Pearson's Correlation Coefficient

- Used to measure the strength of association between two continuous features.
- Both positive and negative correlation are useful.

Steps

1. Compute the Pearson's Correlation Coefficient for each feature.
2. Sort according the score.
3. Retain the highest ranked features, discard the lowest ranked.

Limitation

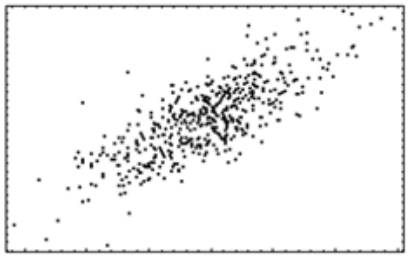
- Pearson assumes all features are independent.
- Pearson identifies only linear correlations

Pearson's Correlation Coefficient

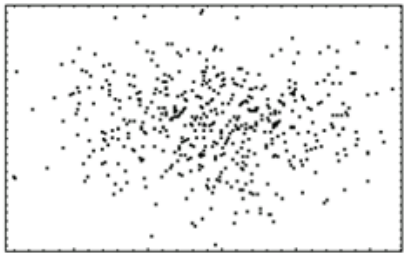
$$\text{Feature: } x_k = \{x_k^{(1)}, \ldots, x_k^{(N)}\}^T$$

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

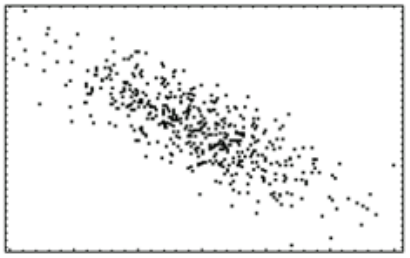
$$\rho(x, y) = \frac{\sum_{i=1}^{N} \left(x^{(i)} - \bar{x}\right)\left(y^{(i)} - \bar{y}\right)}{\sqrt{\left(x^{(i)} - \bar{x}\right)^2}\sqrt{\left(y^{(i)} - \bar{y}\right)^2}}$$



$r = +0.5$



$r = 0.0$



$r = -0.5$

Pearson's Correlation Coefficient

Check whether sale of ice creams and sun glasses are related?

Ice cream sale	Sun glasses sale
A	B
20	30
10	5
23	29
5	10

Pearson's Correlation Coefficient

A	B	$A-\bar{A}$	$(A-\bar{A})^2$	$B-\bar{B}$	$(B-\bar{B})^2$	$(A-\bar{A})(B-\bar{B})$
20	30	5.5	30.25	11.5	132.25	63.25
10	5	-4.5	20.25	-13.5	182.25	60.75
23	29	8.5	72.25	10.5	110.25	89.25
5	10	-9.5	90.25	-8.5	72.25	80.75
58	74		263		497	294

Pearson's Correlation Coefficient

$$\bar{A} = \frac{58}{4} = 14.5$$

$$\bar{B} = \frac{74}{4} = 18.5$$

$$\sigma_A = \sqrt{\frac{213}{3}} = 8.43$$

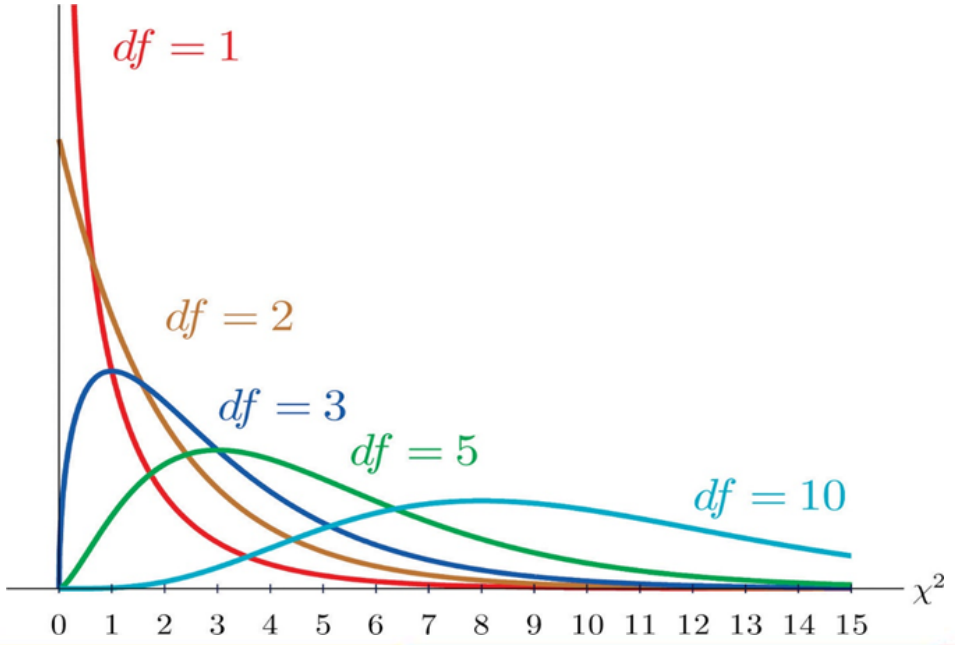
$$\sigma_B = \sqrt{\frac{497}{3}} = 12.87$$

$$r_{A,B} = \frac{294}{4 * 8.43 * 12.87} = 0.68 \approx 1$$

$$= \text{ So positively correlated.}$$

χ^2 Statistic

- Chi-square test of independence allow us to see whether or not two categorical variables are related or not.
- The probability density function for the χ^2 distribution with *r* degrees of freedom (df) .



χ^2 Statistic

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 $3(rows) \times 3(cols)$ contingency table below.

Observed Counts	Colors			
Personality	Red	Yellow	Green	Total
Introvert	11	5	1	17
Extrovert	8	6	8	22
Normal	3	10	12	25
Total	22	21	21	64

χ^2 Statistic

Step 1:

- Set up hypotheses and determine level of significance.
- Null hypothesis(H_0): Color preference is independent of personality.
- Alternative hypothesis(H_A): Color preference is dependent on personality .
- Level of significance: specifies teh probability of error. Generally it is set as 5%.

$$\alpha = 0.005$$

- Assume that H_0 is always true unless the evidence portraits something else in which case we will reject H_0 and accept H_A.

χ^2 Statistic

Step 2:

- Compute the expected count.

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Expected Counts	Colors			
Personality	Red	Yellow	Green	Total
Introvert	5.8	5.6	5.6	17
Extrovert	7.6	7.2	7.2	22
Normal	8.6	8.2	8.2	25
Total	22	21	21	64

χ^2 Statistic

Step 3:

- Compute the Chi-Squared Statistic.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- If H_0 is true, there should not be any difference between the observed values and expected values.

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

χ^2 Statistic

Step 4:

- Use a probability table to find P-Value associated with χ^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.

- $df = (3-1) \times (3-1) = 4$ (contigency table)

	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

- $P(\chi^2 = 14.5) = 0.0058$ (from probability table)

χ^2 Statistic

Step 5:

$$\begin{aligned} \alpha &= 0.05 \\ P(\chi^2 = 14.5) &= 0.0058 \\ &< \alpha \end{aligned}$$

So reject H_0.
Accept H_A.
The two features are independent.

Information Theory Metrics

- Information-theoretic concepts can only be applied to discrete variables.
- For continuous feature values, some data discretization techniques are required beforehand.
- Three metrics
  - Information Gain
  - Gain Ratio
  - Gini Index

Information Gain

- Information Gain $IG(X, Y)$ is a measure of the mutual independence between two random variables X and Y.
- Measures non-linear dependencies.

$$IG(X, Y) = H(Y) - H(Y|X)$$
$$= \sum_{x_i \in X} \sum_{y_i \in Y} P(x_i, y_j) \frac{\log_2 P(x_i, y_j)}{P(x_i)P(y_j)}$$
$$IG(X, Y) = IG(Y, X)$$

- Information Gain is symmetric.
- Higher Information Gain; better prediction of Y given X.
- $I(X, Y) = 0$ if X and Y are independent.
- Biased towards the features having large number of discrete values.

Information Gain

Compute the Information Gain for the attribute Travel Cost.

Gender	Car Ownership	Travel Cost	Income Level	Transport Mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

Information Gain

- Step 1: Compute the Entropy of target.

Transport Mode		
Bus	Car	Train
4	3	3

$$
\begin{aligned}
H(\mathit{Transport}) &= H(4,3,3) \\
&= -\frac{4}{10}\log_2\frac{4}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10} \\
&= 1.571
\end{aligned}
$$

Information Gain

- Step 2: Compute the Entropy of target given one feature.

Feature	Transport Mode		
	Bus	Car	Train
Cheap	4	1	0
Expensive	0	0	3
Standard	0	2	0

$$
\begin{aligned}
H(\mathit{Transport}|\mathit{Cost}) &= H(5,3,2) \\
&= -\frac{5}{10}\left(\frac{4}{5}\log_2\frac{4}{5}+\frac{1}{5}\log_2\frac{1}{5}\right)-\frac{3}{10}\left(\frac{3}{3}\log_2\frac{3}{3}\right)-\frac{2}{10}\left(\frac{2}{2}\log_2\frac{2}{2}\right) \\
&= 0.36
\end{aligned}
$$

Information Gain

- Step 3: Compute the information gain.

$$\begin{aligned} IG(Transport|Cost) &= H(4,3,3) - (H(5,3,2) \\ &= 1.571 - 0.36 \\ &= 1.211 \end{aligned}$$

Gain Ratio

- Gain Ratio $GR(X, Y)$normalizes Information Gain $IG(X, Y)$.
- The information gain ratio is a variant of the mutual information.

$$GR(X, Y) = \frac{IG(A)}{H(A)}$$

- Reduces the bias toward attributes with many discrete values.
- The feature with the maximum gain ratio is selected as the best feature.

Gain Ratio

Compute the Gain Ratio for the attribute Travel Cost.

Gender	Car Ownership	Travel Cost	Income Level	Transport Mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

Gain Ratio

- Step 1: Compute the Entropy of target.

$$\begin{aligned} H(\textit{Transport}) &= H(4,3,3) \\ &= -\frac{4}{10}\log_2\frac{4}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10} \\ &= 1.571 \end{aligned}$$

- Step 2: Compute the Entropy of feature.

$$\begin{aligned} H(\textit{Cost}) &= H(5,3,2) \\ &= -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} \\ &= 1.48 \end{aligned}$$

Gain Ratio

- Step 3: Compute the Entropy of target given one feature.

$$
\begin{aligned}
H(\mathit{Transport}|\mathit{Cost}) &= H(5,3,2) \\
&= -\frac{5}{10}\left(\frac{4}{5}\log_2\frac{4}{5}+\frac{1}{5}\log_2\frac{1}{5}\right)-\frac{3}{10}\left(\frac{3}{3}\log_2\frac{3}{3}\right)-\frac{2}{10}\left(\frac{2}{2}\log_2\frac{2}{2}\right) \\
&= 0.36
\end{aligned}
$$

Gain Ratio

- Step 4: Compute the information gain.

$$\begin{aligned} IG(Transport|Cost) &= H(4,3,3) - (H(5,3,2) \\ &= 1.571 - 0.36 \\ &= 1.211 \end{aligned}$$

- Step 5: Compute Gain Ratio.

$$\begin{aligned} GR(Transport|Cost) &= \frac{IG(Transport|Cost)}{H(Cost)} \\ &= \frac{1.211}{1.48} \\ &= 0.818 \end{aligned}$$

Gini Index

- Gini index minimizes the probability of misclassification.
- Used in CART (Classification and Regression Tree) algorithms.

$$Gini = 1 - \sum_{i=1} Kp_k^2$$

  where p_k denotes the proportion of instances belonging to class k.
- Higher Gini Index; better prediction of Y given X.

Gini Index

Compute the Gini Index for the feature Travel Cost.

Gender	Car Ownership	Travel Cost	Income Level	Transport Mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

Gini Index

- Step 1: Compute the Gini Index for each value of the feature.

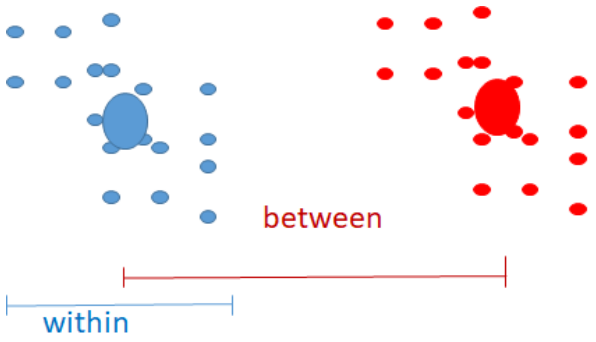
$$Gini(Transport|Cost = Cheap) = 1 - (0.8^2 + 0.2^2) = 0.32$$
$$Gini(Transport|Cost = Expensive) = 1 - (1^2 + 0) = 0$$
$$Gini(Transport|Cost = Standard) = 1 - (1^2 + 0) = 0$$

- Step 2: Compute the Gini Index for feature.

$$Gini(Transport|Cost) = \frac{5}{10} * 0.32 + \frac{3}{10} * 0 + \frac{2}{10} * 0 = 0.16$$

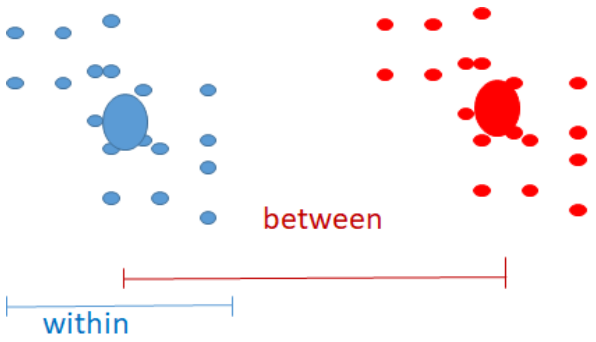
Fisher Score

- Applicable for classification problems with numeric features.
- Metrics can be applied naturally to real-valued features in a binary classification problem or multi-class classification problem.
- Between-class distance -– Distance between the centroids of different classes.
- Within-class distance -– Accumulated distance of an instance to the centroid of its class.



Fisher Score

- Fisher score is the measure the ratio of the average interclass separation to the average intraclass separation.
- The larger the Fisher score, the greater the discriminatory power of the attribute.
- This score is often referred as signal to noise ratio.



Fisher Score

- The Fisher Ratio is defined as the ratio of the variance of the between classes to the variance of within classes.
- Fisher's ratio is a measure for (linear) discriminating power of a variable.
  - Maximum between class variance (difference of means).
  - Minimum within class variance (sum of variances).

$$F = \frac{sum_{j=1}^{k} p_j(\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \sigma_j^2}$$