# MODULE 6
# Different Approaches to Feature Selection

## LESSON 9
### Branch and Bound Schemes

Keywords: Curse of Dimensionality, Peaking Phenomenon, Relaxed Branch and Bound

**Feature Selection**

- The features which are chosen to represent a set of patterns may be large or redundant in some way. In such cases, it may be good to reduce the features. This process is called feature selection or dimensionality reduction.
- If the dimensionality is large, it is necessary to have a large training set to get a good classification accuracy.
- This is due to the **curse of dimensionality** or **peaking phenomenon**.

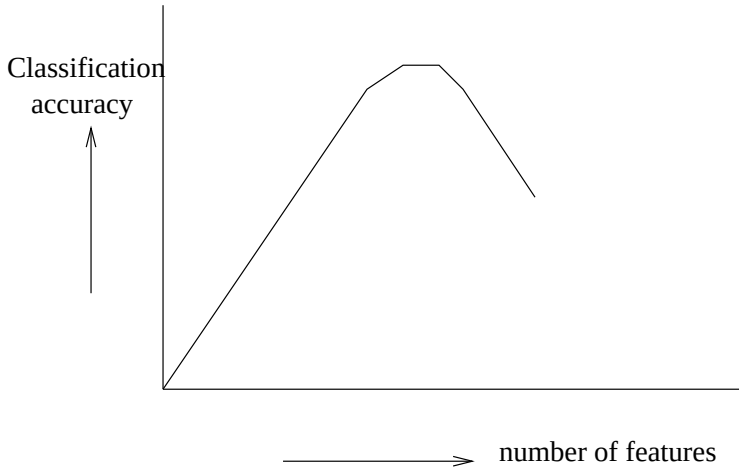


Figure 1: Illustration of curse of dimensionality

- This is illustrated in Figure 1. For a particular number of training patterns, up to a certain point, the increase in dimensionality leads to improvement in the classification accuracy. But after that, there is a reduction in the classification accuracy.
- Besides, reducing the dimensionality, reduces the time required to find the class of a test pattern when using a classifier.
- This may lead to a reduction in the classification accuracy. If it is not very much, then it might be worth doing feature selection which will lead to reduction in time and space requirement of classifiers.

- If the features removed are not discriminatory features, it ensures that the classification accuracy does not decrease.

**Methods of Feature Selection**

- All the methods of feature selection involve searching for the best subset of features.
- Generally after removing a feature or adding a feature, the resulting set of features is evaluated and the decision is taken as to whether to keep that feature or not.
- The evaluation function is called the criterion function J.
- The evaluation usually employed is the classification accuracy obtained on a validation set of patterns.
- This can also be the classification error. If E is the % error, then the criterion function J = 100 - E.
- It could also be based on classification accuracy and on the number of features used which are combined together in some way. If two subsets have the same classification accuracy then the subset which has a smaller number of features has a larger evaluation than the one with the larger number of features.

**Exhaustive search**

- In this method, all combinations of features are tried out and the criterion function J calculated.
- The combination of features which gives the highest value of J is the set of features selected.
- If the number of features is large, this method becomes prohibitively time consuming.
- If the number of features is d and the number of required features is k, then the number of different combinations to be tried out to find the best set of features will be $\begin{pmatrix} d \\ k \end{pmatrix}$.

- For every combination of features X, the criterion function J(X) has to be calculated.
- The combination of features which gives the best criterion value is chosen as the feature subset selected.
- This method is a brute force method and hence gives the optimal solution.
- Because of the time complexity for large feature sets, this method is impractical when the number of features goes up.

**Branch and Bound Search**

- The branch and bound method forms a tree where the root pertains to choosing all features.
- The children of this node pertains to the combinations of features where one feature has been removed. From each of these children, we get the nodes where another feature has been removed and so on.
- If there are $d$ features and it is necessary to retain $f$ features, then the height of the tree will be $d - f$.
- A leaf node represents a combination of features.
- Once a leaf node is reached, it can be evaluated and its criterion function $J$ found.
- This value is stored as the bound $b$.
- While evaluating the future branches, at anytime if the criterion value goes below the bound $b$, that branch is not further expanded.
- When another leaf node is reached, if its $J$ value is larger than $b$, $b$ is updated and that combination of features is stored as the best so far.
- The assumption made here is monotonicity. In the tree generated, it is assumed that the parent node has a higher criterion function value as compared to its children. This means that a feature set has a larger $J$ value as compared to any of its subsets.

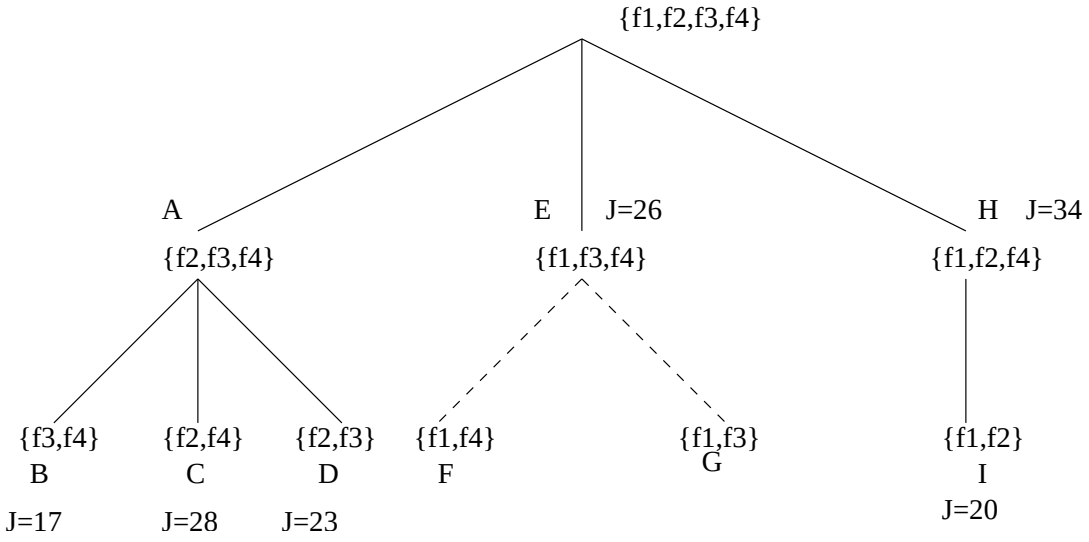


Figure 2: The solution tree for branch and bound with d=4 and f=2

- The tree generated by the branch and bound algorithm for a particular problem is shown in Figure 2.
- When the left-most node B is reached, which corresponds to the feature subset {f3,f4}, let the evaluation of the node $J$ be 17.
- At this stage, since this is a leaf node, the bound $b = 17$.
- The next node generated is C having a $J$ value of 28.
- Since this is greater than the value of $b$, the bound $b$ is updated to 28 and the best subset so far is {f2,f4}.
- The next node generated is D corresponding to the feature subset {f2,f3}.
- This has a $J$ value of 23 which is smaller than $b$, so $b$ remains unchanged.
- The next node generated is E which is found to have a $J$ value of 26.
- Since the $J$ value of this node is less than $b$, this branch is not expanded any further.

- This means that the branches F and G are not generated.
- The next node to be generated is H which has a J value of 34.
- Since 34 is greater than $b$, this node is expanded to give I.
- The $J$ value of I is 20 which is lower than the bound.
- Now the entire tree is completed and the node with the lower bound $b$ as the criterion function is the best subset which is {f2,f4}. This is based on the fact that we are selecting 2 out of 4 features.

**Relaxed Branch and Bound**

- In the Relaxed Branch and Bound, the principle of monotonicity is used but is relaxed.
- A margin $m$ is used and if the $J$ value of any node being considered becomes less than the bound $b$ by a value less $m$, it is still considered for expansion and one of its children can become the optimal subset.
- Consider Figure 2. Let the margin $m = 5$.
- When node E is being considered, the bound $b = 28$. At E, $J = 26$, which is lower than $b$. This is not expanded further, in the branch and bound method. But in the relaxed branch and bound, since the difference between its criterion 26 and the bound 28 is less than the margin, it is expanded further. Now if F has a criterion of 29, then the subset chosen would be {f1,f4} instead of {f2,f4}.
- In this method, the principle of monotonicity is relaxed.

**Selecting Best Individual Features**

- In this method, each individual feature is evaluated individually.
- If $f$ features are to be chosen from $d$ features, after evaluation, the best $f$ features are chosen.
- This method is the simplest method for finding a subset of features and is computationally fast since different combinations of features need not be evaluated.

- However, the subset of features chosen is not likely to give good results since the dependencies between features is not taken into account.
- It is necessary to see how the combination of features perform.
- Consider the example where there are 6 features. It is necessary to choose two features out of the six features. After evaluation of individual features if the evaluation is as follows :

$f_1 = 45$ $\quad$ $f_2 = 56$
$f_3 = 60$ $\quad$ $f_4 = 22$
$f_5 = 65$ $\quad$ $f_6 = 34$

The two features giving the highest criterion function value are features $f_5$ and $f_3$. Therefore the subset to be chosen is $\{f_3, f_5\}$ ,