Expected Information Gain

Firuz Kamalov

Mathematics Department
Canadian University of Dubai
Dubai, UAE

Copyright © 2017 Firuz Kamalov. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Information gain is a widely used metric in feature selection. However, as any other statistic that is based on a random sample, it suffers from the sample variation. In this note, we would like introduce the notion of expected information gain. We analyze and discuss various properties of expected information gain.

Mathematics Subject Classification: 94A15, 62B10

Keywords: mutual information, information gain, feature selection

1 Introduction

Information gain (mutual information) is a popular tool in feature selection [3, 5, 6]. The goal of the feature selection process is to identify the relevant features in data that can be used to classify the target class. Information gain is often used to determine the relevancy of a feature with respect to the target class. One approach to selecting the optimal features is to rank them according to the information gain with respect to the target class and then select the features with the highest information gain [1, 2, 4]. This is an example of a filter approach. Filter approaches rely on various inherent characteristics of features. Thus filter approaches can be widely used in combination with different classifiers.

Information gain is perhaps the most popular metric used in feature ranking. However, as any other metric that is computed based on a random sample
it suffers from the estimation errors. In the process of feature evaluation researchers analyze a sample of the population to reach their conclusions. In particular, information gain is often computed based on sample data from population. Hence, the results of information gain calculations are random in nature. Therefore, we believe that it is appropriate and necessary to study the notion of the expected information gain and its variance.

The notion of expected value of a sample parameter is often employed in statistics to better understand the errors involved in estimating the population parameter based on sample information. For instance, when estimating the population proportion $p$ based on a sample one needs to be mindful of the variations in sample proportion $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Our goal is to extend this approach to the context of information gain. In this note, we introduce the notion of expected information gain between a pair of variables. We analyze various aspects of the expected information gain and highlight possible applications.

In Section 2, we review the notion of information gain in some detail. In Section 3, we define the notion of the expected information gain and variance between a pair of variables. We show that the sample information gain has a nonzero variance.

## 2 Information gain

Let $X$ and $Y$ be a pair of discrete random variables. Information gain between $X$ and $Y$ is a measure of shared information between the two variables and is defined as

$$I(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right).$$

Information gain viewed through Equation 1 represents the Kullback-Leibler divergence between the joint distribution $p(x, y)$ and the product of the marginal distributions. Note that if $X$ and $Y$ are independent variables then $p(x, y) = p(x)p(y)$ for all $x \in X$ and $y \in Y$. It would follow, in this case, that $I(X, Y) = 0$.

We can also view the concept of information gain in terms of entropy as

$$I(X, Y) = H(X) - H(X|Y),$$

where

$$H(X) = \sum_{x \in X} p(x) \log p(x)$$

and

$$H(X|Y) = \sum_{x \in X, y \in Y} p(x, y) \log \left( \frac{p(y)}{p(x, y)} \right).$$
We can see from Equation 2 that $I(X, Y)$ is maximized when $H(X|Y) = 0$. From Equation 3 it follows that $H(X|Y) = 0$ when $p(y) = p(x, y)$ for all $x \in X, y \in Y$.

3 Expected information gain

Feature selection methods are based on the analysis of a sample of the population of features and the target class. The samples used in the analysis are often drawn from a finite population. For instance, samples of DNA are drawn from a finite population of humans. Therefore, the outcome of the feature selection process depends on the specific sample that is used for analysis which is random in nature. In particular, mutual information calculations will depend on the sample. Our goal in this section, is to define the notion the expected mutual between two random variables.

Let $X$ and $Y$ be binary variables each taking on values of 0 and 1. Suppose that $X$ and $Y$ are drawn from a finite population. Let $N_{i,j}$ be number of pairs in the population with $X = i$ and $Y = j$, $i, j = 0$ or 1. Suppose that we draw a sample of pairs $(x_k, y_k)_{k=1}^n$ from the population $X$ and $Y$ where $n$ is the size of the sample. Let $n_{i,j}$ be number of pairs in the sample with $x = i$ and $y = j$, $i, j = 0$ or 1. Let use start by computing the probability of drawing a sample $(n_{0,0}, n_{0,1}, n_{1,0}, n_{1,1})$.

**Lemma 3.1.** Let $X$ and $Y$ be a pair of binary random variables. Suppose that we draw a sample of pairs $(x_i, y_i)_{i=1}^n$ from the population $(X, Y)$. Then, using the above notation, the probability of drawing a sample $(n_{0,0}, n_{0,1}, n_{1,0}, n_{1,1})$ is given by

$$\Pr(A_i, B_j) = \frac{n!}{\prod_{i,j} n_{i,j}!} \frac{(N - \sum_{i,j} n_{i,j})!}{N!} \prod_{i,j} \frac{N_{i,j}!}{(N_{i,j} - n_{i,j})!}. \quad (4)$$

**Proof.** To calculate the probability in Equation 4 we first calculate the total number of ways we can make $n$ selections that would result in the combination $(n_{0,0}, n_{0,1}, n_{1,0}, n_{1,1})$. There are $\binom{n}{n_{0,0}}$ ways to place the combination $(x = 0, y = 0)$. For each arrangement of the combination $(x = 0, y = 0)$ there $\binom{n - n_{0,0}}{n_{0,1}}$ ways to place the combination $(x = 0, y = 1)$. Similarly, for each arrangement of $(x = 0, y = 0)$ and $(x = 0, y = 1)$ there would be $n - n_{0,0} - n_{0,1}$ spots remaining to place the $n_{1,0}$ combinations of $(x = 1, y = 0)$. Since $n_{0,0} + n_{1,0} + n_{0,1} + n_{1,1} = n$ there is only one way to place $n_{1,1}$ combinations of $(x = 1, y = 1)$ in the remaining $n - n_{0,0} - n_{1,0} - n_{0,1}$ spots. It follows that the total number of ways of selecting $\{n_{i,j}\}$ is $\binom{n}{n_{0,0}} \binom{n - n_{0,0}}{n_{0,1}} \binom{n - n_{0,0} - n_{0,1}}{n_{1,0}}$. Furthermore, a simple calculation shows that $\binom{n}{n_{0,0}} \binom{n - n_{0,0}}{n_{0,1}} \binom{n - n_{0,0} - n_{0,1}}{n_{1,0}} = \prod_{i,j} n_{i,j}!$. 
The probability of choosing \(n_{i,j}\) pairs of \((x = i, y = j)\) is
\[
\frac{N_{i,j}}{(N - k_1)} \cdot \frac{N_{i,j} - 1}{(N - k_2)} \cdots \frac{N_{i,j} - n_{i,j} - 1}{(N - k_{n_{i,j}})} = \frac{N_{i,j}!}{n_{i,j}!} \prod_{i,j} \frac{1}{(N - k_i)},
\]
where \(k_i\) is the order of selection. It follows that the probability of selecting a single combination of \(\{n_{i,j}\}\) is \(\frac{(N - \sum_i n_{i,j})!}{N!} \prod_{i,j} \frac{N_{i,j}!}{(N_{i,j} - n_{i,j})!}\).

We can now define the notion of the expected information gain between a pair of variables based on a finite sample.

**Definition 3.2.** Let \(X\) and \(Y\) be a pair of discrete, binary variables each taking a value of either 0 or 1. Let \(S\) be the set of all possible combinations of \(\{n_{i,j}\} = (n_{0,0}, n_{0,1}, n_{1,0}, n_{1,1})\) in a sample of size \(n\). Then we define the expected mutual information between \(X\) and \(Y\) based on a random sample of size \(n\) as
\[
EI(X, Y, n) = \sum_{\{n_{i,j}\} \in S} \Pr(\{n_{i,j}\}) I(Y, X, \{n_{i,j}\}).
\]

We want to show that there exists some \((n'_{0,0}, n'_{0,1}, n'_{1,0}, n'_{1,1}) \in S\) such that \(I(X, Y, (n'_{0,0}, n'_{0,1}, n'_{1,0}, n'_{1,1})) > 0\).

**Theorem 3.3.** Let \(X\) and \(Y\) be a pair of discrete, binary variables. Then \(EI(X, Y, n) > 0\).

**Proof.** Assume, without the loss of generality, that \(n \geq 3\). Then consider the selection \((n'_{0,0} = n - 2, n'_{0,1} = 1, n'_{1,0} = 1, n'_{1,1} = 0)\). Then
\[
I(Y, X) = \frac{n - 2}{n} \log \left[ \frac{n - 2}{n - 1} \frac{n}{n - 1} \frac{n}{n} \right]
+ \frac{1}{n} \log \left[ \frac{1}{n - 1} \frac{1}{n} \frac{1}{n} \right]
+ \frac{1}{n} \log \left[ \frac{1}{n - 1} \frac{1}{n} \frac{1}{n} \right]
= \frac{n - 2}{n} \log \left[ \frac{n(n - 2)}{(n - 1)^2} \right]
+ \frac{2}{n} \log \left[ \frac{n}{n - 1} \right]
= \log \left[ \frac{n(n - 2)}{(n - 1)^2} \left( \frac{n - 1}{n - 2} \right)^{n} \right].
\]

Since \(\frac{n}{n - 1} > \left(\frac{n - 1}{n - 2}\right)^{1 - \frac{2}{n}}\), then
\[
\frac{n(n - 2)}{(n - 1)^2} \left( \frac{n - 1}{n - 2} \right)^{\frac{2}{n}} > 1.
\]
It follows that $I(X, Y, (n'_{0,0} = n - 2, n'_{0,1} = 1, n'_{1,0} = 1, n'_{1,1} = 0)) > 0$. It is known that $I(X, Y, \{n_{i,j}\}) \geq 0$ is always nonnegative. Therefore, $EI(X, Y, n)$ is the average of nonnegative values with at least one value being strictly positive. It follows that $EI(X, Y, n) > 0$. □

The above theorem shows that even if $X$ and $Y$ are independent variables the average mutual information between the variables is greater than zero. In the context of feature selection this implies that adding a new feature will always increase the information gain of the feature subset even if the new feature is unrelated to the target class. Therefore, we must take into account the cost of adding a new feature to the final feature subset.

Our next result shows that the sample expected information gain converges to the population information gain as the size of the sample increases. This is result shows that our definition has the correct properties expected from a such a statistic.

**Theorem 3.4.** Let $X$ and $Y$ be binary variables as above. Let $N$ the size of the population of pairs $(X, Y)$ and $n$ be the size of the sample taken from the population $(X, Y)$. Then $EI(X, Y, n) \to I(X, Y)$ as $n \to N$.

**Proof.** Recall that $EI(X, Y, n)$ is calculated as a weighted average over all possible combinations of $\{n_{i,j}\} = (n_{0,0}, n_{0,1}, n_{1,0}, n_{1,1})$ in a sample of size $n$. Consider the sample $\{n'_{i,j}\}$, where $n'_{i,j} = n \cdot \frac{N_{i,j}}{N}$ for all $i, j$. If $n \to N$, then $n'_{i,j} \to N_{i,j}$ for each $i, j$. It follows from Lemma 1 that $Pr(\{n'_{i,j}\}) \to 1$. Therefore, $Pr(\{n_{i,j}\}) \to 0$ for every $\{n_{i,j}\} \neq \{n'_{i,j}\}$. Furthermore, it is easy to check that $I(X, Y, \{n'_{i,j}\}) \to I(X, Y)$ as $n'_{i,j} \to N_{i,j}$. It follows that

$$\sum_{\{n_{i,j}\} \in S} Pr(\{n_{i,j}\})I(Y, X, \{n_{i,j}\}) \to Pr(\{n'_{i,j}\})I(Y, X, \{n'_{i,j}\}) \to I(X, Y)$$

as $n \to N$. □

**References**


Received: February 27, 2017; Published: September 21, 2017