Information-Theoretic Feature Selection

By: Behzad Asadi

Dimensionality Reduction

Feature Construction

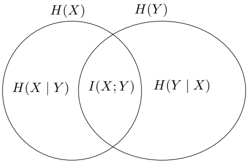
- Principle Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Autoencoders (Neural Nets)

Feature Selection

- Wrappers: classifier dependent
- Embedded Methods: classifier dependent
- Filter Methods: classifier independent

Information-theoretic feature selection is a filter method.

Information-Theoretic Measures



Entropy

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

Conditional Entropy

$$H(X \mid Y) = \sum_{y \in \mathcal{Y}} p(y) H(X \mid y) = -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log p(x \mid y)$$

Relative Entropy (Kullback-Leibler divergence)

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Mutual Information

$$I(X;Y) = H(X) - H(X \mid Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Fano's Inequality

Fano's Inequality: Considering the classification problem where $X=\{X_1,X_2,\ldots,X_m\}$ is the set of features, Y is the true class label, and $\hat{Y}$ is the estimated class label, we have the Markov chain $Y \to X \to \hat{Y}$. Therefore, using Fano's inequality, we have

$$H(Y \mid \hat{Y}) \leq H(P_e) + P_e \log(|\mathcal{Y}| - 1)$$

where $P_e = P(Y \neq \hat{Y})$, and $H(P_e) = -P_e \log(P_e) - (1 - P_e) \log(1 - P_e)$.

Proof: First, an error random variable is defined as

$$E = \begin{cases} 1 & \text{if } \hat{Y} \neq Y, \\ 0 & \text{if } \hat{Y} = Y. \end{cases}$$

Then, $H(E, Y \mid \hat{Y})$ is expanded in two different ways

$$\begin{aligned} H(E, Y \mid \hat{Y}) &= H(Y \mid \hat{Y}) + \underbrace{H(E \mid Y, \hat{Y})}_{=0} \\ &= \underbrace{H(E \mid \hat{Y})}_{\leq H(E) = H(P_e)} + H(Y \mid E, \hat{Y}) \end{aligned}$$

where $H(Y \mid E, \hat{Y}) = (1 - P_e) \underbrace{H(Y \mid E = 0, \hat{Y})}_{=0} + P_e \underbrace{H(Y \mid E = 1, \hat{Y})}_{\leq \log(|\mathcal{Y}| - 1)}$

Cont'd

As a result, we have

$$H(Y \mid \hat{Y}) \leq H(P_e) + P_e \log(|\mathcal{Y}| - 1). \qquad \square$$

Therefore, we can write

$$H(Y) - I(Y;\hat{Y}) = H(Y \mid \hat{Y}) \leq H(P_e) + P_e \log(|\mathcal{Y}| - 1) \leq 1 + P_e \log(|\mathcal{Y}| - 1)$$

$$\frac{H(Y) - I(Y;\hat{Y}) - 1}{\log(|\mathcal{Y}| - 1)} \leq P_e$$

Using the Markov chain $Y \to X \to \hat{Y}$ and the data processing inequality, we have $I(Y;\hat{Y}) \leq I(X;Y)$, therefore

$$\frac{H(Y) - I(X;Y) - 1}{\log(|\mathcal{Y}| - 1)} \leq P_e$$

Feature Selection: Information-Theoretic Metric

The last equation shows that in order to minimize the lower bound on P_e, we need to maximize $I(X_1, X_2, \ldots, X_m; Y)$.

Using this observation, the objective of feature selection is defined as follows

Objective: The objective is to find a subset of features with minimum cardinality which preserves the mutual information between all the features and the class label. In the feature selection process, we try to get rid of the features which are either irrelevant or relevant but redundant in the context of others.

Sequential Search

Forward Selection:

$$\begin{aligned}
X_i' &= \underset{X_k \in X \setminus S_i}{\arg\max}\, I(\{X_k, S_i\}; Y) \\
&= \underset{X_k \in X \setminus S_i}{\arg\max}\, \left(I(S_i; Y) + I(X_k; Y \mid S_i)\right) \\
&= \underset{X_k \in X \setminus S_i}{\arg\max}\, I(X_k; Y \mid S_i) \\
S_{i+1} &\leftarrow S_i \cup \{X_i'\}
\end{aligned}$$

Backward Elimination:

$$\begin{aligned}
X_i' &= \underset{X_k \in S_i}{\arg\max} I(S_i \setminus \{X_k\}; Y) \\
&= \underset{X_k \in S_i}{\arg\max} (I(S_i; Y) - I(X_k; Y \mid S_i \setminus \{X_k\})) \\
&= \underset{X_k \in S_i}{\arg\min} I(X_k; Y \mid S_i \setminus \{X_k\}) \\
S_{i+1} &\leftarrow S_i \setminus \{X_i'\}
\end{aligned}$$

Forward Selection is computationally less expensive than backward elimination.

Mutual Information Maximization (MIM) Algorithm

MIM Selection Criterion:

$$J_{\text{MIM}}(X_k) = I(X_k; Y)$$

This naive algorithm selects the first K features with the largest $I(X_k; Y)$.

Drawbacks:

By re-writing the forward-selection criterion as follows

$$\begin{aligned} I(X_k; Y \mid S_i) &= I(X_k; S_i, Y) - I(X_k; S_i) \\ &= I(X_k; Y) - I(X_k; S_i) + I(X_k; S_i \mid Y), \end{aligned} \tag{1}$$

where the first term is a measure of relevance, the second term is a measure of redundancy, and the last term is measure of conditional redundancy. We can see that the last two terms are missing in the MIM criterion.

Mutual Information Feature Selection (MIFS) Algorithm

MIFS Selection Criterion:

$$J_{\text{MIFS}}(X_k) = I(X_k; Y) - \beta \sum_{X_j \in S_i} I(X_k; X_j),$$

where β is a parameter which determines the penalty concerning the dependency of the feature with already selected ones.

Drawbacks:

Considering equation (1), we can see that the last term is missing in the MIFS criterion.

minimum Redundancy Maximum Relevance (mRMR) Algorithm

mRMR Selection Criterion:

$$J_{\mathsf{mRMR}}(X_k) = I(X_k;Y) - \frac{1}{|S_i|}\sum_{X_j \in S_i} I(X_k;X_j)$$

Drawbacks:

Similar to MIFS, there is a term missing in mRMR criterion.

[1] H. Peng, F. Long, and C. Ding “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy"

Joint Mutual Information (JMI) Algorithm

JMI Selection Criterion:

$$J_{\text{JMI}}(X_k) = \sum_{X_j \in S_i} I(X_k, X_j; Y)$$

This criterion can be rewritten as $\sum_{X_j \in S_i} I(X_j; Y) + \sum_{X_j \in S_i} I(X_k, Y \mid X_j)$. As the first part is constant for all X_k, we can write the selection criterion as follows

$$\begin{aligned}
J_{\text{JMI}}(X_k) &= \sum_{X_j \in S_i} I(X_k; Y \mid X_j) \\
&= \sum_{X_j \in S_i} \left(I(X_k; Y, X_j) - I(X_k; X_j)\right) \\
&= \sum_{X_j \in S_i} \left(I(X_k; Y) + I(X_k; X_j \mid Y) - I(X_k; X_j)\right) \\
&= |S_i| I(X_k; Y) - \sum_{X_j \in S_i} \left(I(X_k; X_j) - I(X_k; X_j \mid Y)\right) \\
&\propto I(X_k; Y) - \frac{1}{|S_i|} \sum_{X_j \in S_i} \left(I(X_k; X_j) - I(X_k; X_j \mid Y)\right)
\end{aligned}$$

A Unified Framework

All the criteria presented in the previous slides can be unified using the following single criterion

Selection Criterion:

$$J(X_k) = I(X_k; Y) - \beta \sum_{X_j \in S_i} I(X_k; X_j) + \gamma \sum_{X_j \in S_i} I(X_k; X_j \mid Y)$$

[1] G. Brown, A. Pocock, M. J. Zhao, and M. Luján, “Conditional likelihood maximisation: A unifying framework for information theoretic feature selection"

References

1. Leonidas Lefakis, Information Theory and Feature Selection (Joint Informativeness and Tractability), Zalando Research Labs
2. G. Brown, A. Pocock, M. J. Zhao, and M. Luján, “Conditional likelihood maximization: A unifying framework for information theoretic feature selection"