NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W10: 10/20/15-10/22/15

NC STATE UNIVERSITY

Admin: Changes in grading

- Please send me an email before 10/23/15, if you want to keep 20% weightage to midterm-1.

NC STATE UNIVERSITY

Today

- Information theory, entropy
- Readings
  - Chapter 2 from : [DM] David MacKay. Information Theory, Inference, and Learning Algorithms. (http://www.inference.phy.cam.ac.uk/itprnn/book.html)

10/24/15 © Raju Vatsavai CSC-591. 3

NC STATE UNIVERSITY

Entropy, MI, …

- Entropy $H[X] = -\sum_x P(X=x)\log_2 P(X=x)$
- Conditional Entropy: $H[C \mid X=x] = -\sum_c P(C=c \mid X=x)\log_2 P(C=c \mid X=x)$
- Mutual Information (Expected Information):

$$I[C;X] = H[C] - H[C \mid X] = H[C] - \sum_x P(X=x)\log_2 H[C \mid X=x]$$

$$I[C;X] = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right); I[C;X] = \int_Y\int_X p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

- Joint Entropy: $H[X,Y] \equiv -\sum_{x,y} P(X=x,Y=y)\log_2 P(X=x,Y=y)$
- $I[C;X_1,X_2,...X_k] = H[C] + H[X_1,X_2,...X_k] - H[C,X_1,X_2,...X_k] = H[C] - H[C \mid X_1,X_2,...X_k]$

10/24/15 © Raju Vatsavai CSC-591. 4

Algorithm to find most informational attribute

- Calculate the entropy of the training set, *T*, using the percentages, p_+ and p_-, of the positive and negative examples:

  $$H(T) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

- For each attribute, a, that divides T into subsets, T_i, with relative sizes P_i, do the following:
  - (i) calculate the entropy of each subset, T_i
  - (ii) calculate the average entropy: $H(T, a) = \Sigma_i P_i H(T_i)$
  - (iii) calculate information gain: $I(T, a) = H(T) – H(T, a)$
- Choose the attribute with the highest value of information gain.

Example

Example	crust size	shape	filling size	Class
*e*1	big	circle	small	**pos**
*e*2	small	circle	small	**pos**
*e*3	big	square	small	**neg**
*e*4	big	triangle	small	**neg**
*e*5	big	square	big	**pos**
*e*6	small	square	small	**neg**
*e*7	small	square	big	**pos**
*e*8	big	circle	big	**pos**

- $H(T) = -p_+ \log_2 p_+ - p_- \log_2 p_- = -(5/8)\log(5/8) – (3/8) \log (3/8) = 0.945$
- Now calculate entropies of subsets defined by attribute (a=shape) (and repeat for all attributes).
  - $H(\text{shape} = \text{square}) = -(2/4) \log (2/4) – (2/4) \log (2/4) = 1$
  - $H(\text{shape} = \text{circle}) = -(3/3) \log (3/3) – (0/3) \log(0/3) = 0$
  - $H(\text{shape} = \text{triangle}) = -(0/1) \log(0/1) - (1/1) \log (1/1) = 0$
- From these, we obtain the average entropy of the system where the class labels and the value of attribute shape is known as
  - $H(T, \text{shape}) = (4/8) \times 1 + (3/8) \times 0 + (1/8) \times 0 = 0.5$

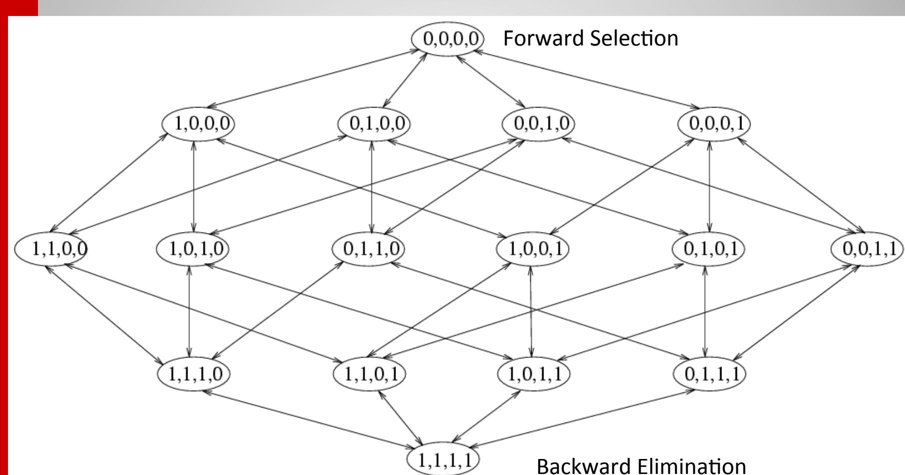
NC STATE UNIVERSITY

Example

- Repeating the process for each attribute, we get
  - H(T, curst-size) = 0.951; H(T, filling-size) = 0.607
- Now compute information gains:
  - I(T, shape) = H(T) – H(T, shape) = 0.954-0.5 = 0.454
  - I(T, curst-size) = 0.954 – 0.951 = 0.003
  - I(T, filling-size) = 0.954 – 0.607 = 0.347
- Therefore, maximum information is contributed by the **shape** attribute

NC STATE UNIVERSITY

Searching for Feature Subsets



NC STATE UNIVERSITY

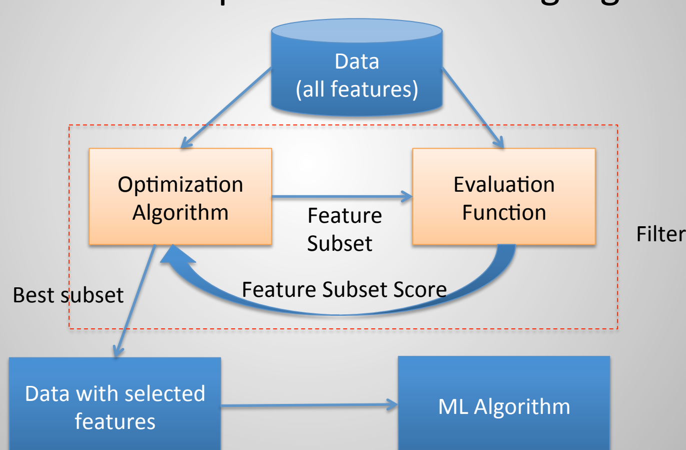
Feature Subset Selection

- Simple Filters
  – Assume features are independent
- Filters
  – Evaluation function is independent of learning algorithm
- Wrappers
  – Evaluation using the machine learning algorithm
- Embedded approaches
  – Feature selection during learning

NC STATE UNIVERSITY

Filtering

- Evaluation independent of learning algorithm



NC STATE UNIVERSITY

Filtering Approaches

- Distribution based*
- Idea: Select minimal subset of features whose probability distribution is close to original distribution, i.e., P(C|F_subset) ~ P(C|F_all)
- Algorithm
  - Start will all features
  - Optimization: Use backward elimination to eliminate predefined number of features
  - Evaluation: the next feature to be eliminated is obtained using cross-entropy measure

*D. Koller and M. Sahami: Towards optimal feature selection, ICML-1996

10/24/15 © Raju Vatsavai CSC-591. 11

NC STATE UNIVERSITY

Filtering Approaches

- Distribution based*
- Cross Entropy: If p and q are two distributions, then cross entropy of p to q is given by

$$D(p,q) = \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)}$$

  - q is approximation of p
  - p is also called right distribution (our desired distribution)
- Search space is exponential in number of attributes
- Use the idea of conditional independence
  - Two sets of variables A, B are conditionally independent given a variable X, if P(A=a|X=x, B=b) = P(A=a|X=x).
  - Intuitively removing a feature that is almost independent will not increase the distance between the desired distribution and new distribution with subset.

10/24/15 © Raju Vatsavai CSC-591. 12

NC STATE UNIVERSITY

Filtering: FOCUS Algorithm

- FOCUS: Almallim and Dietterich: Efficient Algorithms for Identifying Relevant Features, AAAI-1992.
- Evaluation
  - In a selected subset
  - Count conflicts in class value (two example with same feature value but different class labels)
- Search
  - All promising subsets of same size are evaluated until a sufficient (no conflict) subset is found
- Improved approaches
  - Using heuristics to avoid evaluating all subsets

NC STATE UNIVERSITY

Filtering: FOCUS; Example

F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

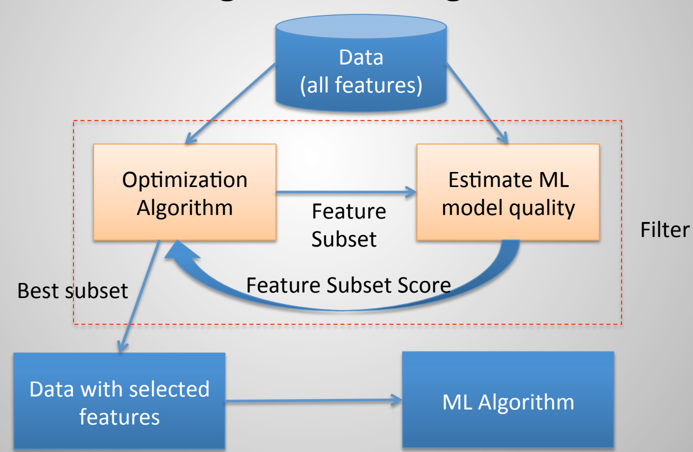
F_4	F_5	C
0	1	0
0	1	1
0	1	1
0	1	1
1	0	0
1	0	1
1	0	1
1	0	1

$C=\{0_1,1_2\}$ Conflict

$C=\{0_1,1_2\}$ Conflict

Wrapper Based Approaches

- Evaluation using same ML algorithm



Wrappers: Instance-based learning

- Evaluation (instance-based learning)
  - Select subset of features
  - Estimate (ML) model quality using cross validation
- Search
  - Start with rand feature subset
  - Use beam search with backward elimination [1]

  or
  - Use random mutation [2]

[1] Aha and Bankert: Feature Selection for case-based classification of cloud-types. AAAI Technical Report WS-94-01.

[2] DB Shalak: Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms.

NC STATE UNIVERSITY

Play with Weka

- Weka
  - http://www.cs.waikato.ac.nz/ml/weka/
- Data
  - UCI Machine Learning Repository
    - https://archive.ics.uci.edu/ml/datasets.html

NC STATE UNIVERSITY

Acknowledgements

- D. Mladenić