CS-E3210- Machine Learning Basic Principles
Home Assignment 1 - "Introduction"

Your solutions to the following problems should be submitted as one single pdf which does not contain any personal information (student ID or name). The only rule for the layout of your submission is that each problem has to correspond to one single page, which has to include the problem statement on top and your solution below. You are welcome to use the LaTeX-file underlying this pdf, available under `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`, and fill in your solutions there.

Problem 1: Let The Data Speak - I

In the folder "Webcam" at `https://version.aalto.fi/gitlab/junga1/MLBP2017Public` you will find $N = 7$ webcam snapshots $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ with filename "shot??.jpg". Import these snapshots into your favourite programming environment (Matlab, Python, etc.) and determine for each snapshot $\mathbf{z}^{(i)}$ its greenness $x_g^{(i)}$ and redness $x_r^{(i)}$ by summing the green and red intensities over all image pixels (cf. `https://en.wikipedia.org/wiki/RGB_color_model`). Produce a scatter plot (cf. `https://en.wikipedia.org/wiki/Scatter_plot`) with the points $\mathbf{x}^{(i)} = (x_r^{(i)}, x_g^{(i)})^T \in \mathbb{R}^2$, for $i = 1, \ldots, N$. Do not forget to label the axes of your plot.

Answer.

Images	**Red**	**Green**
shot1	65069350	65398449
shot2	64797631	64619047
shot3	62725984	63548262
shot4	58760119	58666604
shot5	58960696	57416785
shot6	60470606	61338129
shot7	60728760	59001882

Table 1: Green and red pixel intensities for each image.

1.

Figure 1: Scatter plot - Redness vs Greenness

Problem 2: Let The Data Speak - II

Familiarize yourself with random number generation in your favourite programming environment (Matlab, Python, etc.). In particular, try to generate a data set $\{\mathbf{z}^{(i)}\}_{i=1}^{N}$ containing $N = 100$ vectors $\mathbf{z}^{(i)} \in \mathbb{R}^{10}$, which are drawn from (i.i.d. realizations of) a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with zero mean and covariance matrix being the identity matrix $\mathbf{I}$. For each data point $\mathbf{z}^{(i)}$, compute the two features

$$x_1^{(i)} = \mathbf{u}^T \mathbf{z}^{(i)}, \text{ and } x_2^{(i)} = \mathbf{v}^T \mathbf{z}^{(i)}, \tag{1}$$

with the vectors $\mathbf{u} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{10}$ and $\mathbf{v} = (9/10, 1/10, 0, \ldots, 0)^T \in \mathbb{R}^{10}$. Produce a scatter plot (cf. `https://en.wikipedia.org/wiki/Scatter_plot`) with the points $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)})^T \in \mathbb{R}^2$, for $i = 1, \ldots, N$. Do not forget to label the axes of your plot.

Answer. 2.

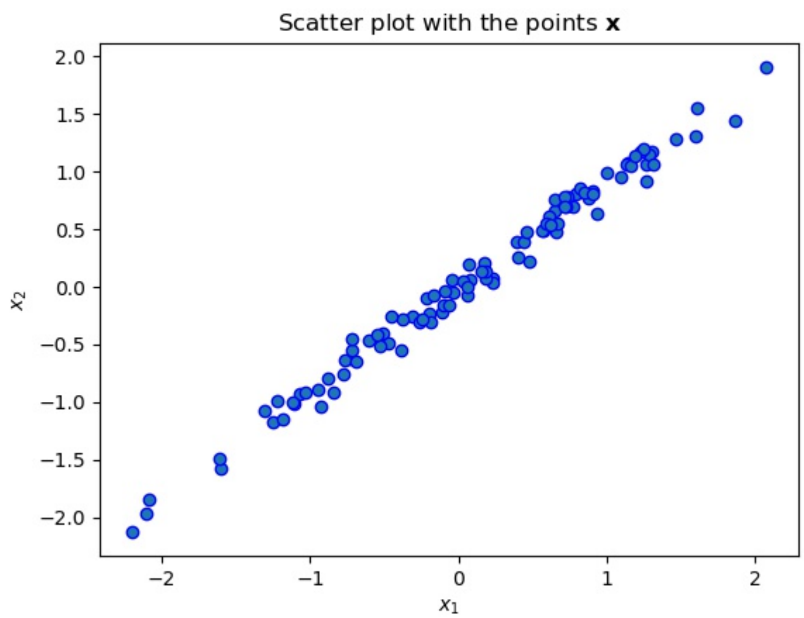

Figure 2: Scatter plot - x_1 vs x_2

Problem 3: Statistician's Viewpoint

Consider you are provided a spreadsheet whose rows contain the data points $\mathbf{z}^{(i)} = (i, y^{(i)})$, with row index $i = 1, \ldots, N$. A statistician might be interested in studying how to model the data using a probabilistic model, e.g.,

$$y^{(i)} = \mu + \sigma e^{(i)} \tag{2}$$

where $e^{(i)}$ are i.i.d. standard normal random variables, i.e., $e^{(i)} \sim \mathcal{N}(0, 1)$.

- Which choice for μ best fits the observed data?
- Given the optimum choice for μ, what would be the best guess for $y^{(N+1)}$?
- Can we somehow quantify the uncertainty in this prediction?

Answer.

The μ value that would best fit the observed data is:
Based on the Maximum Likelihood Estimation method,

$$y^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu_{ML} = arg\,\underset{\mu}{max}\, p(Y \mid \mu) = arg\,\underset{\mu}{max} \prod_{i=1}^{N} p(y^{(i)} \mid \mu)$$
$$\tfrac{dp(Y|\mu)}{d\mu} = 0$$

$$\mu = \bar{y} = \frac{\sum\limits_{i=1}^{N} y^{(i)}}{N} \tag{3}$$

The best prediction for $y^{(N+1)}$ would be:
Based on knowledge about posterior predivtive distributions,

$$y^{(N+1)} = \bar{y} = \frac{\sum\limits_{i=1}^{N} y^{(i)}}{N} \tag{4}$$

The uncertainty of the prediction can be quantified as:
Based on the Mean Square Error and Maximum Likelihood Estimation methods, and we know that $\bar{y}$ is an unbiased estimator,

$$\begin{aligned} MSE =& Bias(\bar{y})^2 + Var(\bar{y}) \\ =& 0 + \frac{\sigma^2}{N} \\ =& \frac{\sigma^2}{N} \\ SE_{\bar{y}} =& \frac{\sigma}{\sqrt{N}} \end{aligned} \tag{5}$$

Problem 4: Three Random Variables

Consider the following table which indicates the presence of a particular property ('A', 'B' or 'C') for a number of items (each item corresponds to one row).

A	B	C
1	0	1
1	1	0
1	0	1
1	1	0

- Can we predict if an item has property 'B' if we know the presence of property 'C' ?
- Can we predict if an item has property 'A' if we know the presence of property 'C' ?

Answer.

$P(\bar{B} \mid C) = 1, P(B \mid C) = 0, P(\bar{B} \mid C) + P(B \mid C) = 1$
$P(B \mid \bar{C}) = 1, P(\bar{B} \mid \bar{C}) = 0, P(B \mid \bar{C}) + P(\bar{B} \mid \bar{C}) = 1$
We can predict if an item has property 'B' if we know the presence of property 'C'. If an item has property 'C', we can predict that it does not have property 'B'; and if it does not have 'C', it will have 'B'.

$P(\bar{A} \mid C) = 0, P(A \mid C) = 1, P(\bar{A} \mid C) + P(A \mid C) = 1$
$P(A \mid \bar{C}) = 1, P(\bar{A} \mid \bar{C}) = 0, P(A \mid \bar{C}) + P(\bar{A} \mid \bar{C}) = 1$
We can predict if an item has property 'A' if we know the presence of property 'C'. The item samples show that an item always has property 'A', no matter whether it has 'C' or not.

(If the size of samples is big enough, we can use fisher's exact test (or chi-square test) to do correspondence analysis.)

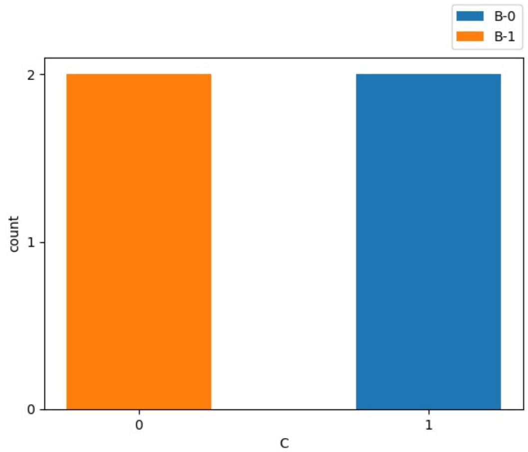

Figure 3: Stacked bar C-B

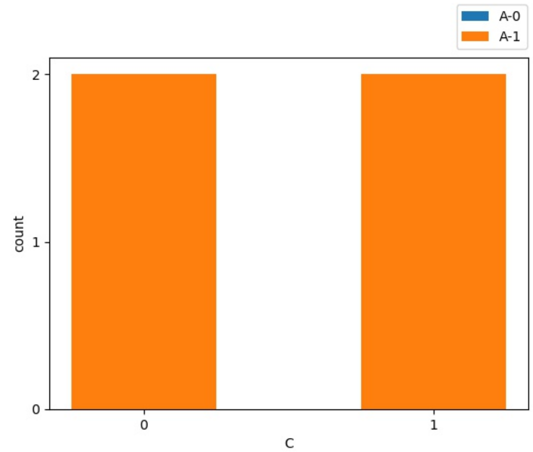

Figure 4: Stacked bar C-A

Problem 5: Expectations

Consider a d-dimensional Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a random variable $e \sim \mathcal{N}(0, \sigma^2)$. For a fixed (non-random) vector $\mathbf{w}_0 \in \mathbb{R}^d$, we construct the random variable $y = \mathbf{w}_0^T \mathbf{x} + e$. Now consider another arbitrary (non-random) vector $\mathbf{w} \in \mathbb{R}^d$. Find a closed-form expression for the expectation $\mathsf{E}[(y - \mathbf{w}^T \mathbf{x})^2]$ in terms of the variance σ^2 and the vectors $\mathbf{w}, \mathbf{w}_0$.

Answer.

$$\begin{aligned}
\mathsf{E}\{(y - \mathbf{w}^T\mathbf{x})^2\} &= \mathsf{E}\{((\mathbf{w}_0^T - \mathbf{w}^T)\mathbf{x} + e) * ((\mathbf{w}_0^T - \mathbf{w}^T)\mathbf{x} + e)\} \\
&= \mathsf{E}\{(\mathbf{w}_0^T - \mathbf{w}^T)^2\mathbf{x}^2 + 2e(\mathbf{w}_0^T - \mathbf{w}^T)\mathbf{x} + e^2\} \\
&= (\mathbf{w}_0^T - \mathbf{w}^T)^2\mathsf{E}[\mathbf{x}^2] + 2(\mathbf{w}_0^T - \mathbf{w}^T)\mathsf{E}[e\mathbf{x}] + \mathsf{E}[e^2]
\end{aligned}$$

$$\begin{gathered}
\mathsf{E}[\mathbf{x}^2] = Var(\mathbf{x}) + \mathsf{E}[\mathbf{x}]^2 = [1, 1, ...1]^T \in \mathbb{R}^d \\
\mathsf{E}[e\mathbf{x}] = \tfrac{\mu_e \sigma_{\mathbf{x}}^2 + \mu_{\mathbf{x}} \sigma_e^2}{\sigma_e^2 + \sigma_{\mathbf{x}}^2} = 0 \\
\mathsf{E}[e^2] = \mathsf{E}[(e - \mathsf{E}[e]) * (e - \mathsf{E}[e])] = \sigma^2
\end{gathered}$$

$$\begin{aligned}
\mathsf{E}\{(y - \mathbf{w}^T\mathbf{x})^2\} &= (\mathbf{w}_0^T - \mathbf{w}^T)^2[1, 1, ...1]^T + \sigma^2 \\
&= (\mathbf{w}_0 - \mathbf{w})^T(\mathbf{w}_0 - \mathbf{w}) + \sigma^2
\end{aligned}$$