Probability Theory

(short reminder)

A discrete random variable x can assume any of a finite number of m different values in $X = \{v_1, v_2, \ldots, v_m\}$

Probability: $$p_i = \Pr[x = v_i], \quad i = 1, \ldots, m$$

$$p_i \geq 0 \qquad \sum_{i=1}^{m} p_i = 1$$

More generally, for a random real variable x, we use a probability function $p(x)$

Expected value, also called *mean* or *average,* of a random variable x

$$\varepsilon[x] = \mu = \sum_{x \in X} x\ p(x) = \sum_{i=1}^{m} v_i p_i$$

More generally, if *f(x)* is a a function of x, the expected value of *f(x)* is

$$\varepsilon[f(x)] = \sum_{x \in X} f(x)\ p(x)$$

Linearity:

$$\varepsilon[\alpha_1 f_1(x)+\alpha_2 f_2(x)]=\alpha_1\varepsilon[f_1(x)]+\alpha_2\varepsilon[f_2(x)]$$

Special cases:

- Second moment

$$\varepsilon[x^2]=\sum_{x\in X} x^2 p(x)$$

- Variance

$$Var[x]=\sigma^2=\varepsilon[(x-\mu)^2]=\sum_{x\in X}(x-\mu)^2 p(x)$$

Useful formula:

$$Var[x] = \varepsilon[x^2] - (\varepsilon[x])^2$$

Unlike the mean, the variance is not linear:

$$y = \alpha\, x \qquad \Rightarrow \qquad Var[y] = \alpha^2\, Var[x]$$

For two random variables

$$x \in X,\ X = \{v_1, v_2, \ldots, v_m\}$$
$$y \in Y,\ \ Y = \{w_1, w_2, \ldots, w_m\}$$

joint probability

$$p_{ij} = \Pr[x = v_i, y = w_j]$$

Joint probability function:

$$p(x,y) \geq 0 \qquad\qquad \sum_{x \in X} \sum_{y \in Y} p(x,y) = 1$$

Marginal probability:

$$p_x(x) = \sum_{y \in Y} p(x,y)$$

$$p_y(y) = \sum_{x \in X} p(x,y)$$

Statistical independence

$$p(x, y) = p_x(x)\ p_y(y)$$

Expected values of functions of two variables

$$\varepsilon[f(x, y)] = \sum_{x \in X} \sum_{y \in Y} f(x, y)\ p(x, y)$$

Linearity:

$$\varepsilon[\alpha_1 f_1(x,y)+\alpha_2 f_2(x,y)]=\alpha_1\varepsilon[f_1(x,y)]+\alpha_2\varepsilon[f_2(x,y)]$$

Special cases:

- Mean $\mu_x = \varepsilon[x]=\sum_{x\in X}\sum_{y\in Y} x\ p(x,y)$

$$\mu_y = \varepsilon[y]=\sum_{x\in X}\sum_{y\in Y} y\ p(x,y)$$

- Variance

$$\sigma_x^2 = Var[x]=\varepsilon[(x-\mu_x)^2]=\sum_{x\in X}\sum_{y\in Y}(x-\mu_x)^2\, p(x,y)$$

$$\sigma_y^2 = Var[y]=\varepsilon[(y-\mu_y)^2]=\sum_{x\in X}\sum_{y\in Y}(y-\mu_y)^2\, p(x,y)$$

Covariance (cross-moment):

$$\sigma_{xy} = \varepsilon[(x-\mu_x)(y-\mu_x)] = \sum_{x\in X}\sum_{y\in Y}(x-\mu_x)(y-\mu_y)\,p(x,y)$$

Matrix notation:

$$\mathrm{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

$$\sum = \varepsilon[(\mathrm{x}-\mu)(\mathrm{x}-\mu)^t]$$

Correlation coefficient

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad \rho \in [-1, 1]$$

$\rho = 1$ - events are maximally positively correlated

$\rho = -1$ - events are maximally negatively correlated

$\rho = 0$ - events are uncorrelated

Conditional probability

$$\Pr[x = v_i \mid y = w_j] = \frac{\Pr[x = v_i, y = w_j]}{\Pr[y = w_j]}$$

or, in terms of probability functions

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

In case of a vector x

mean

$$\mu = \varepsilon[\mathrm{x}] = \begin{bmatrix} \varepsilon[x_1] \\ \varepsilon[x_2] \\ \cdots \\ \varepsilon[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_d \end{bmatrix} = \sum_{\mathrm{x}} \mathrm{x}\ p(\mathrm{x})$$

$$\sum = \varepsilon[(\mathrm{x} - \mu)(\mathrm{x} - \mu)^t]$$

Covariance matrix

$$\sum = \begin{bmatrix} \varepsilon[(x_1-\mu_1)(x_1-\mu_1)] & \varepsilon[(x_1-\mu_1)(x_2-\mu_2)] & \dots & \varepsilon[(x_1-\mu_1)(x_d-\mu_d)] \\ \varepsilon[(x_2-\mu_2)(x_1-\mu_1)] & \varepsilon[(x_2-\mu_2)(x_2-\mu_2)] & \dots & \varepsilon[(x_2-\mu_2)(x_d-\mu_d)] \\ \dots & \dots & & \\ \varepsilon[(x_d-\mu_d)(x_1-\mu_1)] & \varepsilon[(x_d-\mu_d)(x_2-\mu_2)] & \dots & \varepsilon[(x_d-\mu_d)(x_d-\mu_d)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{21} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2d} \\ \dots & \dots & & \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1{}^2 & \sigma_{21} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2{}^2 & \dots & \sigma_{2d} \\ \dots & \dots & & \\ \sigma_{d1} & \sigma_{d2} & \dots & \sigma_d{}^2 \end{bmatrix}$$

$$\sum = \varepsilon[(\mathrm{x}-\mu)(\mathrm{x}-\mu)^t]$$

Properties of the covariance matrix:

- symmetric
- positive-semi-definite
- if the variables are statistically independent, it is diagonal
- its eigenvalues are positive

Distribution of sums of independent variables $\mathrm{z} = \mathrm{x} + \mathrm{y}$

Mean:

$$\mu = \varepsilon[\mathrm{z}] = \varepsilon[\mathrm{x} + \mathrm{y}] = \varepsilon[\mathrm{x}] + \varepsilon[\mathrm{y}] = \mu_x + \mu_y$$

Variance:

$$\sigma^2 = \varepsilon[(\mathrm{z} - \mu_\mathrm{z})^2] = \varepsilon[(\mathrm{x} + \mathrm{y} - (\mu_\mathrm{x} + \mu_\mathrm{y}))^2] =$$

$$= \varepsilon[((\mathrm{x} - \mu_\mathrm{x}) + (\mathrm{y} - \mu_\mathrm{y}))^2] =$$

$$= \varepsilon[(\mathrm{x} - \mu_\mathrm{x})^2] + 2\varepsilon[(\mathrm{x} - \mu_\mathrm{x}) + (\mathrm{y} - \mu_\mathrm{y})] + \varepsilon[(\mathrm{y} - \mu_\mathrm{y})^2]$$

$$= \sigma_x^{\ 2} + \sigma_y^{\ 2}$$

Theorem: The probability distribution function (pdf) of the sum of two independent random variables is the convolution between the concerned pdf's.

(Left as assignment)

Normal (Gaussian) distribution:

$$p(x) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

$$p(x) \sim N(\mu, \sigma^2)$$

$$\varepsilon[x] = \int_{-\infty}^{\infty} x\ p(x)\ dx = \mu$$

$$\varepsilon[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\ p(x)\ dx = \sigma^2$$

The Mahalanobis distance from x to μ is:

$$r = \frac{|x - \mu|}{\sigma}$$

In the one-dimensional case, this distance is also called the *z-score.*

Multivariate normal densities of d independent distributions:

$$p(\mathrm{x}) = \prod_{i=1}^{d} p(x_i) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^{\,2}} \right]$$

$$= \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_i} \exp\left[-\frac{1}{2} \sum_{i=1}^{d} \frac{(x_i - \mu_i)^2}{\sigma_i^{\,2}} \right]$$

The covariance matrix in this case is diagonal:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & & \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & \dots & & \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{bmatrix}$$

General quadratic form:

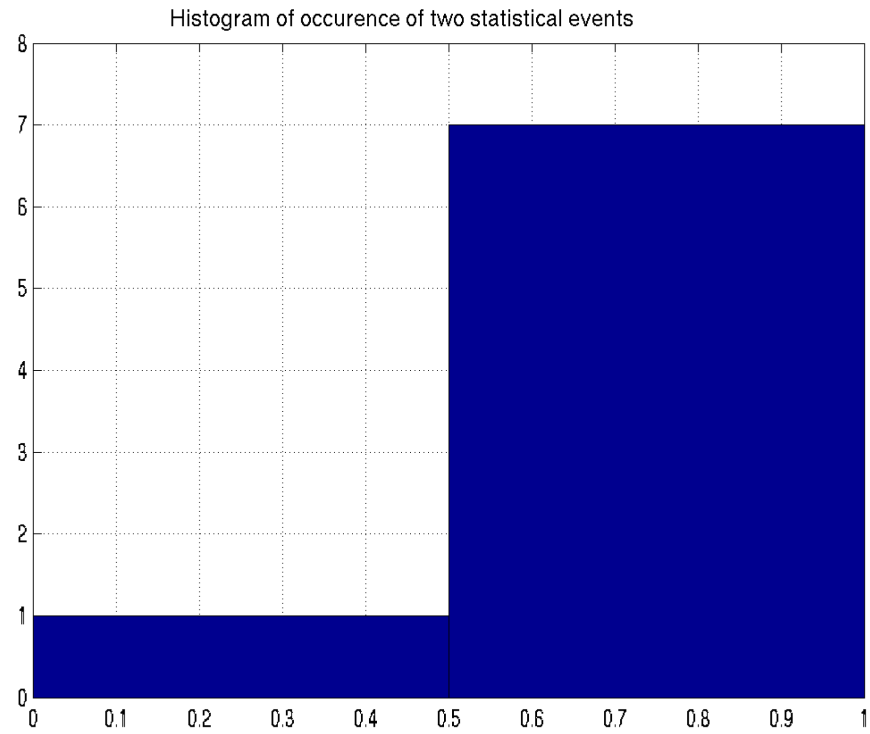
$$p(\mathrm{x}) = \frac{1}{(2\pi)^{d/2} \,|\, \sum |^{1/2}} \exp\left[-\frac{1}{2} (\mathrm{x} - \mu)^t \sum{}^{-1} (\mathrm{x} - \mu) \right]$$

Square of the Mahalanobis distance from x to μ :

$$r^2 = (\mathrm{x} - \mu)^t \sum{}^{-1} (\mathrm{x} - \mu)$$

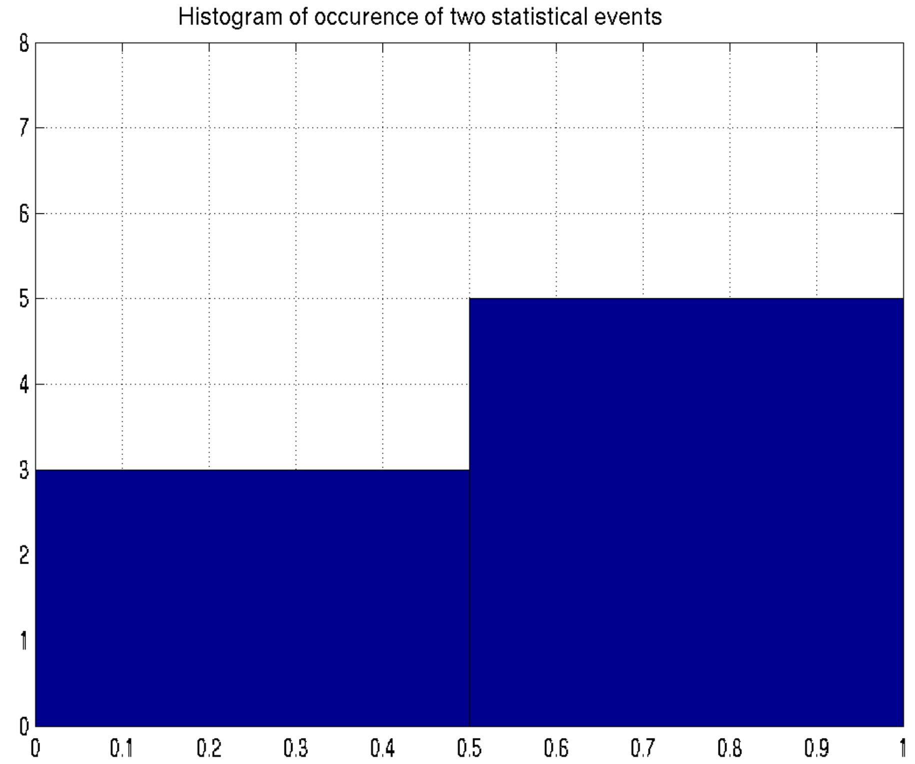
Entropy of a distribution:

$$H(p(x)) = -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx$$

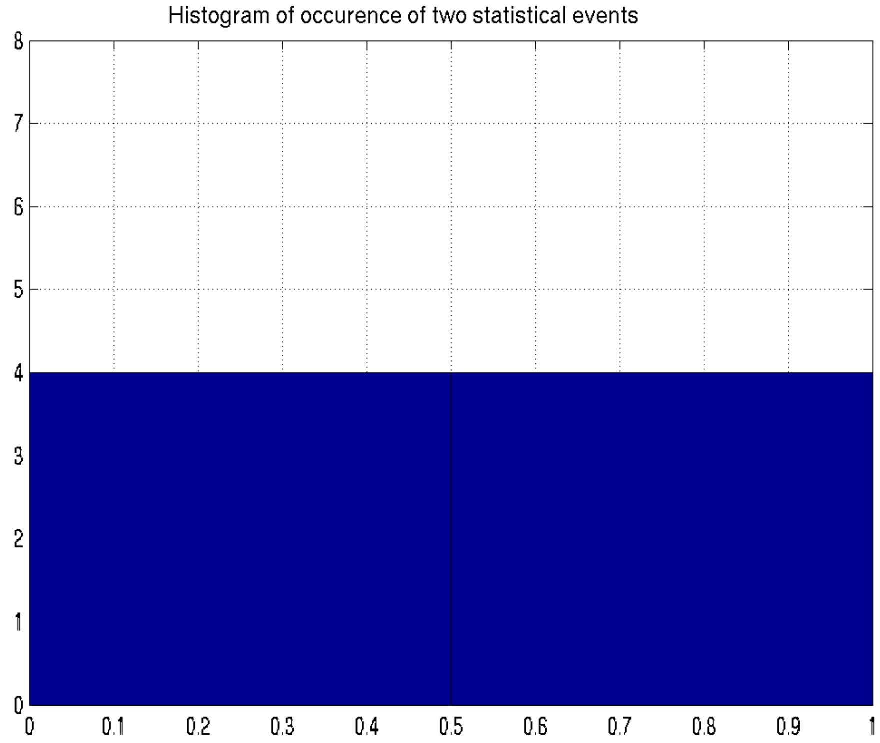


$$H(p(x)) = -0.125\ln(0.125) - 0.875\ln(0.875) = 0.3768$$

Probability Theory



$$H(p(x)) = -0.375\ln(0.375) - 0.625\ln(0.625) = 0.6616$$



$$H(p(x)) = -0.5\ln(0.5) - 0.5\ln(0.5) = 0.6931$$