的概率是多少（十点或一张面牌）的概率是多少？  
假设在最后一次抽签之前剩下34个牌组中有7张十点牌。现在，概率根据前一事件的结果而不同。如果之前的卡十点，则有一个6/33 = 2/11的机会获得一张十点的卡，否则概率为7/33。  
由于前一张卡值十的概率是7/34，因此联合概率或两种事件发生的概率是：  
img  
where Pi is the probability that the previous card was worth ten and P(j | i) is the conditional probability that the next card will be worth ten, given that the previous card was also worth ten.

假设在最后一次抽签之前剩下34个牌组中有7张十点牌。现在，概率根据前一事件的结果而不同。如果之前的卡十点，则有一个6/33 = 2/11的机会获得一张十点的卡，否则概率为7/33。

由于前一张卡值十的概率是7/34，因此联合概率或两种事件发生的概率是：

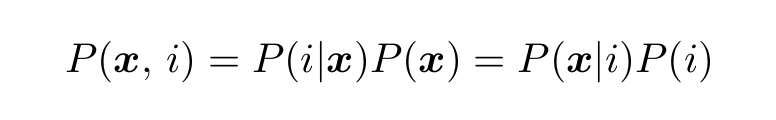


img

where *Pi* is the probability that the previous card was worth ten and *P(j | i)* is the conditional probability that the next card will be worth ten, given that the previous card was also worth ten.

With prior, joint, and conditional probabilities defined, we are set to write down Bayes’ theorem.  
定义了先验概率、联合概率和条件概率，我们就可以写出贝叶斯定理。

Note that these definitions are symmetric in i and j, thus:  
注意，这些定义在i和j中是对称的，因此：

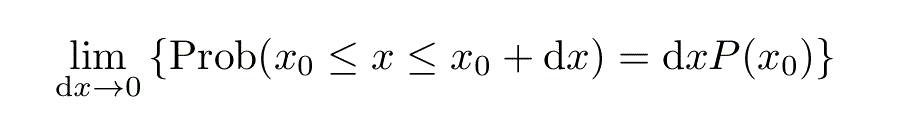


img  
img公司

which is the symmetric form of Bayes’ Theorem.  
这是贝叶斯定理的对称形式。

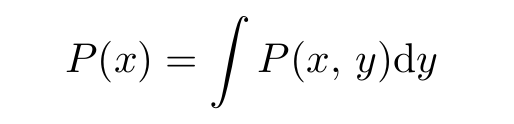
### Continuous probabilities 连续概率

The extension to continuous probabilities or probability densities is straightforward. Imagine we have a continuous random variable, x, governed by a probability distribution, P(x). The probability that x takes on a value between xₒ and xₒ+dx is given:  
对连续概率或概率密度的扩展是直接的。假设我们有一个连续的随机变量x，由概率分布P（x）控制。x取xₒ和xₒ+dx之间的值的概率如下：



img  
img公司

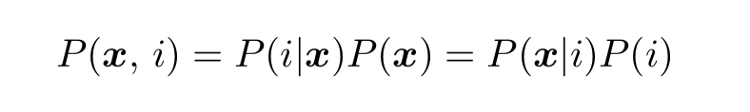
When working with continuous random variables, summations become integrals so that Equation (2) becomes:  
当使用连续随机变量时，求和变为积分，因此方程（2）变为：



img  
img公司

where P(x, y) is the joint probability of both x and y and the integral is over all of x.  
其中P（x，y）是x和y的联合概率，积分在x上。

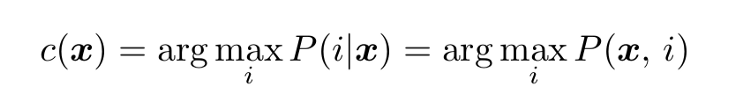
In statistical classification, we are dealing with probabilities having a very specific form. One of the variables is scalar and discrete, while the other is vector and continuous:  
在统计分类中，我们处理的是具有非常特殊形式的概率。其中一个变量是标量和离散变量，而另一个是矢量和连续变量：



img  
img公司

Where i is the class or class labeland x is a vector of attributes or features.  
其中i是类或类labeland x是属性或特征的向量。

Typically, the goal of Bayesian-based statistical classification is to estimate either the joint probability, P(x, i), or the conditional probability, P(i | x). Classifications are normally done on the basis of maximum likelihood:  
通常，基于贝叶斯的统计分类的目标是估计联合概率P（x，i）或条件概率P（i | x）。分类通常在最大似然的基础上进行：



img  
img公司

where c is the most likely estimate for the class, that is the index of the largest value of the conditional probability.  
其中c是类的最可能估计，即条件概率最大值的索引。

Note that because P(x) is the same for a given test point, using either the joint or the conditional probability will produce the same result. The conditional probabilities of the feature space, P(x | i), are important also as these describe the distributions of each isolated class: that is, if you remove all other class labels leaving only i, this is the distribution you are left with.  
注意，因为P（x）对于给定的测试点是相同的，所以使用联合概率或条件概率将产生相同的结果。特征空间的条件概率P（x | i）也很重要，因为它们描述了每个独立类的分布：也就是说，如果移除所有只留下i的类标签，这就是剩下的分布。

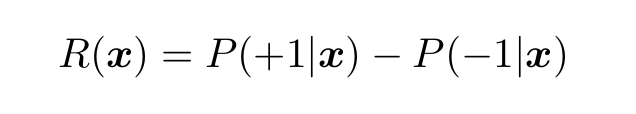
We can use the definition of probability density in (4) to derive one of the oldest and most sophisticated statistical classification techniques by simply removing the limit sign. Consider picking a radius from the test point, x, then counting the number of training samples of one class or another within that distance.  
我们可以使用（4）中的概率密度定义，通过简单地去掉极限符号，来推导最古老和最复杂的统计分类技术之一。考虑从测试点x选取一个半径，然后计算该距离内一个或另一个类别的训练样本数。

The problem with this is that sometimes the enclosed volume will contain no samples, while other times it may contain a great many. So rather than distance, we instead fix the number of samples and implicitly choose the distance on this basis. This is what is known as a k-nearest-neighbors (KNN) classifier, where k is the number of neighboring samples used in each classification.  
这样做的问题是，有时封闭的卷将不包含任何样本，而其他时候它可能包含大量样本。因此，我们不是确定距离，而是确定样本数，并在此基础上隐式地选择距离。这就是所谓的k近邻（KNN）分类器，其中k是每个分类中使用的相邻样本数。

### Binary classifiers 二进制分类器

A binary classifier is special because you can, in many cases, draw a single hyperplane in the feature space that separates the two classes. A hyperplane is a subspace having dimension one less than the embedding dimension. So for a two-dimensional feature space the boundary would be a line, while in three-dimensions, a plane.  
二进制分类器是特殊的，因为在许多情况下，您可以在分隔这两个类的特征空间中绘制一个超平面。超平面是维数小于嵌入维数的子空间。所以对于二维特征空间，边界是一条直线，而在三维空间，边界是一个平面。

Most binary classifiers return not an integer having only two values, but a continuous, decision function. A convenient form of the decision function would be the difference in conditional probabilities:  
大多数二进制分类器返回的不是只有两个值的整数，而是一个连续的决策函数。决策函数的一种方便形式是条件概率的差异：



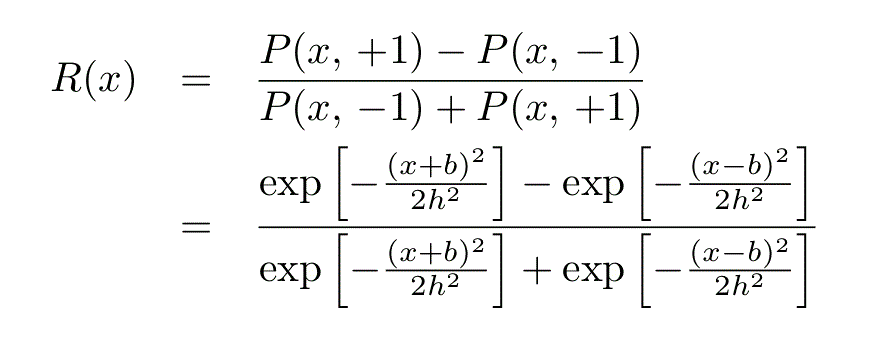
img  
img公司

where, for convenience, we have chosen the class values as -1 and +1.  
为了方便起见，我们选择了类值-1和+1。

Unfortunately, most statistical classifiers do not return a decision function that estimates this quantity well, so a significant chunk of this article will be dedicated towards describing methods of calibrating it so that it does.  
不幸的是，大多数统计分类器并没有返回一个能够很好地估计这个数量的决策函数，因此本文的很大一部分将致力于描述校准它的方法，以便它能够。

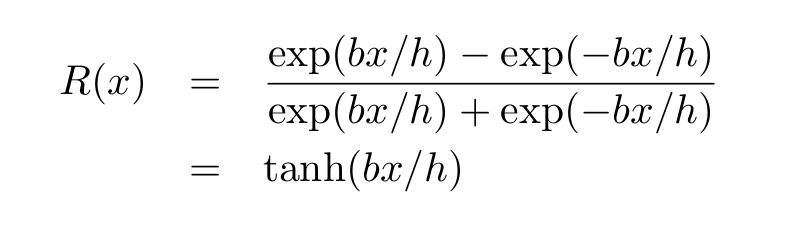
Consider a pair of equally-sized, one-dimensional Gaussian functions of equal width, h, and spaced an equal distance, b, from the origin. The difference in  
考虑一对同样大小的一维高斯函数，其宽度h相等，与原点的距离b相等。不同的是

conditional probabilities is given:  
给出了条件概率：



img  
img公司

which, with some manipulation, works out to:  
通过一些操作，它可以：



img  
img公司

In other words, for a pair of equal-size Gaussians, the decision function in one dimension is a hyperbolic tangent.  
换言之，对于一对等尺寸高斯函数，一维的判定函数是双曲正切。

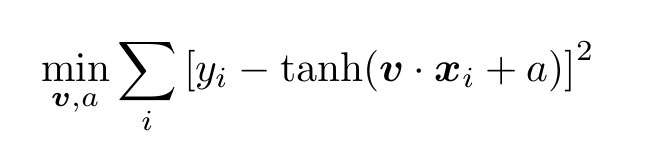
This may seem like a trivial example, however, the tanh function is found throughout the field of machine learning. In statistical classification, it is often used to correct the decision function to better estimate the conditional probabilities.  
这似乎是一个微不足道的例子，然而，tanh函数在整个机器学习领域都可以找到。在统计分类中，为了更好地估计条件概率，通常采用修正决策函数的方法。

This is applied in the library, for instance, as well as in my own machine learning library, . The example illustrates why: the difference in conditional probabilities, R, is, more often than not, sigmoidal close to the class borders.  
例如，这在库中以及在我自己的机器学习库中都有应用。这个例子说明了原因：条件概率R的差异，通常是，靠近类边界的乙状结肠。

Consider logistic regression. In logistic regression, we use the following decision function:  
考虑逻辑回归。在logistic回归中，我们使用以下决策函数：

Where v is a vector and a is a constant.  
其中v是向量，a是常数。

The function parameters are fitting by minimizing a cost function, for instance a least squares:  
函数参数通过最小化成本函数进行拟合，例如最小二乘法：

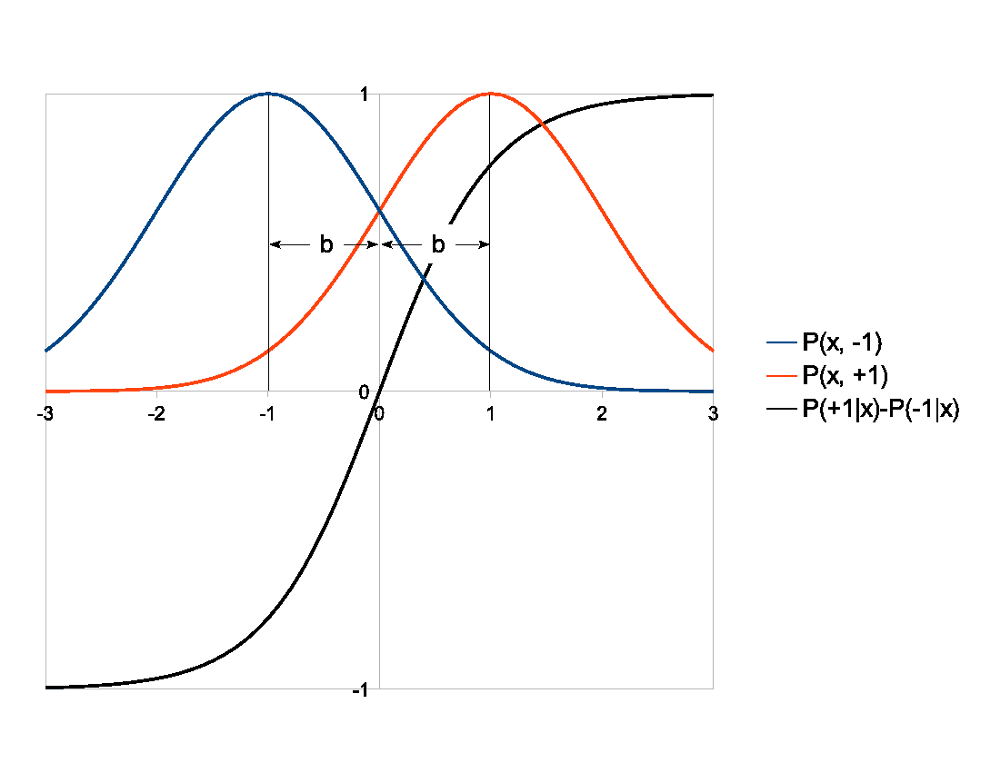


img  
img公司

To fit or “train” the thing, we need some training data. This comprises a set of ordered pairs of a vector in the feature space mapping onto its corresponding class value: {xᵢ : yᵢ}. Here, yᵢ takes on one of two values: -1 or +1, that is, yᵢ ∈ {-1, +1}.  
为了适应或“训练”这个东西，我们需要一些训练数据。这包括特征空间中向量的一组有序对，映射到其对应的类值：{xᵢ：yᵢ}。这里，yᵢ取两个值中的一个-1或+1，即yᵢ∈{-1，+1}。

The training data represents the “ground truth’’ and could be obtained in a variety of ways. Consider a land classification problem: a satellite instrument measures upwelling electromagnetic radiation in several bands and on this basis, we are interested in classifying the surface type, whether field, forest, city, or water, for instance.  
训练数据代表“基本事实”，可以通过多种方式获得。考虑一个土地分类问题：卫星仪器测量多个波段的上升流电磁辐射，在此基础上，我们有兴趣对地表类型进行分类，例如，场、森林、城市或水。

The data could have been painstakingly measured by hand: an aircraft carried a terrestrial version of the instrument aloft and measured radiances, while observers in the craft noted the type of land they were flying over.  
这些数据本可以用手工进行艰苦的测量：一架飞机将这种仪器的地面版本带到高空并测量辐射，而飞机上的观测者则注意到它们飞过的陆地类型。



img  
img公司

It could have been modelled: perhaps we have an algorithm that we trust that returns modelled radiances depending on different parameters describing the land surface. In this case, the resulting training data is potentially infinite, although not necessarily all that accurate.  
它可以被建模：也许我们有一个我们信任的算法，它根据描述陆地表面的不同参数返回建模的辐射。在这种情况下，得到的训练数据可能是无限的，尽管不一定都那么准确。

Or perhaps it was measured by the actual instrument but classified by hand. You have a simple app that brings up an image and each pixel can be classified with a mouse click on the basis of color.  
或者也许它是用实际的仪器测量的，但是用手工分类。你有一个简单的应用程序，它会显示一个图像，每个像素都可以根据颜色用鼠标点击进行分类。

Equations (10) and (11) provide a succinct illustration of the entire process of statistical classification. There is a training phase, given by (11), in which a model is derived. In this case the model is defined by a small set of function parameters which makes this an exercise in parametric statistics.  
式（10）和式（11）简洁地说明了统计分类的整个过程。有一个训练阶段，由（11）给出，在这个阶段中导出一个模型。在这种情况下，模型由一组小的函数参数定义，这使得它成为参数统计中的一个练习。

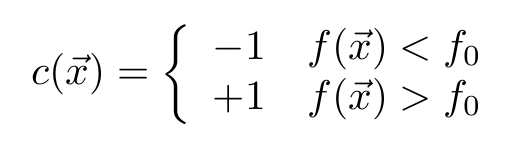
Contrast this with a non-parametric statistical model, such as KNN, which uses all of the training data for each classification. For the logistic classifier, the fitting will be nonlinear, another common technique in machine learning.  
将其与非参数统计模型（如KNN）进行对比，KNN使用每个分类的所有训练数据。对于logistic分类器，拟合将是非线性的，这是机器学习中的另一种常用技术。

Nonlinear optimization is normally performed with an iterative, numerical algorithm, assuming the problem cannot be reduced to a closed-form, analytic solution. It is in itself a broad and diverse field, so we won’t go into the exact details. See the problem set for more info.  
非线性优化通常是用迭代的数值算法进行的，假设问题不能归结为一个封闭的解析解。它本身是一个广泛而多样的领域，所以我们不会详细讨论。有关详细信息，请参见问题集。

The model is then applied to classify a series of test points using Equation (10).  
然后，利用方程（10）对一系列测试点进行分类。

### Calibration 校准

The nice thing about using a continuous decision function for binary classification is that it allows some degree of calibration. Consider the following classifier:  
使用连续决策函数进行二值分类的好处是它允许一定程度的校准。考虑以下分类器：



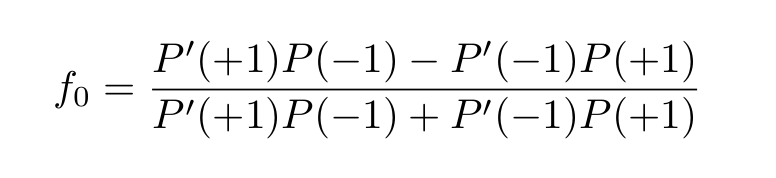
img  
img公司

Varying the classification threshold, f₀, allows us to adjust the sensitivity of the classifier.  
通过改变分类阈值f₀，我们可以调整分类器的灵敏度。

This is particularly useful in medical diagnostics.  
这在医学诊断中特别有用。

Note that the case f=f₀ is left undefined. To remove bias, a random value should be returned when this occurs in numerical computations.  
注意，情况f=f₀未定义。为了消除偏差，当数值计算中出现这种情况时，应返回一个随机值。

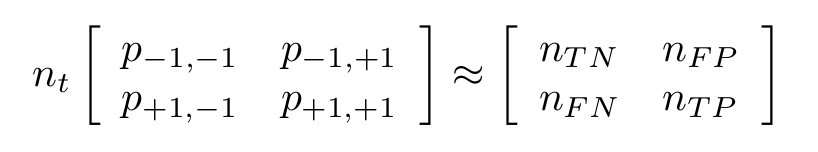
Suppose the classifier is trained using data with a prior class distribution of P’(i) while the population distribution is actually P(i). Assuming that faccurately estimates R, we want to find the value for f₀ such that the sample statistics are corrected to those of the population:  
假设分类器是用先验分布为P’（i）的数据训练的，而种群分布实际上是P（i）。假设faccurally估计R，我们希望找到f₀的值，这样样本统计数据就可以修正为总体统计数据：



img  
img公司

To make this more detailed, consider the confusion matrix. The element of the confusion matrix in the ith row and jth column tells us: for all of the test data, how many test samples had the ith class but the classifier returned the jth class?  
为了使这个更详细，考虑混淆矩阵。第i行和第jth列中的混淆矩阵元素告诉我们：对于所有的测试数据，有多少测试样本具有第i类，但是分类器返回了第jth类？

By dividing by the number of test samples, the confusion matrix can be expressed as an approximate joint probability. Consider the confusion matrix for a binary classifier:  
通过划分测试样本的数量，混淆矩阵可以表示为近似联合概率。考虑二进制分类器的混淆矩阵：

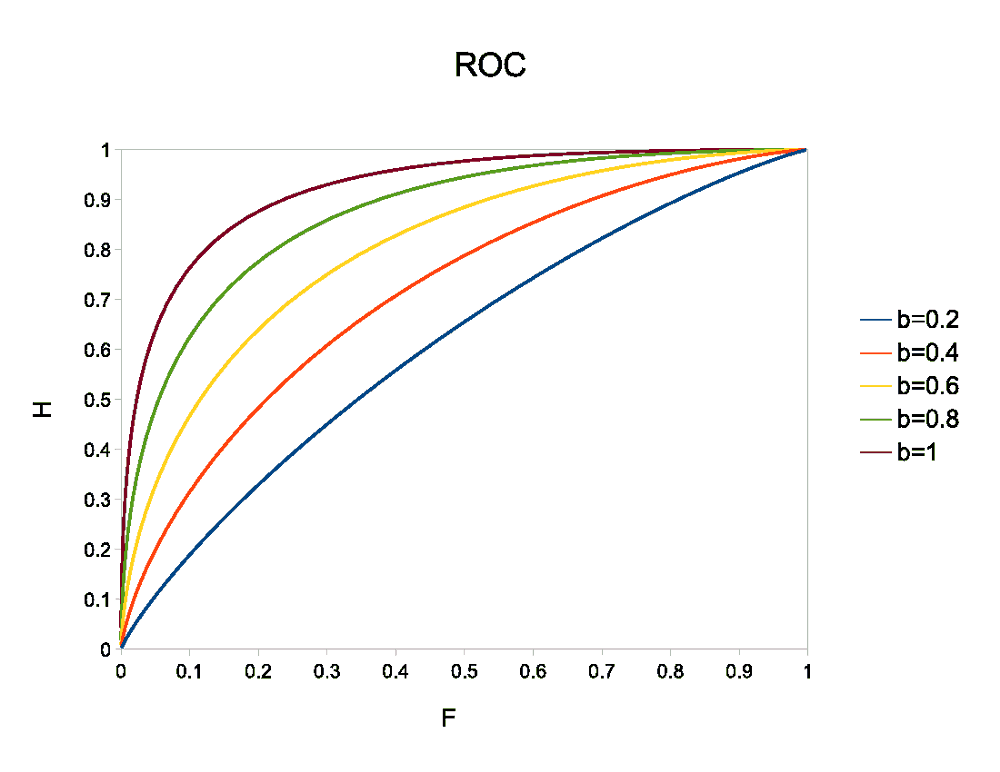


img  
img公司

where:  
哪里：

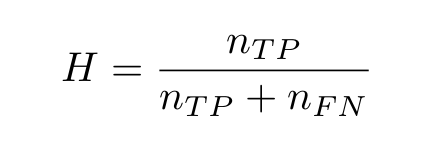
* nt=nTN+nFP+nFN+nTP is the total number of test samples  
  nt=nTN+nFP+nFN+nTP为试样总数
* nTN is the number of true negatives  
  nTN是真负片的数目
* nFP is the number of false positives  
  nFP是误报的数目
* nFN is the number of false negatives  
  nFN是假阴性的数目
* nTP is the number of true positives  
  nTP是真阳性数

A perfect classifier would return a diagonal matrix: an element is non-zero only when i=j.  
一个完美的分类器会返回一个对角矩阵：只有当i=j时元素才是非零的。



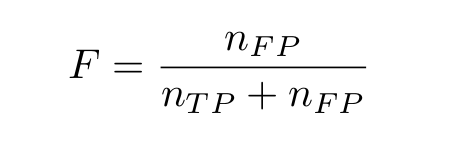
img  
img公司

From these five parameters, you can write down all possible skill scores for a simple binary classifier. The receiver operating characteristic (ROC) curve is produced by plotting two such skill scores against one another while varying the classification threshold. These are the hit rate:  
从这五个参数中，您可以为一个简单的二进制分类器写下所有可能的技能分数。接收机工作特性（ROC）曲线是在改变分类阈值的同时，通过绘制两个这样的技能得分来生成的。以下是命中率：



img  
img公司

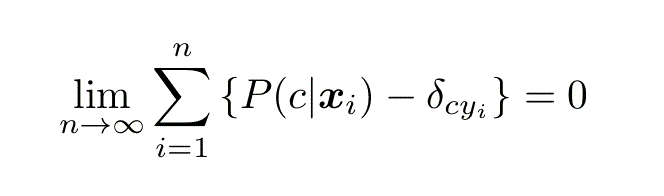
and the false alarm rate:  
错误报警率：



img  
img公司

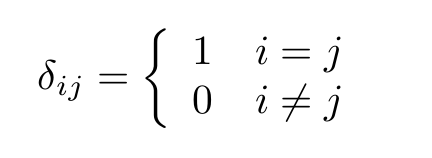
The figure plots the ROC curve for the one-dimensional logistic classifier in (9) for h=1 and for different values of b. The classifier is assumed to be a perfect estimator for the conditional probabilities.  
图绘制了一维logistic分类器在（9）中h=1和b的不同值的ROC曲线。该分类器被认为是条件概率的完美估计。

A more sophisticated calibration exercise would transform the decision function such that it accurately represents the difference in conditional probabilities. Consider the following equation, derived strictly from the background material presented in the first two sections:  
更复杂的校准工作将转换决策函数，使其准确地表示条件概率的差异。请考虑以下方程式，该方程式严格来自前两节中介绍的背景材料：



img  
img公司

Where δ is the Dirac delta function:  
其中δ是狄拉克δ函数：



img  
img公司

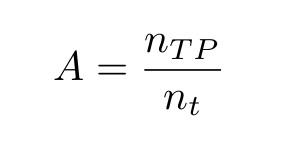
A well-calibrated estimator for the conditional probabilities should obey this equation.  
一个经过良好校准的条件概率估计值应该遵循这个方程。

### Validation 验证

Once we have derived a statistical classifier, we need to validate it on some test data. This data should be different from that used to train the classifier, otherwise skill scores will be unduly optimistic. This is known as cross-validation.  
一旦我们导出了一个统计分类器，我们需要在一些测试数据上验证它。这些数据应该不同于用来训练分类器的数据，否则技能得分会过于乐观。这就是所谓的交叉验证。

The confusion matrix expresses everything about the accuracy of a discrete classifier over a given database and you can use it to compose any possible skill score. Here, we are going to cover two that are rarely seen in the literature, but are nonetheless important for reasons that will become clear.  
混淆矩阵表达了关于给定数据库上离散分类器准确性的所有内容，您可以使用它来组成任何可能的技能得分。在这里，我们将讨论两个在文献中很少出现的问题，尽管如此，这些问题之所以重要，原因将变得很清楚。

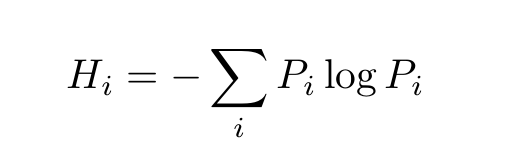
The most basic skill score is accuracy:  
最基本的技能得分是准确性：



img  
img公司

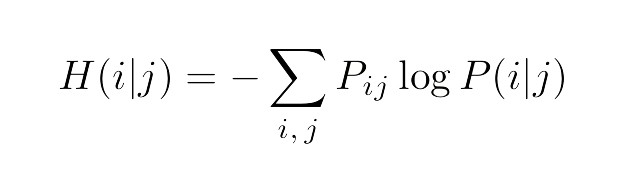
With a maximum-likelihood classification algorithm, accuracy will be maximized. Accuracy, however, has several limitations which can be mitigated by using the following, alternative measures.  
利用最大似然分类算法，将最大化精度。然而，准确性有几个限制，可以通过使用以下替代措施来减轻这些限制。

The first is the uncertainty coefficient. This measure is based on Shannon’s channel capacity and requires, first, a definition of the information entropy. For a discrete probability, this is:  
一是不确定度系数。这种测量基于香农的信道容量，首先需要定义信息熵。对于离散概率，这是：



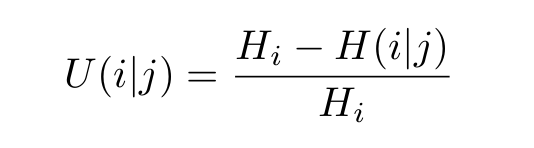
img  
img公司

and tells us how many bits we need to represent i, given that its prior distribution is Pᵢ. The measure can be extended to multivariate distributions. The conditional entropy is given:  
并告诉我们需要多少位来表示i，因为它的先验分布是Pᵢ。该测度可推广到多元分布。条件熵如下：



img  
img公司

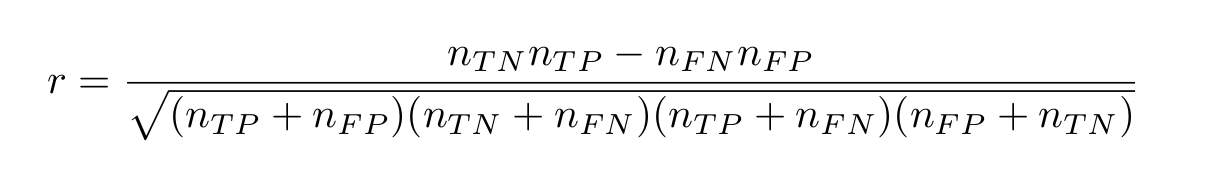
Once we have these two definitions out of the way, we can write down the uncertainty coefficient:  
一旦我们有了这两个定义，我们就可以写下不确定系数：



img  
img公司

which tells us how many bits of information a single classification result in jgives us of the true class value, i. This makes it a good skill score since the lowest possible value is 0, meaning the classifier provides, on average, no information on the true class values, while the highest is 1, meaning the classifier provides full information.  
它告诉我们一个分类中有多少位信息产生了真正的类值，这使得它成为一个很好的技能得分，因为最低的可能值是0，这意味着分类器平均不提供关于真正类值的信息，而最高的可能值是1，这意味着分类器提供了完整的信息。

For binary classifiers, I also recommend the Pearson correlation coefficient:  
对于二进制分类器，我还建议使用皮尔逊相关系数：



img  
img公司

Finally, for binary classifiers that return a continuum decision function rather than a discrete, binary value, we can use the ROC curve to measure the average skill for all possible thresholds by calculating the area under the curve.  
最后，对于返回连续决策函数而不是离散二值的二值分类器，我们可以使用ROC曲线通过计算曲线下的面积来测量所有可能阈值的平均技巧。

For a perfect discriminator, the ROC curve will follow the unit square, rising to H=1 at F=0 and staying there for the duration, thus the area will be 1. An area of 0 is also a perfect classifier, but the sign is reversed, while a classifier with no discrimination value will follow the diagonal with an area of 0.5.  
对于一个完美的鉴别器，ROC曲线将遵循单位平方，在F=0时上升到H=1，并保持在那里一段时间，因此面积将为1。面积为0也是一个完美的分类器，但符号是相反的，而没有判别值的分类器将跟随面积为0.5的对角线。

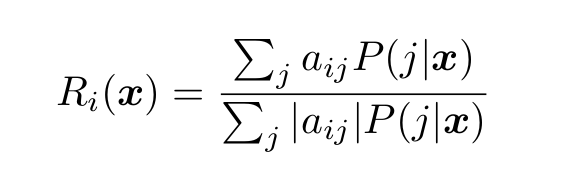
Note for instance, how the area under the example curves gets larger as the separation between the classes increases.  
请注意，例如，示例曲线下的区域如何随着类之间的间隔增加而变大。

### Multi-class classification 多类分类

We have spent a considerable amount of time discussing binary classifiers. Assuming the only suitable statistical classifier we have at our disposal is a binary classifier, how do we generalize it to classification problems with more than two classes, that is, multi-class classifiers? We can use probability theory to derive an answer.  
我们花了大量时间讨论二进制分类器。假设我们能使用的唯一合适的统计分类器是二元分类器，我们如何将其推广到两个以上类的分类问题，即多类分类器？我们可以用概率论来得出答案。

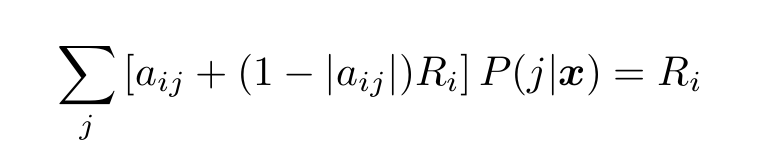
Suppose we design a set of binary classifiers by multiply partitioning the classes into two sets. A coding matrix, A, describes how this partitioning is done: the ith row of the matrix describes the partitioning of the ith binary classifier with a -1/+1 in the jth column, meaning that the jth class label was transformed to a -1/+1 for the training and a 0, meaning it was excluded entirely.  
假设我们设计了一组二进制分类器，将类分成两组。编码矩阵A描述了如何进行分区：矩阵的第i行描述了第i个二进制分类器的分区，在第j列中为-1/+1，这意味着第j个类标签被转换为-1/+1用于训练，为0，这意味着它被完全排除在外。

The conditional probabilities of the multi-class problem are related to those of the binary classifiers as follows:  
多类问题的条件概率与二进制分类器的条件概率有关，如下所示：



img  
img公司

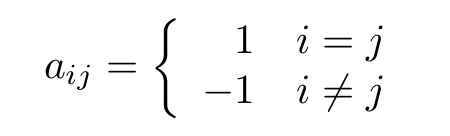
With some rearrangement, we can transform this into a linear system:  
通过重新排列，我们可以将其转换为线性系统：



img  
img公司

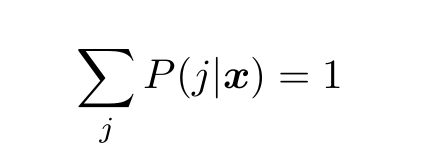
where Ri is the difference in conditional probabilities for the ith binary classifier.  
式中，Ri是第i个二进制分类器的条件概率差。

As an example, consider the “one-versus-the-rest” approach to multi-class classification. Here, we compare each class with all the others. The coding matrix is given (similar to the Dirac delta function):  
例如，考虑多类分类的“一对一”方法。在这里，我们将每一个班级与其他班级进行比较。给出了编码矩阵（类似于Dirac delta函数）：



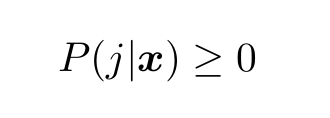
img  
img公司

The preceding assumes that the conditional probabilities for the binary classifiers are estimated correctly. Otherwise we need to constrain the resulting multi-class probabilities. Neglecting the second argument, a conditional probability has the same properties as a univariate probability. First, they all ought to sum to one:  
前面假设二元分类器的条件概率被正确估计。否则我们需要约束得到的多类概率。忽略第二个参数，条件概率具有与单变量概率相同的性质。首先，它们都应该归结为一个：



img  
img公司

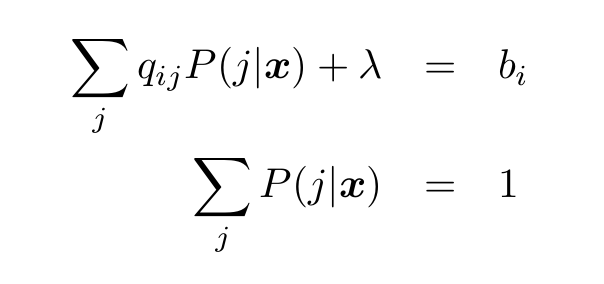
Second, they should all be positive:  
第二，它们都应该是积极的：



img  
img公司

The normalization constraint in (18), being an equality constraint, is the easiest to enforce.  
（18）中的规范化约束是一个等式约束，最容易实施。

One way is to introduce a “slack” variable:  
一种方法是引入一个“slack”变量：



img  
img公司

where Qp = b is the linear system for the unconstrained problem and 𝜆 is the slack variable.  
其中Qp=b是无约束问题的线性系统，而𝜆是松弛变量。

For the “one-versus-one’’ method of multi-class classification, where we compare each class with each of the others in turn, this is all we need. It turns out that once the normalization constraint is enforced, all the others fall into place and the solution has only positive elements.  
对于多类分类的“一对一”方法，我们依次比较每个类和其他类，这就是我们所需要的。结果发现，一旦规范化约束被强制执行，所有其他约束都会就位，解决方案只有积极的元素。

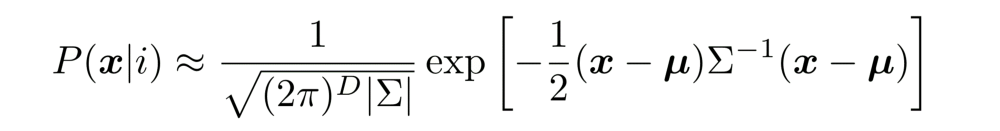
Note that because the system of equations is overdetermined, it will need to be solved as a least-squares problem and there is one other caveat: the normalization must be done separately from the least squares minimization.  
注意，由于方程组是超定的，它将需要作为一个最小二乘问题来求解，还有一个警告：标准化必须与最小二乘最小化分开进行。

In other words, we form the normal equations from (17) first, then plug these into (20). To learn about the normal equations, please see my upcoming article, “Mastering Singular Value Decomposition.”  
换言之，我们先从（17）建立正规方程，然后把它们插入（20）。要了解正规方程，请参阅我即将发表的文章“掌握奇异值分解”

### Problems 问题

This list of problems is provided to help you with Bayesian learning and probability theory and derive useful formulas related to statistical classification. They will also get you thinking about some of the fundamental issues in the field.  
提供此问题列表是为了帮助您学习贝叶斯学习和概率论，并导出与统计分类相关的有用公式。它们也会让你思考这个领域的一些基本问题。

1. Why does the fitting for the logistic classifier in (10) have to be nonlinear? What advantage does this have?  
   为什么（10）中logistic分类器的拟合必须是非线性的？这有什么好处？
2. Do some research online to find nonlinear optimization algorithms to fit the logistic classifier.  
   在线研究寻找适合logistic分类器的非线性优化算法。
3. Derive Equation (12). (It’s surprisingly difficult.) How important do you think it is, on average, to correct for the class distribution? Explain.  
   导出方程（12）。（出人意料的困难）平均来说，你认为纠正班级分布有多重要？解释一下。
4. How would you calculate the ROC curves shown in the figure? Fill in the missing steps going from Equation (8) to (9) and then to calculation of the ROC curves.  
   如何计算图中所示的ROC曲线？填写从方程式（8）到（9）的缺失步骤，然后计算ROC曲线。
5. Derive Equation (13).  
   导出方程（13）。
6. List the advantages of the uncertainty coefficient and the correlation coefficient (for binary classifiers) as a measure of classification skill. What happens when a) the class labels are rearranged and b) the distribution of the class labels is changed in the test data? How does this affect the outcome?  
   列出不确定性系数和相关系数（对于二进制分类器）作为分类技巧的度量的优点。当a）重新排列类标签，b）更改测试数据中类标签的分布时，会发生什么情况？这对结果有何影响？
7. From the general formula for Pearson correlation, derive Equation (15). Note: this is not trivial.  
   从皮尔逊相关的一般公式出发，导出方程（15）。注意：这不是小事。
8. Correlation is normally not appropriate for multi-class classification problems. Why not? What types of problems would be the exceptions?  
   相关性通常不适用于多类分类问题。为什么不？例外情况是什么类型的问题？
9. Derive Equation (17) from Equation (16). Hint: what property of P(i|x̄) do you need to complete the derivation?  
   从方程（16）导出方程（17）。提示：P（i | x̄）的什么性质需要完成推导？
10. The one-versus-the-rest coding matrix in (18) can use a simplified version of Equation (17). Explain.  
    （18）中的一对其余编码矩阵可以使用方程式（17）的简化版本。解释一下。
11. Write down the coding matrix for the one-versus-one multi-class classifier.  
    写下一对一多类分类器的编码矩阵。
12. Find some statistical classification data online or create some on your own, e.g., by classifying pixels in an image. Perform statistical classifications by fitting a multi-dimensional Gaussian to each of the classes:  
    在线查找一些统计分类数据或自行创建一些数据，例如，通过对图像中的像素进行分类。通过将多维高斯拟合到每个类来执行统计分类：



img  
img公司

Where Σ is the covariance matrix, 𝜇 is the arithmetic mean, and D is the dimension of the features data. Measure the accuracy of your results. Don’t forget to divide the data into a test set and a training set.  
其中∑为协方差矩阵，𝜇为算术平均数，D为特征数据的维数。测量你的结果的准确性。别忘了把数据分成测试集和训练集。

### Conclusion 结论

I hope you got the idea of bayesian learning for statistical classification. Mathematical models are not closed systems. They can be expanded, re-purposed, and recombined.  
我希望你有贝叶斯学习统计分类的想法。数学模型不是封闭系统。它们可以被扩展、重新设计和组合。

The applications of probability and Bayes’ theorem and the problems we can put them to are limited only by the imagination. Here we have presented just a few of the ways to use these tools to help gain a foothold in the complex world of computational learning algorithms.  
概率论和贝叶斯定理的应用及其所能解决的问题仅限于想象。在这里，我们只介绍了一些使用这些工具帮助在复杂的计算学习算法世界中站稳脚跟的方法。