Data Mining and Machine Learning: Fundamental Concepts and Algorithms

datamininingbook.info

Mohammed J. Zaki[1]   Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 3: Categorical Attributes

Univariate Analysis: Bernoulli Variable

Consider a single categorical attribute, X, with domain $dom(X) = \{a_1, a_2, \ldots, a_m\}$ comprising m symbolic values. The data $\boldsymbol{D}$ is an $n \times 1$ symbolic data matrix given as

$$\boldsymbol{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

where each point $x_i \in dom(X)$.

Bernoulli Variable: Special case when $m = 2$

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

i.e., $dom(X) = \{0, 1\}$.

Bernoulli Variable: Mean and Variance

The probability mass function (PMF) of X is given as

$$P(X=x) = f(x) = p^x(1-p)^{1-x}$$

The expected value of X is given as

$$\mu = E[X] = 1 \cdot p + 0 \cdot (1-p) = p$$

and the variance of X is given as

$$\sigma^2 = var(X) = p(1-p)$$

Assume that each symbolic point has been mapped to its binary value. The set $\{x_1, x_2, \ldots, x_n\}$ is a random sample drawn from X.

The sample mean is given as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n} = \hat{p}$$

where n_i is the number of points with $x_j = i$ in the random sample (equal to the number of occurrences of symbol a_i).

The sample variance is given as

$$\hat{\sigma}^2 = \hat{p}(1-\hat{p})$$

Binomial Distribution: Number of Occurrences

Given the Bernoulli variable X, let $\{x_1, x_2, \ldots, x_n\}$ be a random sample of size n. Let N be the random variable denoting the number of occurrences of the symbol a_1 (value $X = 1$). N has a binomial distribution, given as

$$f(N = n_1 \mid n, p) = \binom{n}{n_1} p^{n_1} (1-p)^{n-n_1}$$

N is the sum of the n independent Bernoulli random variables x_i IID with X, that is, $N = \sum_{i=1}^{n} x_i$. The mean or expected number of occurrences of a_1 is

$$\mu_N = E[N] = E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x_i] = \sum_{i=1}^{n} p = np$$

The variance of N is

$$\sigma_N^2 = var(N) = \sum_{i=1}^{n} var(x_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Multivariate Bernoulli Variable

For the general case when $dom(X) = \{a_1, a_2, \ldots, a_m\}$, we model X as an m-dimensional or *multivariate Bernoulli random variable* $\boldsymbol{X} = (A_1, A_2, \ldots, A_m)^T$, where each A_i is a Bernoulli variable with parameter p_i denoting the probability of observing symbol a_i.

However, X can assume only one of the symbolic values at any one time. Thus,

$$\boldsymbol{X}(v) = \boldsymbol{e}_i \text{ if } v = a_i$$

where $\boldsymbol{e}_i$ is the i-th standard basis vector in m dimensions. The range of $\boldsymbol{X}$ consists of m distinct vector values $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m\}$.

The PMF of $\boldsymbol{X}$ is

$$\boxed{P(\boldsymbol{X} = \boldsymbol{e}_i) = f(\boldsymbol{e}_i) = p_i = \prod_{j=1}^{m} p_j^{e_{ij}}}$$

with $\sum_{i=1}^{m} p_i = 1$.

Multivariate Bernoulli: Mean

The mean or expected value of $\boldsymbol{X}$ can be obtained as

$$\boldsymbol{\mu} = E[\boldsymbol{X}] = \sum_{i=1}^{m} \boldsymbol{e}_i f(\boldsymbol{e}_i) = \sum_{i=1}^{m} \boldsymbol{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \boldsymbol{p}$$

The sample mean is

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = \sum_{i=1}^{m} \frac{n_i}{n} \boldsymbol{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\boldsymbol{p}}$$

where n_i is the number of occurrences of the vector value $\boldsymbol{e}_i$ in the sample, i.e., the number of occurrences of the symbol a_i. Furthermore, $\sum_{i=1}^{m} n_i = n$.

Bins	Domain	Counts
$[4.3, 5.2]$	Very Short (a_1)	$n_1 = 45$
$(5.2, 6.1]$	Short (a_2)	$n_2 = 50$
$(6.1, 7.0]$	Long (a_3)	$n_3 = 43$
$(7.0, 7.9]$	Very Long (a_4)	$n_4 = 12$

We model `sepal length` as a multivariate Bernoulli variable $\boldsymbol{X}$

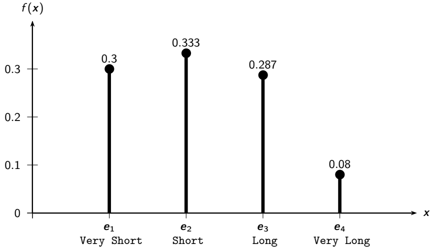
$$\boldsymbol{X}(v) = \begin{cases} \boldsymbol{e}_1 = (1,0,0,0) & \text{if } v = a_1 \\ \boldsymbol{e}_2 = (0,1,0,0) & \text{if } v = a_2 \\ \boldsymbol{e}_3 = (0,0,1,0) & \text{if } v = a_3 \\ \boldsymbol{e}_4 = (0,0,0,1) & \text{if } v = a_4 \end{cases}$$

For example, the symbolic point $x_1 = \texttt{Short} = a_2$ is represented as the vector $(0,1,0,0)^T = \boldsymbol{e}_2$.

Probability Mass Function

The total sample size is $n = 150$; the estimates $\hat{p}_i$ are:

$$\hat{p}_1 = 45/150 = 0.3$$
$$\hat{p}_2 = 50/150 = 0.333$$
$$\hat{p}_3 = 43/150 = 0.287$$
$$\hat{p}_4 = 12/150 = 0.08$$



Multivariate Bernoulli Variable: Covariance Matrix

We have $\boldsymbol{X} = (A_1, A_2, \ldots, A_m)^T$, where A_i is the Bernoulli variable corresponding to symbol a_i. The variance for each Bernoulli variable A_i is

$$\boxed{\sigma_i^2 = var(A_i) = p_i(1-p_i)}$$

The covariance between A_i and A_j is

$$\boxed{\sigma_{ij} = E[A_iA_j] - E[A_i]\cdot E[A_j] = 0 - p_ip_j = -p_ip_j}$$

Negative relationship since A_i and A_j cannot both be 1 at the same time.
The covariance matrix for $\boldsymbol{X}$ is given as

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_m & -p_2p_m & \cdots & p_m(1-p_m) \end{pmatrix}$$

More compactly $\Sigma = diag(\boldsymbol{p}) - \boldsymbol{p}\cdot\boldsymbol{p}^T$ where $\boldsymbol{\mu} = \boldsymbol{p} = (p_1, \cdots, p_m)^T$.

Categorical, Mapped Binary and Centered Dataset

Modeling as multivariate Bernoulli variable is equivalent to treating $\boldsymbol{X}(x_i)$ as a new $n \times m$ binary data matrix

	X
x_1	Short
x_2	Short
x_3	Long
x_4	Short
x_5	Long

	A_1	A_2
$\mathbf{x}_1$	0	1
$\mathbf{x}_2$	0	1
$\mathbf{x}_3$	1	0
$\mathbf{x}_4$	0	1
$\mathbf{x}_5$	1	0

	Z_1	Z_2
$\mathbf{z}_1$	-0.4	0.4
$\mathbf{z}_2$	-0.4	0.4
$\mathbf{z}_3$	0.6	-0.6
$\mathbf{z}_4$	-0.4	0.4
$\mathbf{z}_5$	0.6	-0.6

X is the multivariate Bernoulli variable

$$\boldsymbol{X}(v) = \begin{cases} \mathbf{e}_1 = (1,0)^T & \text{if } v = \texttt{Long}(a_1) \\ \mathbf{e}_2 = (0,1)^T & \text{if } v = \texttt{Short}(a_2) \end{cases}$$

The sample mean and covariance matrix are

$$\hat{\boldsymbol{\mu}} = \hat{\mathbf{p}} = (2/5, 3/5)^T = (0.4, 0.6)^T \qquad \widehat{\Sigma} = diag(\hat{\mathbf{p}}) - \hat{\mathbf{p}}\hat{\mathbf{p}}^T = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

From the centered data, we have $\boldsymbol{Z} = (Z_1, Z_2)^T$ and

$$\widehat{\Sigma} = \frac{1}{5}\boldsymbol{Z}^T\boldsymbol{Z} = \begin{pmatrix} 0.24 & -0.24 \\ -0.24 & 0.24 \end{pmatrix}$$

Multinomial Distribution: Number of Occurrences

Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be a random sample from $\boldsymbol{X}$. Let N_i be the random variable denoting number of occurrences of symbol a_i in the sample, and let $\boldsymbol{N} = (N_1, N_2, \ldots, N_m)^T$.
$\boldsymbol{N}$ has a multinomial distribution, given as

$$f\left(\boldsymbol{N} = (n_1, n_2, \ldots, n_m) \mid \boldsymbol{p}\right) = \binom{n}{n_1 n_2 \ldots n_m} \prod_{i=1}^{m} p_i^{n_i}$$

The mean and covariance matrix of $\boldsymbol{N}$ are:

$$\boldsymbol{\mu}_{\boldsymbol{N}} = E[\boldsymbol{N}] = nE[\boldsymbol{X}] = n \cdot \boldsymbol{\mu} = n \cdot \boldsymbol{p} = \begin{pmatrix} np_1 \\ \vdots \\ np_m \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{N}} = n \cdot (diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_m \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_m & -np_2p_m & \cdots & np_m(1-p_m) \end{pmatrix}$$

The sample mean and covariance matrix for $\boldsymbol{N}$ are

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{N}} = n\hat{\boldsymbol{p}} \qquad \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{N}} = n\left(diag(\hat{\boldsymbol{p}}) - \hat{\boldsymbol{p}}\hat{\boldsymbol{p}}^T\right)$$

Bivariate Analysis

Assume the data comprises two categorical attributes, X_1 and X_2,

$$dom(X_1) = \{a_{11}, a_{12}, \dots, a_{1m_1}\}$$
$$dom(X_2) = \{a_{21}, a_{22}, \dots, a_{2m_2}\}$$

We model X_1 and X_2 as multivariate Bernoulli variables $\mathbf{X}_1$ and $\mathbf{X}_2$ with dimensions m_1 and m_2, respectively. The joint distribution of $\mathbf{X}_1$ and $\mathbf{X}_2$ is modeled as the $m_1 + m_2$ dimensional vector variable $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$

$$\mathbf{X}\left((v_1, v_2)^T\right) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \mathbf{X}_2(v_2) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2j} \end{pmatrix}$$

provided that $v_1 = a_{1i}$ and $v_2 = a_{2j}$.
The joint PMF for $\mathbf{X}$ is given as the $m_1 \times m_2$ matrix

$$\mathbf{P}_{12} = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1m_2} \\ p_{21} & p_{22} & \dots & p_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} & p_{m_1 2} & \dots & p_{m_1 m_2} \end{pmatrix}$$

Bivariate Empirical PMF: sepal length and sepal width

X_1:sepal length

Bins	Domain	Counts
$[4.3, 5.2]$	Very Short (a_1)	$n_1 = 45$
$(5.2, 6.1]$	Short (a_2)	$n_2 = 50$
$(6.1, 7.0]$	Long (a_3)	$n_3 = 43$
$(7.0, 7.9]$	Very Long (a_4)	$n_4 = 12$

X_2:sepal width

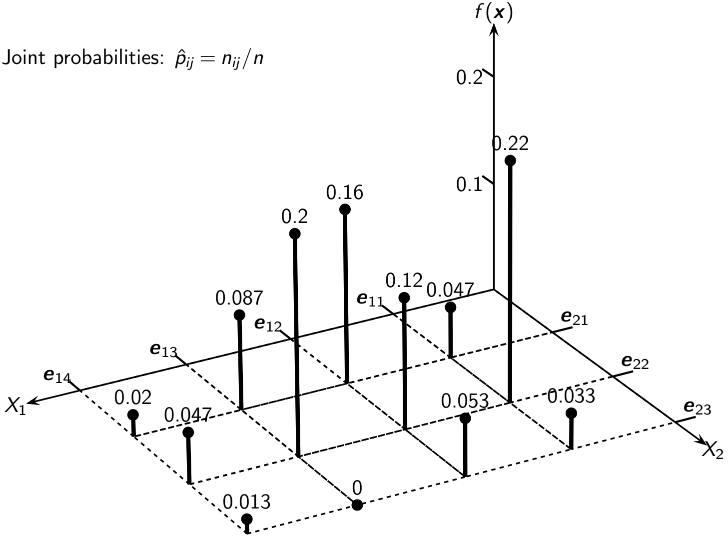
Bins	Domain	Counts
$[2.0, 2.8]$	Short (a_1)	47
$(2.8, 3.6]$	Medium (a_2)	88
$(3.6, 4.4]$	Long (a_3)	15

Observed Counts (n_{ij})

		X_2		
		Short ($\mathbf{e}_{21}$)	Medium ($\mathbf{e}_{22}$)	Long ($\mathbf{e}_{23}$)
X_1	Very Short ($\mathbf{e}_{11}$)	7	33	5
	Short ($\mathbf{e}_{22}$)	24	18	8
	Long ($\mathbf{e}_{13}$)	13	30	0
	Very Long ($\mathbf{e}_{14}$)	3	7	2

Bivariate Empirical PMF: sepal length and sepal width

Joint probabilities: $\hat{p}_{ij} = n_{ij}/n$



Attribute Dependence: Contingency Analysis

The *contingency table* for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is the $m_1 \times m_2$ matrix of observed counts n_{ij}

$$\boldsymbol{N}_{12} = n \cdot \widehat{\boldsymbol{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

where $\widehat{\boldsymbol{P}}_{12}$ is the empirical joint PMF for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The contingency table is augmented with row and column marginal counts, as follows:

$$\boldsymbol{N}_1 = n \cdot \hat{\boldsymbol{p}}_1 = \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \end{pmatrix} \qquad \boldsymbol{N}_2 = n \cdot \hat{\boldsymbol{p}}_2 = \begin{pmatrix} n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix}$$

$\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ have a multinomial distribution with parameters $\boldsymbol{p}_1 = (p_1^1, \ldots, p_{m_1}^1)$ and $\boldsymbol{p}_2 = (p_1^2, \ldots, p_{m_2}^2)$, respv.
$\boldsymbol{N}_{12}$ also has a multinomial distribution with parameters $\boldsymbol{P}_{12} = \{p_{ij}\}$, for $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$.

Contingency Table: `sepal length` vs. `sepal width`

Sepal length (X_1)	Sepal width (X_2)				
		Short a_{21}	Medium a_{22}	Long a_{23}	Row Counts
	Very Short (a_{11})	7	33	5	$n_1^1 = 45$
	Short (a_{12})	24	18	8	$n_2^1 = 50$
	Long (a_{13})	13	30	0	$n_3^1 = 43$
	Very Long (a_{14})	3	7	2	$n_4^1 = 12$
	Column Counts	$n_1^2 = 47$	$n_2^2 = 88$	$n_3^2 = 15$	$n = 150$

Chi-Squared Test for Independence

Assume $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. Then, their joint PMF is

$$\hat{p}_{ij} = \hat{p}^1_i \cdot \hat{p}^2_j$$

The expected frequency for each pair of values is

$$e_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}^1_i \cdot \hat{p}^2_j = n \cdot \frac{n^1_i}{n} \cdot \frac{n^2_j}{n} = \frac{n^1_i n^2_j}{n}$$

The χ^2 statistic quantifies the difference between observed and expected counts

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

The sampling distribution for the χ^2 statistic follows the *chi-squared* density function:

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{\frac{q}{2}-1} e^{-\frac{x}{2}}$$

where q is the degrees of freedom

$$\begin{aligned} q &= |dom(X_1)| \times |dom(X_2)| - (|dom(X_1)| + |dom(X_2)|) + 1 \\ &= m_1 m_2 - m_1 - m_2 + 1 \\ &= (m_1 - 1)(m_2 - 1) \end{aligned}$$

Chi-Squared Test: sepal length and sepal width

	Expected Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Short (a_{23})
X_1	Very Short (a_{11})	14.1	26.4	4.5
	Short (a_{12})	15.67	29.33	5.0
	Long (a_{13})	13.47	25.23	4.3
	Very Long (a_{14})	3.76	7.04	1.2

	Observed Counts	X_2		
		Short (a_{21})	Medium (a_{22})	Long (a_{23})
	Very Short (a_{11})	7	33	5
	Short (a_{12})	24	18	8
	Long (a_{13})	13	30	0
	Very Long (a_{14})	3	7	2

The chi-squared statistic value is $\chi^2 = 21.8$.

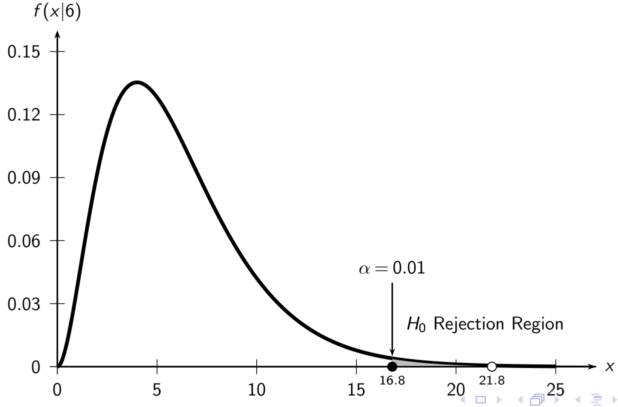
The number of degrees of freedom are

$$q = (m_1 - 1) \cdot (m_2 - 1) = 3 \cdot 2 = 6$$

Chi-Squared Distribution ($q = 6$).

The *p-value* of a statistic θ is defined as the probability of obtaining a value at least as extreme as the observed value.

The null hypothesis, that X_1 and X_2 are independent, is rejected if $p\text{-value}(z) \le \alpha$, say $\alpha = 0.01$. We have $p\text{-}value(21.8) = 0.0013$. Thus, we reject the null hypothesis, and conclude that X_1 and X_2 are dependent.



Multiway Contingency Analysis

Given $\boldsymbol{X} = (X_1, X_2, \cdots, X_d)^T$. The chi-squared statistic is given as

$$\boxed{\chi^2 = \sum_{\boldsymbol{i}} \frac{(n_{\boldsymbol{i}} - e_{\boldsymbol{i}})^2}{e_{\boldsymbol{i}}} =} \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \cdots \sum_{i_d=1}^{m_d} \frac{(n_{i_1,i_2,\ldots,i_d} - e_{i_1,i_2,\ldots,i_d})^2}{e_{i_1,i_2,\ldots,i_d}}$$

Under the null hypothesis, that attributes are independent, the expected number of occurrences of the symbol tuple $(a_{1i_1}, a_{2i_2}, \ldots, a_{di_d})$ is given as

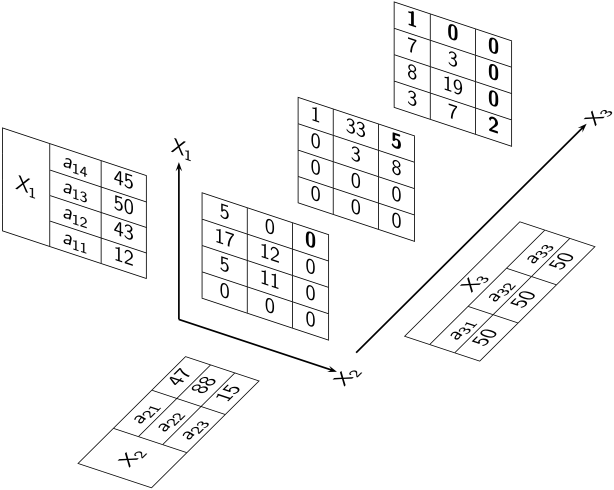
$$e_{\boldsymbol{i}} = n \cdot \hat{p}_{\boldsymbol{i}} = n \cdot \prod_{j=1}^{d} \hat{p}_{i_j}^{j} = \frac{n_{i_1}^1 n_{i_2}^2 \ldots n_{i_d}^d}{n^{d-1}}$$

The total number of degrees of freedom for the chi-squared distribution is given as

$$\begin{aligned} q &= \prod_{i=1}^{d} |dom(X_i)| - \sum_{i=1}^{d} |dom(X_i)| + (d-1) \\ &= \left(\prod_{i=1}^{d} m_i\right) - \left(\sum_{i=1}^{d} m_i\right) + d - 1 \end{aligned}$$

3-Way Contingency Table

X_1: sepal length, X_2: sepal width and X_3: Iris type



3-Way Contingency Analysis

		$X_3(a_{31}/a_{32}/a_{33})$		
		X_2		
		a_{21}	a_{22}	a_{23}
X_1	a_{11}	1.25	2.35	0.40
	a_{12}	4.49	8.41	1.43
	a_{13}	5.22	9.78	1.67
	a_{14}	4.70	8.80	1.50

The value of the χ^2 statistic is $\chi^2 = 231.06$, and the number of degrees of freedom is $q = 4 \cdot 3 \cdot 3 - (4+3+3) + 2 = 36 - 10 + 2 = 28$.

For a significance level of $\alpha = 0.01$, the critical value of the chi-square distribution is $z = 48.28$.

The observed value of $\chi^2 = 231.06$ is much greater than z, and it is thus extremely unlikely to happen under the null hypothesis. We conclude that the three attributes are not 3-way independent, but rather there is some dependence between them.

Distance and Angle

With the modeling of categorical attributes as multivariate Bernoulli variables, it is possible to compute the distance or the angle between any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{e}_{1i_1} \\ \vdots \\ \boldsymbol{e}_{d\, i_d} \end{pmatrix} \qquad \boldsymbol{x}_j = \begin{pmatrix} \boldsymbol{e}_{1j_1} \\ \vdots \\ \boldsymbol{e}_{d\, j_d} \end{pmatrix}$$

The different measures of distance and similarity rely on the number of matching and mismatching values (or symbols) across the d attributes $\boldsymbol{X}_k$.
The number of matching values s is given as:

$$s = \boldsymbol{x}_i^T \boldsymbol{x}_j = \sum_{k=1}^{d} (\boldsymbol{e}_{ki_k})^T \boldsymbol{e}_{kj_k}$$

The number of mismatches is simply $d - s$. Also useful is the norm of each point:

$$\|\boldsymbol{x}_i\|^2 = \boldsymbol{x}_i^T \boldsymbol{x}_i = d$$

Distance and Angle

The *Euclidean distance* between $\mathbf{x}_i$ and $\mathbf{x}_j$ is given as

$$\delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j} = \sqrt{2(d-s)}$$

The *Hamming distance* is given as

$$\delta_H(\mathbf{x}_i, \mathbf{x}_j) = d - s$$

Cosine Similarity: The cosine of the angle is given as

$$\cos\theta = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|} = \frac{s}{d}$$

The *Jaccard Coefficient* is given as

$$J(\mathbf{x}_i, \mathbf{x}_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

Discretization

Discretization, also called *binning*, converts numeric attributes into categorical ones.

Equal-Width Intervals: Partition the range of X into k *equal-width* intervals. The interval width is simply the range of X divided by k:

$$w = \frac{x_{\max} - x_{\min}}{k}$$

Thus, the ith interval boundary is given as

$$v_i = x_{\min} + iw, \text{ for } i = 1, \ldots, k-1$$

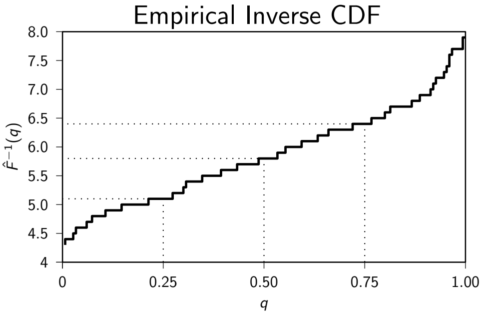
Equal-Frequency Intervals: We divide the range of X into intervals that contain (approximately) equal number of points. The intervals are computed from the empirical quantile or inverse cumulative distribution function

$$\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}$$

We require that each interval contain $1/k$ of the probability mass; therefore, the interval boundaries are given as follows:

$$v_i = \hat{F}^{-1}(i/k) \text{ for } i = 1, \ldots, k-1$$

Equal-Frequency Discretization: sepal length (4 bins)



Quartile values:

$$\hat{F}^{-}1(0.25) = 5.1$$
$$\hat{F}^{-}1(0.5) = 5.8$$
$$\hat{F}^{-}1(0.75) = 6.4$$

Range: $[4.3, 7.9]$

Bin	Width	Count
$[4.3, 5.1]$	0.8	$n_1 = 41$
$(5.1, 5.8]$	0.7	$n_2 = 39$
$(5.8, 6.4]$	0.6	$n_3 = 35$
$(6.4, 7.9]$	1.5	$n_4 = 35$

Data Mining and Machine Learning: Fundamental Concepts and Algorithms

datamininingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 3: Categorical Attributes