Feature Selection with Mutual Information

Clayton W. Seitz

February 21, 2022

Outline

Feature Selection

What is it?

A special type of dimensionality reduction where we select a subset of features, in contrast with *feature extraction*

Why do we do it?

- Quality of the input data is just as important as the algorithm you choose
- The volume of a feature space grows exponentially in the number of dimensions n
- But we often have a small number of samples $p << n$

Mutual Information

Mutual information comes from information theory and statistics.

$$
\begin{aligned}
I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y)) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

where H denotes the entropy

- It quantifies the amount of information one variable carries about another
- Captures nonlinear correlation and is not limited to continuous variables
- Y could be categorical e.g., cellular phenotypes

An Example

Using Mutual Information for Feature Selection

$$X^* = \underset{X}{\operatorname{argmax}} \; I(\boldsymbol{X}; Y)$$

For phenotyping, we might want to find the optimal X which is most informative about the cell type Y

This is an optimization problem (NP-hard) on maximizing the *joint mutual information*

There are approximate solutions

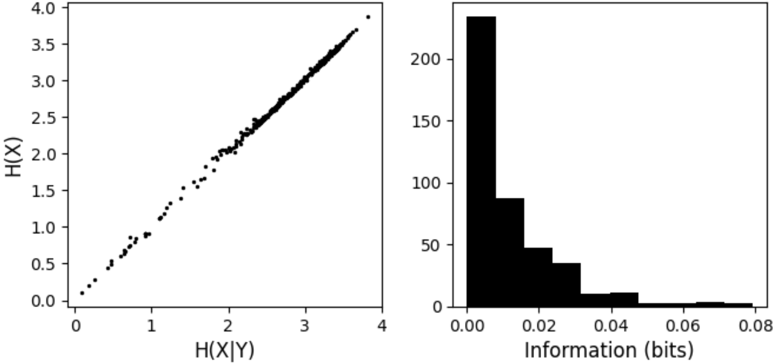
Maximum Relevancy Minimum Redundancy (MRMR)

By making some approximations, we can rewrite $I(\boldsymbol{X}; Y)$ as

$$I(\boldsymbol{X}; Y) \approx \sum_i \left(I(X_i; Y) - \alpha \sum_j I(X_i; X_j) \right)$$

where the parameter α determines how strongly we consider redundancy

Results for T1D dataset



Algorithm Details

The chain-rule for mutual information tells us that

$$I(\boldsymbol{X};Y) = \sum_i I(X_i;Y|\boldsymbol{X}_{\setminus i}) \tag{1}$$

To simplify notation let $Z = \boldsymbol{X}_{\setminus i}$. The chain rule for info can also be used to show that

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$$

Solving for $I(X;Y|Z)$ says we can rewrite (1) as

$$\begin{aligned} I(\boldsymbol{X};Y) &= \sum_i I(X_i;Y|Z) \\ &= \sum_i I(X_i;Y,Z) - I(X_i;Z) \end{aligned}$$

Algorithm Details

Applying the chain rule one more time gives

$$\begin{aligned} I(\boldsymbol{X}; Y) &= \sum_i I(X_i; Y, Z) - I(X_i; Z) \\ &= \sum_i I(X_i; Y) - I(X_i; Z) + I(X_i; Z|Y) \end{aligned}$$

We maximize the sum by maximizing the each term s_i

$$\begin{aligned} s_i &= I(X_i; Y) - I(X_i; Z) + I(X_i; Z|Y) \\ &\approx I(X_i; Y) - \alpha \sum_j I(X_i; X_j) + \beta \sum_k I(X_i; X_k|Y) \end{aligned}$$

Setting $\beta = 0$ gives the so-called maximum relevancy minimum redundancy (MRMR) features

Algorithm Details

$$s_i \approx I(X_i; Y) - \alpha \sum_j I(X_i; X_j)$$

Algorithm 1 Pseudocode for Greedy MRMR

```
$features = \{\}$
for $i = 1$ to $N$ do
    if $i = 1$ then
        add $x_i$ to features
    else
        if $s_i > s_{i-1}$ then
            add $x_i$ to features
        end if
    end if
end for
```

References I