

Maximum Mutual Information Criterion Tutorial

Bill McNeill

April 13, 2005

1 Class Separation by Feature

Say you have a group of objects divided into two classes. For a given object, you know the values of a set of features that pertain to it. It may be the case that certain features are more useful than others in distinguishing between the classes. For example, say we have a group of people divided into a set of doctors and non-doctors. About each person we know

1. How many years they spent in school
2. Their annual salary
3. Their height
4. Whether they are left-handed or right-handed

For the purposes of distinguishing doctors from non-doctors, we'd imagine that (1) and (2) would be useful—because doctors spend a long time in school and earn high salaries—but (3) and (4) would not, because height and handedness have nothing to do with being a doctor. We could safely ignore these last two pieces of information. (Though the story for feature (3) might be different if we were trying to identify professional basketball players instead of doctors.)

It would be nice to have a rigorous, data-driven criterion for ranking features according to their usefulness in distinguishing classes. One way to do this is by using mutual information.

2 Maximum Mutual Information Criterion

Say we are given the following.

- A set of classes $G = \{g_1, g_2 \dots g_k\}$
- An m -dimensional finite valued feature space $F_m = V_1 \times V_2 \dots \times V_m$ where the V_i are all finite sets
- A set of n data points of the form $\langle \mathbf{X}, g \rangle$ where $\mathbf{X} \in F$ and $g \in G$

We want to determine the components of \mathbf{X} that are most useful in distinguishing between the different classes. A reasonable metric for the degree to which a particular component can predict a class is the mutual information between the values of that component and the values of the classes in the data.

To calculate this quantity, define two random variables

- X_i , the i th component of \mathbf{X} in a given data point
- C the class of a given data point

Find the mutual information between X_i and C for all the data points.

$$I(X_i, C) = H(C) - H(C|X_i) \tag{1}$$

Since $H(C)$ is the same for all values of i , for the purposes of determining the ordering it is sufficient to calculate $H(C|X_i)$.

$$H(C|X_i) = - \sum_{c \in G} \sum_{x_i \in V_i} p(c, x_i) \log p(c|x_i) \tag{2}$$

Here $p(c, x_i)$ is the joint probability of seeing a data point of class c with i th component x , and $p(c|x_i)$ is the probability of seeing a point of class c given that the i th component is x . The most discriminative component will be the value of (2) that has the largest value, the next most discriminative component will be the value with the second largest value, and so on. This method can be extended to continuous-valued components of \mathbf{X} by changing sums to integrals in (2) as appropriate.

3 A Simple Example

Say that we have two classes, A and B, and the feature vectors are two dimensional, where each value can be either 0 or 1. Then we have

- $G = \{A, B\}$
- $F_2 = V_1 \times V_2$ where $V_1 = V_2 = \{0, 1\}$

Now suppose we have the following data.

Class A	Class B
$\langle 0, 0 \rangle$	$\langle 0, 1 \rangle$
$\langle 1, 0 \rangle$	$\langle 1, 1 \rangle$
$\langle 1, 0 \rangle$	$\langle 1, 1 \rangle$
$\langle 0, 0 \rangle$	$\langle 1, 1 \rangle$
$\langle 0, 1 \rangle$	$\langle 0, 0 \rangle$

From just looking at the data we can tell that X_2 is a more discriminative component than X_1 because the distribution of 0s and 1s is almost equal in the latter, while in the former almost all the 0s are in class A and almost all the 1s are in class B.

The joint and conditional probabilities for each of the components can be calculated directly from the frequencies.

$p(x_0, C)$	A	B	$p(C x_0)$	A	B
0	0.3	0.2	0	0.6	0.4
1	0.2	0.3	1	0.4	0.6
$p(x_1, C)$	A	B	$p(C x_1)$	A	B
0	0.4	0.1	0	0.8	0.2
1	0.1	0.4	1	0.2	0.8

From these we can use (2) to calculate the relative entropy.

$$H(C|X_1) = -0.97 \tag{3}$$

$$H(C|X_2) = -0.72 \tag{4}$$

Since $H(C|X_2) > H(C|X_1)$, the second component is more discriminative.