Exercise Sheet 2

1. Exercise: Shannon Entropy

Shannon entropy quantifies the amount of uncertainty (or randomness) to a variable's possible outcomes. Given a discrete random variable X, which takes the values $x \in \mathcal{X}$ and is distributed according to $p\colon \mathcal{X} \to [0,1]$, the Shannon entropy is

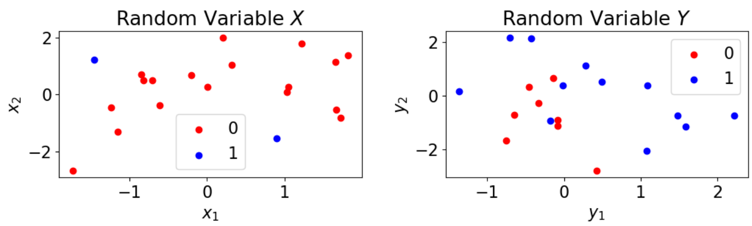
$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{1}$$

The higher the entropy, the more uncertain, and thus in some sense the more random, and hence less predictable, the outcomes. The maximal entropy of an event with n different possible outcomes is

$$H_{\max} = \log n, \tag{2}$$

If the outcomes are completely determined and there is no randomness at all, the entropy is zero ($H_{\min} = 0$).

(a) Load the data `data_entropy_binary.csv`. You can find the file in the Moodle course. Generate the following plot (one plot with two axes) for the random variable $X = [\boldsymbol{x}_1, \boldsymbol{x}_2]$ with labels $\boldsymbol{x}_l$ and random variable $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2]$ with labels $\boldsymbol{y}_l$.



(b) Implement the function `compute_probabilites()` to calculate the probabilities of each label for a given set of labels of a random variable.

(c) Implement the function `compute_entropy()` that computes the entropy according to Eq. 1.

(d) Compute and print the entropy and maximal entropy for the random variables X, Y.

(e) Load the data `data_entropy_multi.csv`. You can find the file in the Moodle course. Generate a similar plot as in (b) for the random variable $X = [\boldsymbol{x}_1, \boldsymbol{x}_2]$ with labels $\boldsymbol{x}_l$ and random variable $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2]$ with labels $\boldsymbol{y}_l$. Note, that the labels are not binary in this dataset.

(f) Compute and print the entropy and maximal entropy for the random variables X, Y.

(g) Interpret the results of (d) and (f). What is the influence of the quantity of labels on the entropy calculation?

2. Exercise: Mahanalobis distance

The Mahalanobis distance is a measure of the distance between a point and a distribution. Given a probability distribution $p = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C}_x)$, with mean $\boldsymbol{\mu} \in \mathbb{R}^N$ and covariance matrix $\mathbf{C}_x \in \mathbb{R}^{N\times N}$, the Mahalanobis distance of a point $\boldsymbol{x} \in \mathbb{R}^N$ from p is

$$\delta_m(\boldsymbol{x}, p) = \sqrt{(\boldsymbol{\mu} - \boldsymbol{x})^{\mathrm{T}} \boldsymbol{C}_x^{-1} (\boldsymbol{\mu} - \boldsymbol{x})}. \tag{3}$$

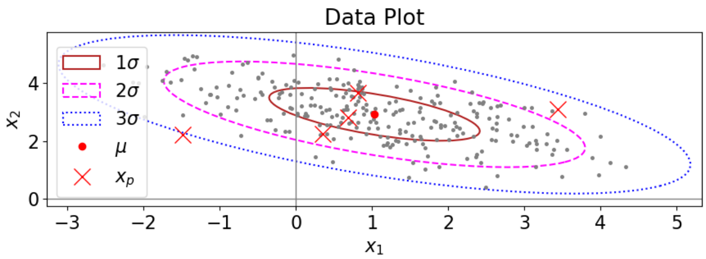
(a) Implement the function `mahanalobis_distance()`, which takes as input the mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{C}_x$ and a point $\boldsymbol{x}$ and return the Mahanalobis distance.

(b) Load the file `data_mdistance.csv`. You can find the file in the Moodle course. Compute the mean $\hat{\boldsymbol{\mu}}$ and the covariance matrix $\hat{\boldsymbol{C}}_x$.

(c) Compute the Pearson Correlation Coefficient and print it in the console. What does the coefficient tell about the correlation?

(d) Load the file `data_mdistance_points.csv`. You can find the file in the Moodle course. Compute the Mahanalobis distance (based on $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{C}}_x$ computed in (b)) and the Euclidean distance for each data point $\boldsymbol{x}_p$ of that file. Print the results in the console.

(e) Generate the following figure by plotting all sample points and the computed statistical quantities using the function `ax_ellipse.py`. You can find the file in the Moodle course. The plot should also include the data points $\boldsymbol{x}_p$.



(f) What is the advantage of using Mahanalobis distance over Euclidean distance?

3. Exercise: Kullback–Leibler (KL) divergence

KL divergence is a type of statistical distance. It is a measure of how one probability distribution P is different from a second, reference probability distribution Q. For discrete probability distributions P and Q, defined on the same sample space $\mathcal{X}$, the entropy relative from Q to P is defined to be

$$D_{\mathrm{KL}}(P||Q) = \sum_{x\in\mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right). \tag{4}$$

(a) What are the mathematical properties of KL divergence?

(b) Let two discrete probability distributions be $p = [0.05, 0.20, 0.40, 0.20, 0.15]^{\mathrm{T}}$ and $q = [0.10, 0.60, 0.05, 0.15, 0.10]^{\mathrm{T}}$, where each number in the array represents the probability of an event e happening. Compute the KL divergence $D_{\mathrm{KL}}(p||q)$ and $D_{\mathrm{KL}}(q||p)$.

(c) What is the analytic formula for the KL divergence $D_{\mathrm{KL}}(P||Q)$ when comparing two univariate Gaussian distributions $P = \mathcal{N}(\mu_p, \sigma_p^2)$ and $Q = \mathcal{N}(\mu_q, \sigma_q^2)$?

Note that $\mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and $\log(\frac{a}{b}) = \log(a) - \log(b)$.