Transformations

CSCI 347, Data Mining

6 Methods

1. Attribute selection
2. Attribute discretization
3. Projections
4. Sampling
5. Dirty data
6. Transforming multiple classes to binary ones

Attribute Selection

Two ways to select best attributes:

- Filter method - select subset of attributes based on general characteristic of the data
- Wrapper method – evaluate the subset using the machine learning algorithm

Filter Attribute Selection

- Select the smallest set of attributes that distinguish all instances uniquely
- Given m attributes there would be 2^m possible subsets
- Search space forward, adding attributes, or backwards, subtracting

Some say: Statistically unwarranted and can lead to overfitting

Principle Component Analysis (PCA)

Principle Component Analysis – identifying the important "directions" in the data

- In the hyperspace, determine which axis gives the greatest variance in the data
- Rotate the coordinate system into the directions that seem to be the most important.
- Typically reduce the number of axis also
- Find the direction of greatest variance that is perpendicular to this direction and repeat

Algorithm for PCA

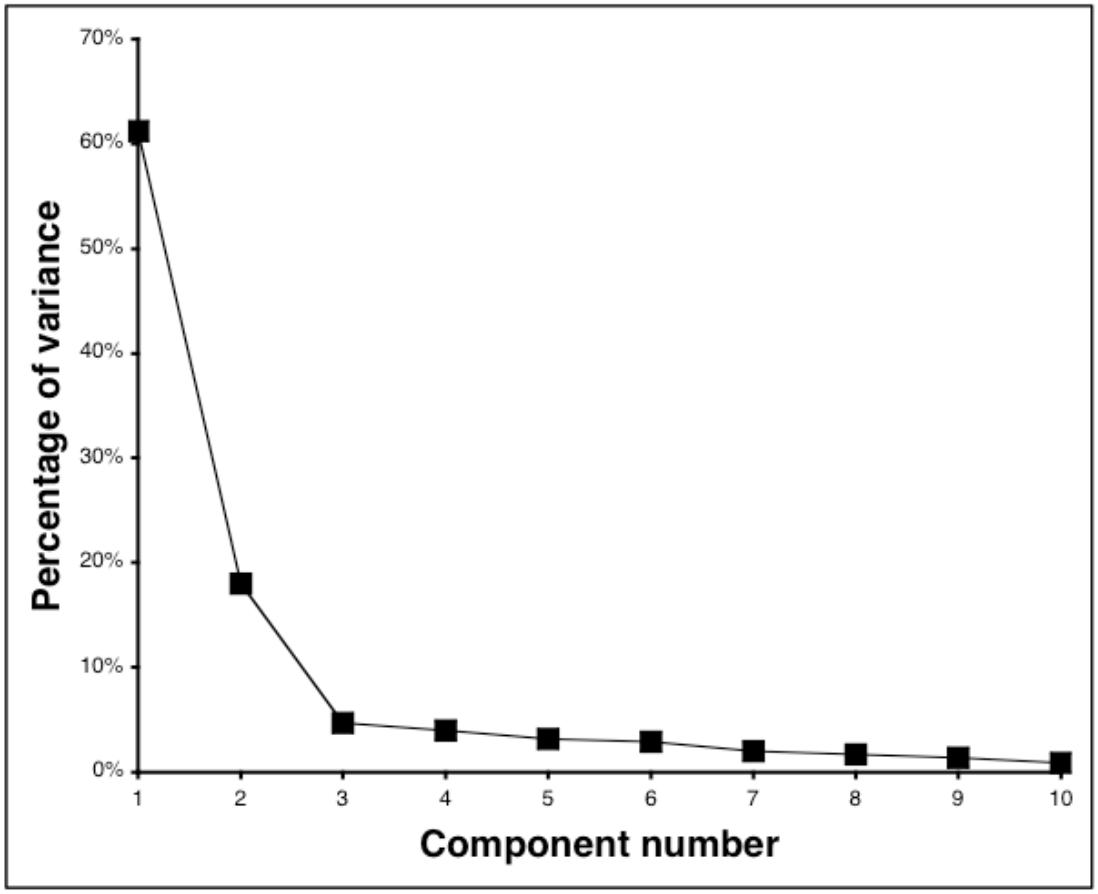
Algorithm:

1. Find direction (axis) of greatest variance
2. Find direction of greatest variance that is perpendicular to previous direction and repeat

Implementation: find eigenvectors of covariance matrix by diagonalization (Eigenvectors , sorted by eigenvalues),are the directions)

Example: 10-Dimensional Data

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%



Wrapper Attribute Selection

- Apply decision tree algorithm to full dataset and use those attributes that are used in the tree
- Use 1R repeatedly
- Use an algorithm that builds a linear model, and rank the attributes based on their weights
- Do learning method with randomly chosen attributes and see which learns best

Discretization

- Unsupervised
  - Equal interval binning
  - Equal frequency binning
- Supervised
  - Entropy based
    - Build decision tree using MDL principle as stopping criteria

Weather Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot			
Overcast	Hot			
Rainy	Mild			
Rainy	Cool			
Rainy	Cool			
…	…			

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	70	96	False	Yes
Rainy	68	80	False	Yes
Rainy	65	70	True	No
…	…	…	…	…

```
If outlook = sunny an
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above th
```

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = no
If none of the above then play = yes
```

Example

Split on temperature attribute:

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	**No**	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Yes**	**Yes**	**No**	**Yes**	**Yes**	**No**

E.g. temperature < 71.5: yes/4, no/2
temperature ≥ 71.5: yes/5, no/3

Info([4,2],[5,3])
= 6/14 info([4,2]) + 8/14 info([5,3])
= 0.939 bits

- Place split points halfway between values
- Can evaluate all split points in one pass!

Projections

Simple transformations can often make a large difference in performance

Example transformations:

- Difference of two data attributes
- Ratio of two numeric (ratio-scale) attributes
- Encode cluster membership
- Adding noise to the data
- Removing data randomly or selectively obfuscating the data