BITS Pilani

BITS Pilani
Hyderabad Campus

Prof.Aruna Malapati
Department of CSIS

innovate achieve lead

BITS Pilani
Hyderabad Campus

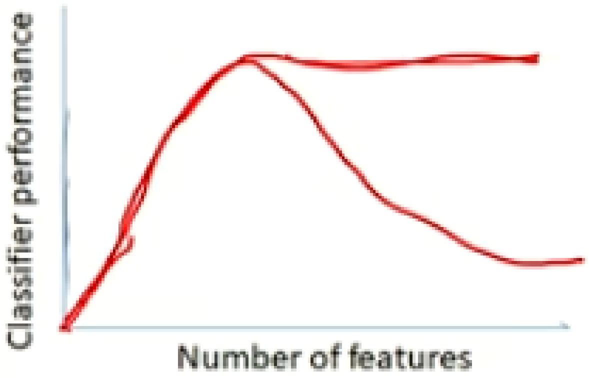
Feature selection

Today's Learning objective

- Curse of dimensionality
- Euclidean distance vs. cosine similarity
- Importance of Feature Selection
- Feature Selection Approaches
- Filter Methods
- Types of filters
- Chi-Squared test

Curse of Dimensionality

- As dimensionality increases the number of data points required for a classification model also increase exponentially.
- Hughes Phenomenon: For a fixed number of training samples(N) in the data set the performance of the models decreases as dimensionality increase.

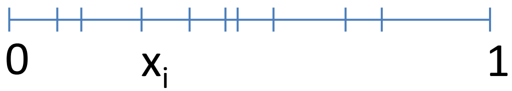


Reasons for this phenomenon:

- ✓ Redundant Features – Carry same data in some other form
- ✓ Correlation between features – the presence of one feature influence the other.
- ✓ Irrelevant Features - those that are simply unnecessary

Curse of Dimensionality (Contd..)

- ✓ The intuitions of distances in 3D are invalid in higher dimensions.
- ✓ For example, consider a data point x_i from N samples in 1D



$\text{dist}_{min}(x_i)=\min_{xi \neq xj}\{\text{dist}(x_i,x_j)\}$ The minimum of distance between x_i and x_j such that $x_i \neq x_j$

$\text{dist}_{max}(x_i)=\max_{xi \neq xj}\{\text{dist}(x_i,x_j)\}$ The maximum of distance between x_i and x_j such that $x_i \neq x_j$

$$\lim_{d \to \infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} = 0$$

- ✓ $\text{dist}_{max}(x_i) \approx \text{dist}_{min}(x_i)$ that means every pair of points are approximately at the same distance from each other.

➢ Distance measures become meaningless in higher dimensions.

Euclidean distance VS Cosine similarity

- Euclidean distance in high dimensionality does not make a sense solution for this is using cosine similarity for high dimensional spaces.
- The impact of dimensionality on cosine similarity is lower than the Euclidean distance.
- If the data is dense, then its impact will be high, and if it is sparse, then the impact will be lower. That means in sparse, most of the values are 0, so the data is non-uniformly spread.

Feature Subset Selection
Dimension Reduction

Importance of Feature Selection

- The objective of feature selection is three-fold:

  ✓ Improving the prediction performance of the models

  ✓ Reduction in the training time required to build model

  ✓ Providing a better understanding of the underlying process that generated the data

What is Feature Selection for classification?

- Given: A set of predictors ("features") $F=\{f_1,f_2,f_3\ldots f_D\}$ and target class label T.
- Find: Minimum subset $F'=\{f_1',f_2',f_3'\ldots f_M'\}$ that achieves maximum classification performance where $F' \subseteq F$.

- Feature subset selection
  - ✓ Given D initial set of features
  - ✓ There are 2^D possible subsets.
  - ✓ Need a criteria to decide which subset is the best:
    - ✓ Classifier based on these m features has the lowest probability of error of all such classifiers.
  - ✓ Evaluating 2^D possible subsets is time consuming and expensive.
  - ✓ Use heuristics to reduce the search space.

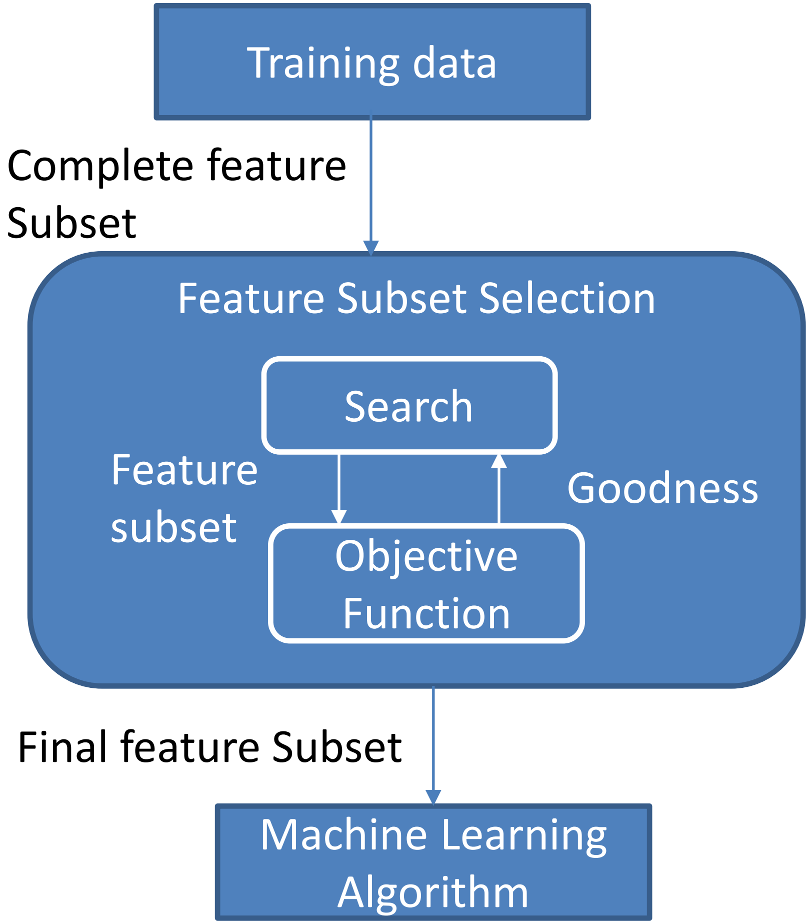
Feature Selection approaches

Three approaches to evaluate 2^D possible subsets

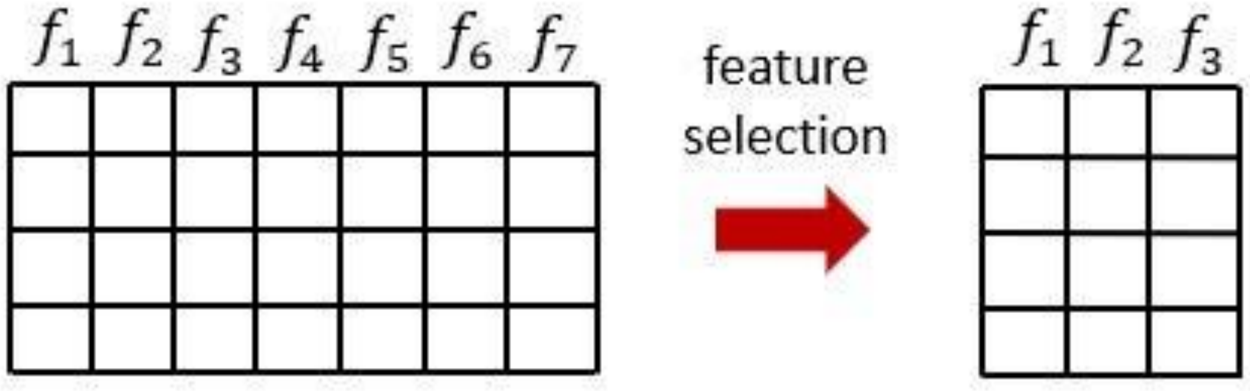
- ➢Unsupervised (Filter Methods)
  - ✓Use only features/predictor variables
  - ✓Select the features that have the most information
- ➢Supervised: Wrapper Methods
  - ✓Train using the selected subset
  - ✓ Estimate error on the validation set
- ➢Embedded Methods
  - ✓ Feature selection is done while training the model

Steps in Feature Selection

- Feature selection is an optimization problem having the following steps:
- Step1: Search the space of all possible features
- Step2: Pick the optimal subset using an objective function



Effect of Feature Subset selection



Filter Methods

- The Predictive power of individual features is evaluated.
- Rank each feature according to some univariate metric and select the highest ranking features.
- The score should reflect the discriminative power of each feature.

Input: large feature set Ω

1 Identify candidate subset S ⊆ Ω

2 While !stop criterion()

Evaluate utility function J using S.

Adapt S

3 Return S.

Pros: fast, provides generically useful feature set

Cons: cause higher error than wrappers

Types of Filters

- **Univariate filters** evaluate each feature independently with respect to the target variable.
  - ✓ Correlation
  - ✓ Fisher Score
  - ✓ Mutual Information (Information Gain)
  - ✓ Gini index
  - ✓ Gain Ratio
  - ✓ Chi-Squared test

Filter Methods

Set of all Features → **Selecting the Best Subset** → **Learning Algorithm** → **Performance**

- Features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.

Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

Types of filters

- Correlation-based
  - Pearson product-moment correlation
  - Spearman rank correlation
  - Kendall concordance
- Information-theoretic metrics
  - Mutual Information (Information Gain)
  - Gain Ratio
- Statistical/probabilistic independence metrics
  - Chi-square statistic
  - F-statistic
  - Welch's statistic
- Others
  - Fisher score
  - Gini index
  - Cramer's V

How “useful” is a single feature? : Univariate filters

Trying to predict someone’s Data Mining exam grade from various possible indicators (a.k.a. features):

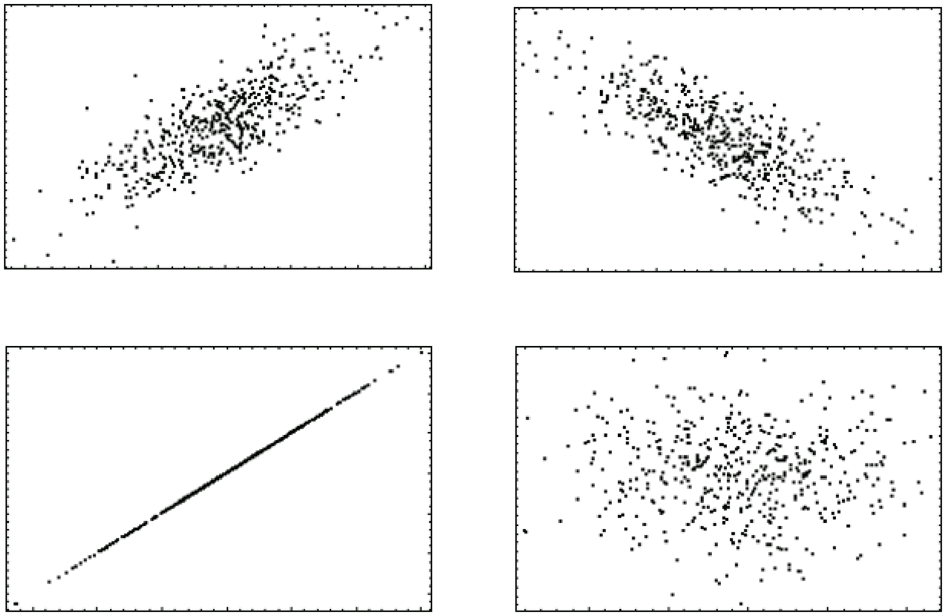
1) Statistics grade,

2) Biology grade,

3) Linear Algebra grade, or

4) Height ...

Which one would you pick?



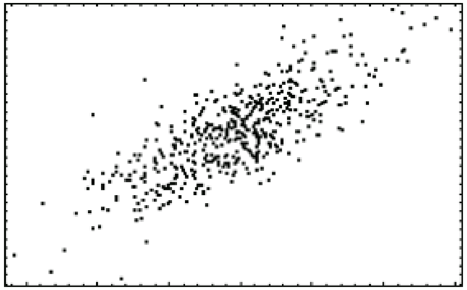
Pearson's Correlation Coefficient

- Used to measure the strength of association between two continuous random variables.

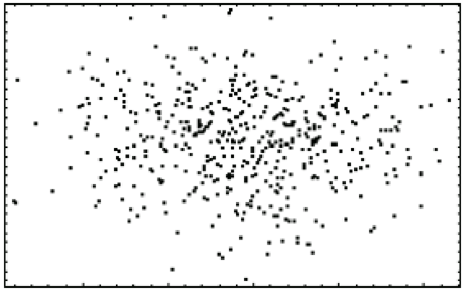
$$\text{Feature}: \mathbf{x}_k = \{x_k^{(1)}, ..., x_k^{(N)}\}^T$$

$$\text{Target}: \mathbf{y} = \{y^{(1)}, ..., y^{(N)}\}^T$$

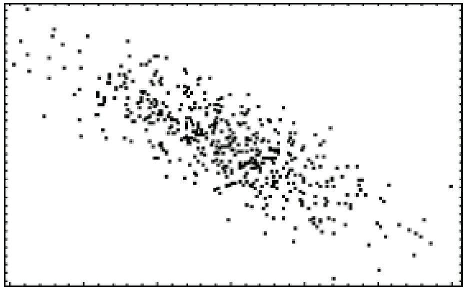
$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{N}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x^{(i)} - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y^{(i)} - \bar{y})^2}}$$



$r = +0.5$



$r = 0.0$



$r = -0.5$

Both positive and negative correlation is useful!

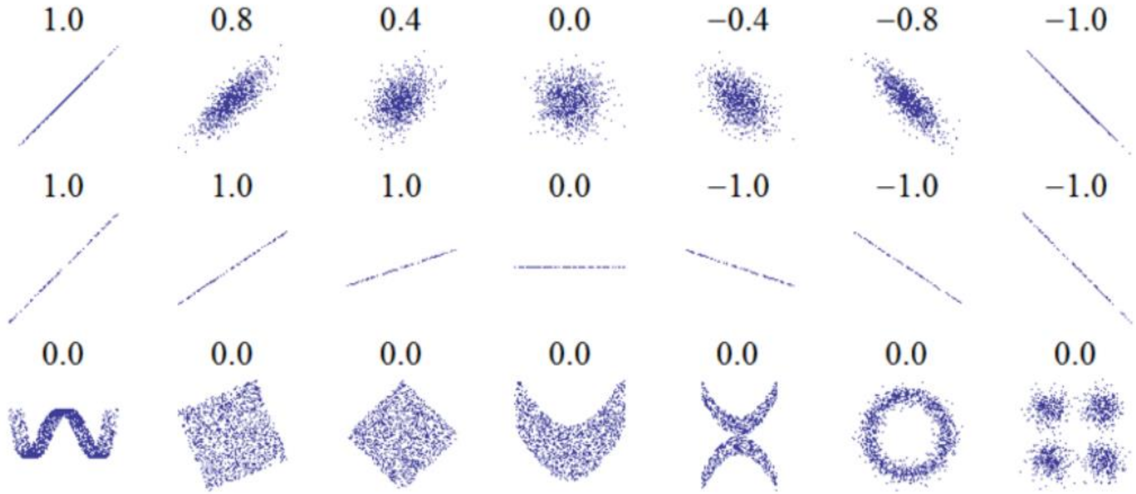
Ranking with Filter Criteria

- Rank features X_i , $\forall i$ by their values of $J(X_k)$.
- Retain the highest ranked features, discard the lowest ranked.

Cut-off point decided by user, e.g. |S| = 5,
S = {35, 42, 10, 654, 22}.

k	$J(X_k)$
35	0.846
42	0.811
10	0.810
654	0.611
22	0.443
59	0.388
...	...
212	0.09
39	0.05

Limitation: Pearson assumes all features are INDEPENDENT !
and... only identifies LINEAR correlations.



There are LOTS of ranking criteria...

- Pearson, Fisher, Mutual Info, Jeffreys-Matsusita, Gini Index, AUC, F-measure, Kolmogorov distance, Chi-squared, CFS, Alpha-divergence, Symmetrical Uncertainty,.... etc, etc

➢ How do I pick the right filter ? Unfortunately, quite complex.... depends on:

  - ✓ type of variables/targets (continuous, discrete, categorical).
  - ✓ class distribution
  - ✓ degree of nonlinearity/feature interaction

➢ The **"No Free Lunch"** theorem states that there is no universal model that works best for every problem.

Hypothesis Testing

- Hypothesis is a premise or claim that we want to investigate.
- Test whether the two random variables (categorical) are independent or not.
- Test Statistic
  - ✓Chi-Squared Test
  - ✓T-Test
  - ✓ANNOVA-Test

Test Statistic

Example

A group of customers were classified in terms of personality (introvert, extrovert or normal) and in terms of color preference (red, yellow or green) with the purpose of seeing whether there is an association (relationship) between personality and color preference.

Data was collected from 400 customers and presented in the 3 (rows) x 3 (cols) contingency table below:

(Observed counts)	Colors			
Personality	**Red**	**Yellow**	**Green**	**Totals**
Introvert personality	11	5	1	17
Extrovert personality	8	6	8	22
Normal	3	10	12	25
Total	22	21	21	64

Five-step approach for Chi-Squared test of independence

Step 1. Set up hypotheses and determine the level of significance.

- ✓ Null hypothesis(H0): Color preference is independent of personality.
- ✓ Alternative hypothesis(H_A): Color preference is dependent on personality
- ✓ $\alpha=0.05$

Five-step approach for Chi-Squared test of independence (contd..)

Step 2. Compute the expected frequency (under the null hypothesis) in each cell using E = (Row Total * Column Total)/N

(Expected counts)	Colors			
Personality	**Red**	**Yellow**	**Green**	**Totals**
Introvert personality	**5.8**	**5.6**	**5.6**	**17**
Extrovert personality	**7.6**	**7.2**	**7.2**	**22**
Normal	**8.6**	**8.2**	**8.2**	**25**
Total	**22**	**21**	**21**	**64**

Step 3:Select the test statistic

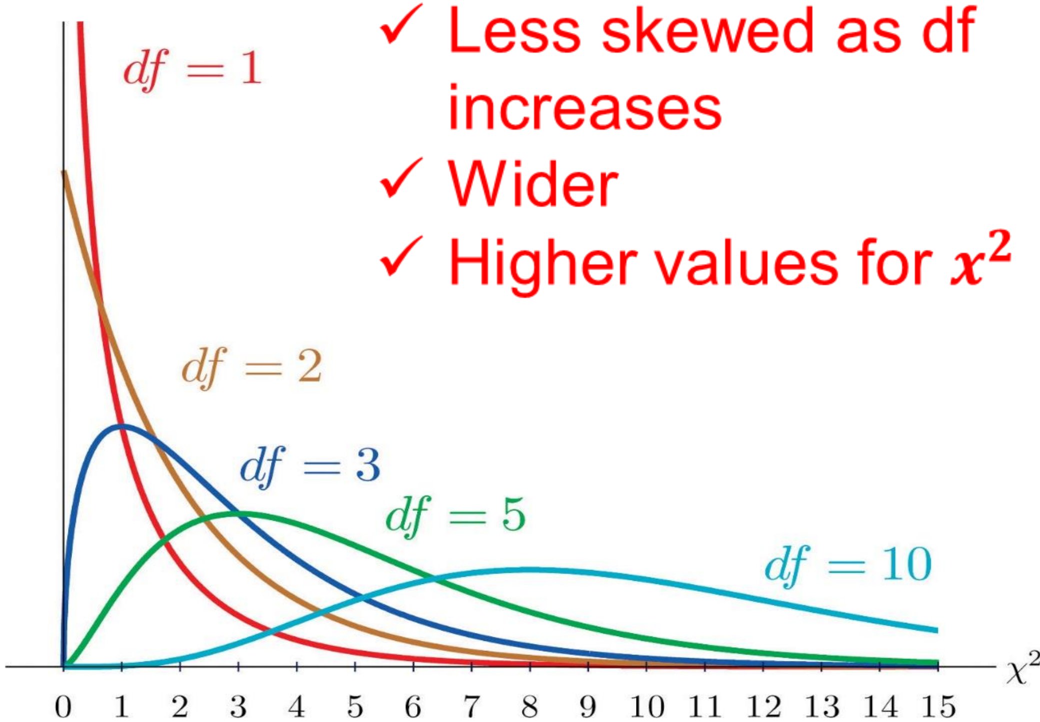
$$x^2 = \Sigma \frac{(0 - E)^2}{E}$$

$$x^2 = \frac{(11-5.8)^2}{5.8} + \frac{(5-5.6)^2}{5.6} + \frac{(1-5.6)^2}{5.6} + \ldots.. + \frac{(12-8.2)^2}{8.2} = 14.5$$

Chi-Squared distribution

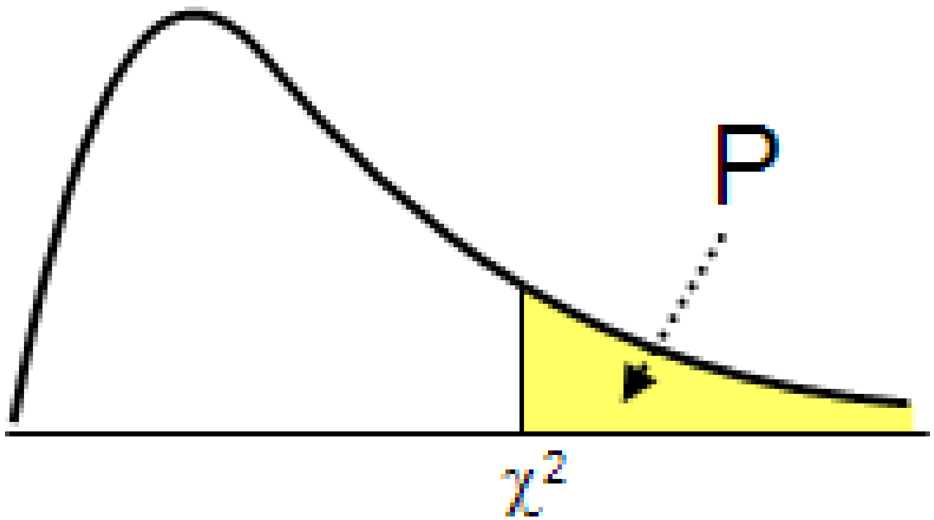
The probability density function for the x^2 distribution with r degrees of freedom(df) is given by

$$P_r(x) = \frac{x^{r/2-1}\, e^{-x/2}}{\Gamma\left(\frac{1}{2}r\right) 2^{r/2}}$$



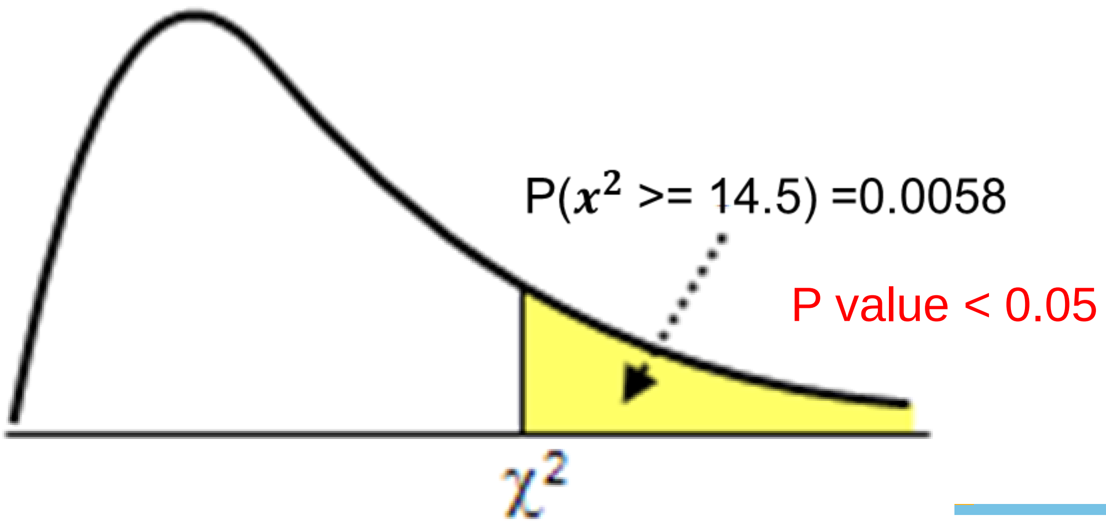
- ✓ Less skewed as df increases
- ✓ Wider
- ✓ Higher values for x^2

Significance of P value



Five-step approach for Chi-Squared test of independence (Contd..)

- Step 4: Use a probability table to find P-Value associated with x^2 value for with degrees of freedom df = (r − 1) (c − 1), r is the number of categories in one variable and c is the number of categories in the other.



	Significance Level				
df	0.10	0.05	0.025	0.01	0.005
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2104	10.5965
3	6.2514	7.8147	9.3484	11.3449	12.8381
4	7.7794	9.4877	11.1433	13.2767	14.8602
5	9.2363	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5475
7	12.017	14.0671	16.0128	18.4753	20.2777

Take home message

- The accuracy of a model depends on selecting the right features.
- Feature subset selection(FSS) helps identify the best subset of features for building the model.
- Three approaches for FSS are Filter, Wrapper, and Embedded approaches.
- Filter based approaches used univariate measure to compute the relationship between Input and output aviable.