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

We would like to thank Karanko Pihla and Matilde Tristany Farinha for their support in preparing the reference solution.

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.

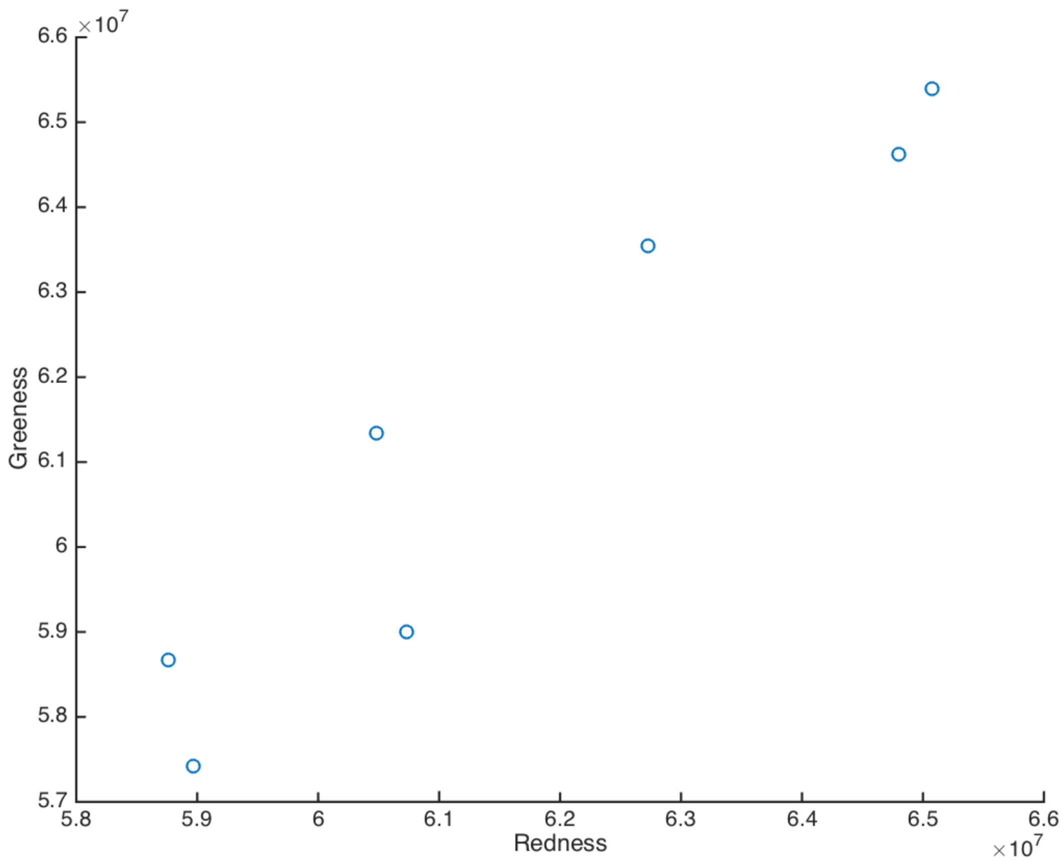

Figure 1: Scatter plot of the greenness $x_g^{(i)}$ and redness $x_r^{(i)}$

From Fig. 1, we conclude that the redness seems well-correlated with the greenness. In particular, if an image has large redness, it also has large 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.

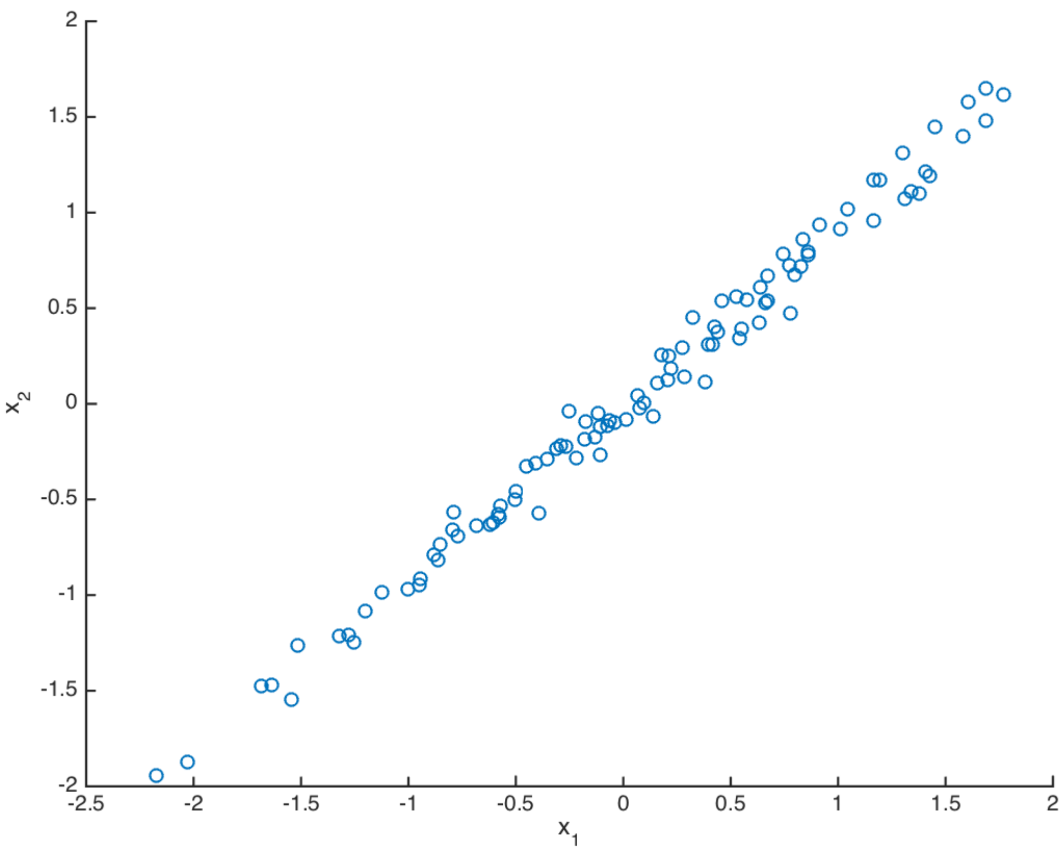

Figure 2: Scatter plot of $x_1^{(i)}$ and $x_2^{(i)}$

From Fig. 2, we obtain that the features $x_1^{(i)}$ and $x_2^{(i)}$ are strongly correlated.

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.,

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

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 likelihood of the observed data can be written as

$$\mathcal{L}(\mu, \sigma, y^{(1)}, \ldots, y^{(N)}) = \prod_{i=1}^{N} f(y^{(i)}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\left(y^{(i)} - \mu\right)^2}{2\sigma^2}\right). \tag{3}$$

  A principled approach of estimating the parameters from observed data is maximum likelihood, or equivalently maximum log-likelihood, i.e., $(\hat{\mu}, \hat{\sigma}) = \arg\max_{\mu,\sigma} l(\mu, \sigma)$, with

$$l(\mu, \sigma) = \log \mathcal{L}(\mu, \sigma, y^{(1)}, \ldots, y^{(N)}) \overset{(3)}{=} -(N/2)\log(2\pi) - N\log\sigma - \sum_{i=1}^{N} \frac{\left(y^{(i)} - \mu\right)^2}{2\sigma^2}. \tag{4}$$

  The partial derivatives of $l(\mu, \sigma)$ w.r.t. μ and σ are

$$\frac{\partial}{\partial\mu} l(\mu, \sigma) \overset{(4)}{=} \sum_{i=1}^{N} \frac{y^{(i)} - \mu}{\sigma^2}; \qquad \frac{\partial}{\partial\sigma} l(\mu, \sigma) \overset{(4)}{=} -N/\sigma + \sum_{i=1}^{N} \frac{(y^{(i)} - \mu)^2}{\sigma^3}.$$

  By the optimality condition, we have $\sum_{i=1}^{N} \frac{y^{(i)} - \hat{\mu}}{\hat{\sigma}^2} = 0$. Therefore,

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} y^{(i)}. \tag{5}$$

  Moreover, $\hat{\sigma}$ satisfies $-N/\hat{\sigma} + \sum_{i=1}^{N} \frac{(y^{(i)} - \hat{\mu})^2}{\hat{\sigma}^3} = 0$. Hence,

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \hat{\mu})^2. \tag{6}$$

- Given the parameters $\hat{\mu}, \hat{\sigma}$, the best guess for $y^{(N+1)}$ is the maximum of $\mathcal{N}(\hat{\mu}, \hat{\sigma})$, i.e., $\hat{\mu}$ (cf. (5)).
- A reasonable measure for the uncertainty is the variance of the distribution, i.e., $\hat{\sigma}^2$ (cf. (6)).

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.

- **Yes**. From the table, we obtain that the property B is complementary to the property C, i.e., B = 1 - C. Therefore, given C, we can predict B by setting B = 1 - C.
- **Yes**. From the table, we obtain that A is always 1. Therefore, we can predict A by setting A=1.

Problem 5: Expectations

Consider a d-dimensional Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a random variable $e \sim \mathcal{N}(0, \sigma^2)$ which is independent of $\mathbf{x}$. Given an arbitrary 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 $\mathbb{E}[(y - \mathbf{w}^T \mathbf{x})^2]$ in terms of the variance σ^2 and the vectors $\mathbf{w}, \mathbf{w}_0$.

Answer.

Since $y = \mathbf{w}_0^T \mathbf{x} + e$, using the linearity property of the expectation $\mathbb{E}[\cdot]$, we have

$$\begin{aligned}
\mathbb{E}[(y - \mathbf{w}^T \mathbf{x})^2] &= \mathbb{E}[(\mathbf{w}_0^T \mathbf{x} - \mathbf{w}^T \mathbf{x} + e)^2] \\
&= \mathbb{E}[((\mathbf{w}_0 - \mathbf{w})^T \mathbf{x} + e)^2] \\
&\overset{(a)}{=} \mathbb{E}[((\mathbf{w}_0 - \mathbf{w})^T \mathbf{x})^2] + \mathbb{E}[e^2] + 2\mathbb{E}[(\mathbf{w}_0 - \mathbf{w})^T \mathbf{x} e] \\
&\overset{(b)}{=} \mathbb{E}[(\mathbf{w}_0 - \mathbf{w})^T \mathbf{x}\mathbf{x}^T (\mathbf{w}_0 - \mathbf{w})] + \mathbb{E}[e^2] + 2(\mathbf{w}_0^T - \mathbf{w}^T)\mathbb{E}[\mathbf{x}]\mathbb{E}[e] \\
&\overset{(c)}{=} (\mathbf{w}_0 - \mathbf{w})^T \mathbb{E}[\mathbf{x}\mathbf{x}^T](\mathbf{w}_0 - \mathbf{w}) + \mathbb{E}[e^2] \\
&\overset{(d)}{=} (\mathbf{w}_0 - \mathbf{w})^T (\mathbf{w}_0 - \mathbf{w}) + \sigma^2 \\
&= \|\mathbf{w}_0 - \mathbf{w})\|^2 + \sigma^2,
\end{aligned}$$

where (a) follows the linearity property of $\mathbb{E}[\cdot]$; (b) is due to the fact that $\mathbf{x}$ and e are independent; (c) from $\mathbb{E}[e] = 0$; and (d) from the assumptions $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mathbf{I}$ and $\mathbb{E}[e^2] = \sigma^2$.