Birla Institute of Applied Sciences
बिरला इंस्टिट्यूट ऑफ़ अप्लाइड साइंसेस
Bhimtal, Distt: Nainital, Uttarakhand- 263136

Pattern Recognition

- S. S. Samant

Feature Extraction and Feature Selection

- Feature extraction

- Feature selection

Feature Extraction and Feature Selection

- Feature extraction
  - Process of determining the features to be used for learning.
  - Usually involves arithmetic operation or computing function

- Feature selection
  - Discard some features
  - Use only a subset of features

Feature Extraction and Feature Selection

- Importance:
  - Using discriminating features enhances performance
  - Avoids unnecessary computation
  - With less features, a smaller dataset is needed
  - Intended similarity can be captured by distances in a smaller set of features

Types of Feature Selection

- Filter methods
- Wrapper methods
- Embedded methods

Types of Feature Selection

- Filter methods: scores each feature
- Wrapper methods: scores subsets of features on a validation set
- Embedded methods: selects features during training itself

Filter methods of Feature Selection

- Consider a dataset with d-dimensional patterns. Feature selection is the task of selecting k features where $1 \leq k < d$.
- So if we have a feature set $\{f_1, f_2, \ldots, f_d\}$, this entails finding the score of each feature f_i. This score represents the degree of importance of the feature.
- The similarity between any two features f_i and f_j is then found to remove the redundancy in the selected features.

Mutual Information (MI) for Feature Selection

- Information gain based
- Measures dependency between features and classes
- MI between term t and class l measures how much information the presence or absence of a term contributes to making the correct classification decision on the class l

Mutual Information (MI) for Feature Selection

- Information gain based
- Measures dependency between features and classes
- MI between term t and class l measures how much information the presence or absence of a term contributes to making the correct classification decision on the class l

$$MI = P(u_t, u_l)\log_2\frac{P(u_t, u_l)}{P(u_t)P(u_l)} + P(\bar{u}_t, u_l)\log_2\frac{P(\bar{u}_t, u_l)}{P(\bar{u}_t)P(u_l)}$$
$$+ P(u_t, \bar{u}_l)\log_2\frac{P(u_t, \bar{u}_l)}{P(u_t)P(\bar{u}_l)} + P(\bar{u}_t, \bar{u}_l)\log_2\frac{p(\bar{u}_t, \bar{u}_l)}{P(\bar{u}_t)P(\bar{u}_l)}$$

where
u_t means that the document contains the term t; and
$\bar{u}_t$ means the document does not contain the term t;
u_l means the document is in class l and;
$\bar{u}_l$ means the document is not in class l.

Mutual Information (MI) for Feature Selection

- Information gain based
- Measures dependency between features and classes
- MI between term t and class l measures how much information the presence or absence of a term contributes to making the correct classification decision on the class l

$$MI = P(u_t, u_l)\log_2\frac{P(u_t, u_l)}{P(u_t)P(u_l)} + P(\bar{u}_t, u_l)\log_2\frac{P(\bar{u}_t, u_l)}{P(\bar{u}_t)P(u_l)}$$
$$+ P(u_t, \bar{u}_l)\log_2\frac{P(u_t, \bar{u}_l)}{P(u_t)P(\bar{u}_l)} + P(\bar{u}_t, \bar{u}_l)\log_2\frac{p(\bar{u}_t, \bar{u}_l)}{P(\bar{u}_t)P(\bar{u}_l)}$$

$$MI = \frac{N_{u_t u_l}}{N}\log_2\frac{NN_{u_t u_l}}{(N_{u_t\bar{u}_l} + N_{u_t u_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{N_{\bar{u}_t u_l}}{N}\log_2\frac{NN_{\bar{u}_t u_l}}{((N_{\bar{u}_t u_l} + N_{\bar{u}_t\bar{u}_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{NN_{u_t\bar{u}_l}}{N}\log_2\frac{NN_{u_t\bar{u}_l}}{(N_{u_t u_l} + N_{u_t\bar{u}_l})(N_{u_t\bar{u}_l} + N_{\bar{u}_t\bar{u}_l})}$$
$$+ \frac{N_{\bar{u}_t\bar{u}_l}}{N}\log_2\frac{NN_{\bar{u}_t\bar{u}_l}}{(N_{\bar{u}_t u_l} + N_{\bar{u}_t\bar{u}_l})(N_{u_t\bar{u}_l} + N_{\bar{u}_t\bar{u}_l})},$$

Mutual Information (MI) for Feature Selection

- Information gain based
- Measures dependency between features and classes
- MI between term t and class l measures how much information the presence or absence of a term contributes to making the correct classification decision on the class l

$$MI = P(u_t, u_l)\log_2\frac{P(u_t, u_l)}{P(u_t)P(u_l)} + P(\bar{u}_t, u_l)\log_2\frac{P(\bar{u}_t, u_l)}{P(\bar{u}_t)P(u_l)}$$
$$+ P(u_t, \bar{u}_l)\log_2\frac{P(u_t, \bar{u}_l)}{P(u_t)P(\bar{u}_l)} + P(\bar{u}_t, \bar{u}_l)\log_2\frac{p(\bar{u}_t, \bar{u}_l)}{P(\bar{u}_t)P(\bar{u}_l)}$$

$$MI = \frac{N_{u_t u_l}}{N}\log_2\frac{NN_{u_t u_l}}{(N_{u_t\bar{u}_l} + N_{u_t u_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{N_{\bar{u}_t u_l}}{N}\log_2\frac{NN_{\bar{u}_t u_l}}{((N_{\bar{u}_t u_l} + N_{\bar{u}_t\bar{u}_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{NN_{u_t\bar{u}_l}}{N}\log_2\frac{NN_{u_t\bar{u}_l}}{(N_{u_t u_l} + N_{u_t\bar{u}_l})(N_{u_t\bar{u}_l} + N_{\bar{u}_t\bar{u}_l})}$$
$$+ \frac{N_{\bar{u}_t\bar{u}_l}}{N}\log_2\frac{NN_{\bar{u}_t\bar{u}_l}}{(N_{\bar{u}_t u_l} + N_{\bar{u}_t\bar{u}_l})(N_{u_t\bar{u}_l} + N_{\bar{u}_t\bar{u}_l})},$$

A filter is used to discard features with low MI

- Backward filter – discards features if MI is below a threshold
- Forward filter - keeps feature if MI is above a threshold

Problem: Mutual Information

$$MI = \frac{N_{u_t u_l}}{N} \log_2 \frac{NN_{u_t u_l}}{(N_{u_t \bar{u}_l} + N_{u_t u_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{N_{\bar{u}_t u_l}}{N} \log_2 \frac{NN_{\bar{u}_t u_l}}{((N_{\bar{u}_t u_l} + N_{\bar{u}_t \bar{u}_l})(N_{u_t u_l} + N_{\bar{u}_t u_l})}$$
$$+ \frac{NN_{u_t \bar{u}_l}}{N} \log_2 \frac{NN_{u_t \bar{u}_l}}{(N_{u_t u_l} + N_{u_t \bar{u}_l})(N_{u_t \bar{u}_l} + N_{\bar{u}_t \bar{u}_l})}$$
$$+ \frac{N_{\bar{u}_t \bar{u}_l}}{N} \log_2 \frac{NN_{\bar{u}_t \bar{u}_l}}{(N_{\bar{u}_t u_l} + N_{\bar{u}_t \bar{u}_l})(N_{u_t \bar{u}_l} + N_{\bar{u}_t \bar{u}_l})},$$

Problem: In Reuters corpus, if term is *export* and class is *poultry*. Let's call non-*export* term as *other* term and non-*poultry* class as other class.
The term *export* is present in 49 documents of class poultry and in 27652 documents of other class. There are 141 other terms in documents of poultry class and 774106 other terms in documents of other classes.
Compute MI.

Chi-square Statistic for Feature Selection

- Used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies.

Thank You!