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.

The images are rode using SciPy and for each image the sum of every channel is calculated. The scatter plot sum of red channel intensities versus the green channel intensities are shown in figure 1 . According to figure 1 the red and green channels are clearly positively correlated.

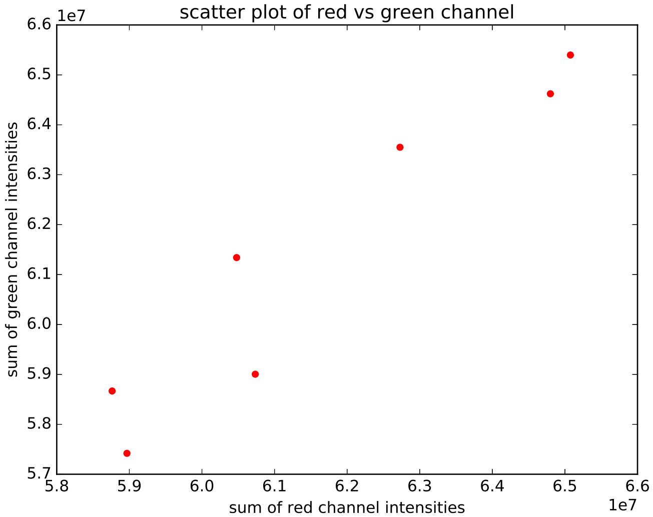

Figure 1: Scatter plot of sum of red channel intensities vs sum of green channel intensities.

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.

Because the covariance-matrix is diagonal the random vectors can be directly sampled from the one-dimensional gaussian with variance $\sigma^2 = \frac{1}{2}$. One hundread random vectors of length 10 are generated and for each vector $\mathbf{z}^{(i)}$ scalars $x_1^{(i)} = u^T z^{(i)}$ and $x_2^{(i)} = v^T z^{(i)}$ are generated and stored. The resulting scatter plot $x_1^{(i)}$ vs $x_2^{(i)}$ is shown in figure 2 . Again theres is a clear positive correlation between variables $x_1^{(i)} = u^T z^{(i)}$ and $x_2^{(i)} = v^T z^{(i)}$

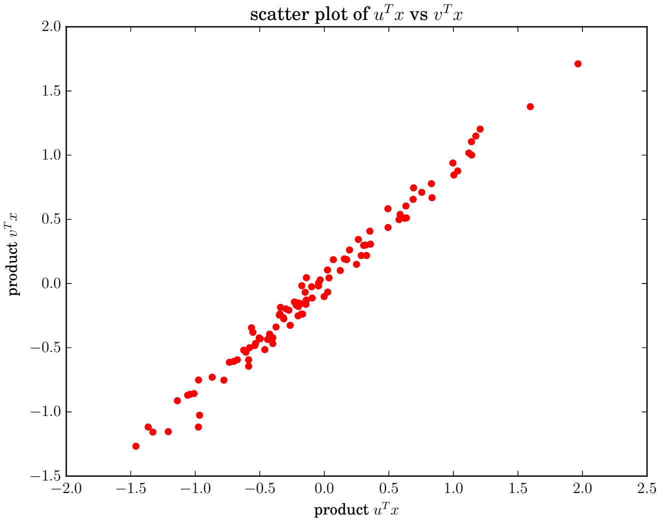

Figure 2: Scatter plot of $x_1^{(i)}$ vs $x_2^{(i)}$.

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.
If $e^{(i)}$ is a normally distributed random variable $e^{(i)} \sim \mathcal{N}(0, 1)$ then $y = \mu + \sigma e^{(i)}$ is also a normally distributed random variable $y \sim \mathcal{N}(\mu, \sigma)$.
(a) The optimal μ can be found by maximizing the log-likelihood $\mathcal{L}(\{\mu, \sigma\}, \mathbf{x})$ (assuming independent data points)

$$\begin{aligned}
\log\mathcal{L}(\mu, \sigma|\mathbf{x}) &= \log \prod_i^N P(x_i|\mu, \sigma) \\
&= \sum_i^N \log P(x_i|\mu, \sigma) && (3) \\
&= \sum_i^N \left(-\frac{1}{2}\log(2\pi\sigma) - \frac{(x_i - \mu)^2}{2\sigma^2} \right). && (4)
\end{aligned}$$

The value of μ that maximizes the log-likelihood is obtained by setting the derivative with respect to μ to zero

$$\frac{\partial \log\mathcal{L}(\mu, \sigma|\mathbf{x})}{\partial \mu} = \sum_i^N \frac{(x_i - \mu)}{\sigma^2}, \tag{5}$$

and setting this to zero give $\mu = \frac{1}{N}\sum_{i=1}^N x_i$, i.e., the optimal μ is simply the average of the datapoints.
(b) The best quess for the next data-point y^{i+1} is the expectation value of the distribution, which for the normal distribution $y \sim \mathcal{N}(\mu, \sigma)$ is simply the mean value μ.
(c). The standard error of the mean $\frac{\mu}{\sqrt(N)}$ can be used to estimate the error of the guess μ.

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.
Based on the data it seems that properties B and C are complementary to each other, *i.e.*, if B does not exist then C exist and if B exist then C does not exist and vice versa. Thus we can predict that B is always opposite to C. However property A seems to be independent of property C and seems to be always one. Thus we can predict that A is one regardless of C.

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.

The distribution's, $\mathcal{N}(0, I)$, covariance matrix is an identity matrix and thus the distribution can be expressed as a product of d one-dimensional gaussian distributions sampled from $\mathcal{N}(0, \frac{1}{2})$ and thus each element of vector $\mathbf{x}$ is normally distributed with $\mu = 0$ and $\sigma^2 = \frac{1}{2}$. Let y_1 be $y_1 = \mathbf{w_0^T} \mathbf{x} + e$ and $y_2 = \mathbf{w^T} \mathbf{x}$. Because y_1 and y_2 are linear combinations of gaussian random variables it hold for both y_1 and y_2 that $\mu_{y_1} = 0$ and $\mu_{y_2} = 0$. For the variances it holds that $Var(y_1) = \frac{1}{2} \sum_{i=1}^{d} (w_0^i)^2 + \sigma^2$ and $Var(y_2) = \frac{1}{2} \sum_{i=1}^{d} (w^i)^2$. Now the expectation $E[(y_1 - y_2)^2]$ can be written as

$$\begin{aligned}
E[(y_1 - y_2)^2] &= E[y_1^2] - 2E[y_1 y_2] + E[y_2^2] && (6)\\
&= Var(y_1) + E[y_1]^2 + 2E[y_1]E[y_2] + Var(y_2) + E[y_2]^2 && (7)\\
&= Var(y_1) + Var(y_2) && (8)\\
&= \frac{1}{2} \sum_{i=1}^{d} \left((w_0^i)^2 + (w^i)^2 \right) + \sigma^2. && (9)
\end{aligned}$$

In above the identities $Var(\mathbf{X}) = E[X^2] - E[X]^2$ and $E[y_1] = \mu_{y1} = E[y_2] = \mu_{y2} = 0$ were used.