Machine Learning (708.064)
Assignment 1 - Gaussians and Baysian regression

March 22, 2019

Lecturers: G. Bellec and A. Subramoney
Tutors: L. Lindner (lydia.lindner@student.tugraz.at), D. Narnhofer (narnhofer@student.tugraz.at)
Submission: Each group should submit on the Teach Center:

- a report in pdf (hand-written or tex file) **(please do not put the pdf into a zip file)**,
- and the code producing the figures used in the report.

The students should use python 3.6 or higher and up to date versions of numpy and matplotlib libraries. It is forbidden to use any other libraries such as scipy or scikit-learn.
Deadline: April 19^{th}, 2019 at 23:55h.

1 Probabilistic calculus on Gaussian distributions (12.5 points)

1. **Basic probability calculus in one dimension** We consider a probability distribution defined by the probability density function (pdf) $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. This probability distribution is often referred as a Gaussian distribution of mean μ and variance σ^2 and this pdf is often denoted by $\mathcal{N}(x|\mu, \sigma)$.

   a) Show that the entropy of the distribution is equal to $H(p) = \frac{1}{2}\log\left(2\pi e\sigma^2\right)$ (hint: use the expression of the mean and variance of a Gaussian to avoid compute too many integrals).

   b) Given x and y two independent random variables (not necessarily Gaussian in this question) of means μ_x and μ_y and variances σ_x^2 and σ_y^2. Compute the mean and variance of the random variable $z = x + y$.

   c) Compute analytically the cross-entropy $H(p_x, p_y)$ between two Gaussian distributions of means μ_x, μ_y and variance σ_x^2, σ_y^2.

   d) Recall the relationship between the Kullback Leibler divergence $D_{KL}\left(p \mid\mid q\right)$, the cross entropy $H(p,q)$ and the entropy $H(q)$ for any pair of distributions p and q, and finally compute the Kullback Leibler divergence $D_{KL}\left(p_x \mid\mid p_y\right)$ between the two Gaussian distributions defined above.

2. **Conditional Gaussians** The goal of this exercise is to show that when a Gaussian variable $z \in \mathbb{R}^2$ is partially conditioned on some coordinates, the conditioned variable is also distributed as a Gaussian variable. Let's consider the joint distribution defined by the mean $\mu \in \mathbb{R}^2$ and covariance $\Sigma \in \mathbb{R}^{2\times 2}$, we write the scalar coordinate of the vector z as x and y, and we condition the random variable on a fixed value on the second coordinate y. Accordingly, we define notations:

$$z = \begin{bmatrix} x \\ y \end{bmatrix}\ ,\ \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{1}$$

To simplify notations further we also define the *precision matrix* $\Lambda \overset{\text{def}}{=} \Sigma^{-1}$, and it can be written as:

$$\Lambda = \begin{bmatrix} \lambda_{xx} & \lambda_{xy} \\ \lambda_{yx} & \lambda_{yy} \end{bmatrix} \tag{2}$$

a) compute analytically $p(y)$ using the definition of a marginal probability distribution,

b) compute $p(x|y)$ using the product rule of probability,

c) and conclude that the conditional probability distribution $p(x|y)$ is a Gaussian distribution and give its mean $\mu_{x|y}$ and variance $\sigma_{x|y}$ as a function of the variables defined above.

3. **Sampling and visualization** of Gaussian variables,

   a) **Sampling a 2D Gaussian variable** Write a function that transforms samples generated from Gaussian of mean 0 and co-variance identity *numpy.random.randn* into samples of a the Gaussian distribution $p(z)$ with mean and co-variance:

   $$\mu = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 3 & 0.75 \\ 0.75 & 1.5 \end{bmatrix} \tag{3}$$

   Explain in two sentences how did you generate these points. Report an empirical estimation of the mean and the co-variance matrix. Show a scatter plot with 200 datapoints.

   b) **Visualization of marginal and conditional distribution** Using as previously the same notation x and y for the two coordinates of $z = [x, y]^T$, sample points from the marginal distribution $p(x)$ using question 3.a and display its pdf using 2.a. Display the distribution $p(x|y = -0.1)$ using question 2.b.

2 Linear regression and Bayesian linear regression (12.5 points)

1. **Different probabilistic views of regression problems** Consider a set of N data points $\mathbf{x}_i \in \mathbb{R}^D$ and associated targets $t_i \in \mathbb{R}$. We assume that the target are generated according to a linear model, such that there exist optimal parameters $\mathbf{w}$ for which the targets t_i are noisy observations of the underlying linear transformation $y(\mathbf{x}_i, \mathbf{w}) = \mathbf{x}_i^T \mathbf{w}$ with additive noise ϵ. Mathematically this means:

   $$t_i = y(\mathbf{x}_i, \mathbf{w}) + \epsilon_i \tag{4}$$

   a) Assuming that ϵ_i are independent Gaussian random variables of mean 0 and variance σ^2, show that the log-likelihood of the data $\log p(t_1, \ldots t_N | \mathbf{x}_1, \ldots \mathbf{x}_N, \mathbf{w}, \sigma)$ is given by:

   $$\log p(t_1, \ldots t_N | \mathbf{x}_1, \ldots \mathbf{x}_N, \mathbf{w}, \sigma) = \alpha E + \beta, \tag{5}$$

   where $E = \frac{1}{2N} \sum_{i=1}^{N} (t_i - w^T \mathbf{x}_i)^2$ is the mean squared error. Report the constants α and β as functions of N and σ.

   b) Compute the parameter σ^{ML} which maximizes the log-likelihood with respect to σ. In one sentence, interpret what it means in terms of uncertainty. In a second sentence describe how can it be beneficial to use the maximum likelihood interpretation instead of directly solving the least square problem: $\min_{\mathbf{w}} E$

2. **Bayesian regression** In Baysian regression, instead of looking for the parameters that maximize the likelihood $p((\mathbf{x}_1, t_1), \ldots (\mathbf{x}_N, t_N) | \mathbf{w})$ we look for “the maximum a posteriori” which refer to the parameters that maximize the posterior distribution $p\left(\mathbf{w} | (\mathbf{x}_1, t_1), \ldots (\mathbf{x}_N, t_N)\right)$. The posterior distribution also includes a hypothesis on the preferred values of the parameters $\mathbf{w}$, this hypothesis is formalized using a prior distribution $p(\mathbf{w})$ on the parameters of the form $p(\mathbf{w}) = \mathcal{N}(\mu_{\mathbf{w}}, \frac{1}{\alpha}\mathbf{I})$ ($\mathbf{I}$ is the identity matrix of $\mathbb{R}^{D \times D}$). To simplify notation we define $\mathbf{X} \in \mathbb{R}^{N \times D}$ as the matrix formed by the vertical concatenation of the vectors $\mathbf{x}_i^T$ and $\mathbf{t} \in \mathbb{R}^N$ as the vertical concatenation of all the targets t^i, and the posterior can hence be written as $p\left(\mathbf{w} | \mathbf{X}, \mathbf{t}\right)$

a) Show that the log of the posterior distribution can be written as:

$$\log p\left(\mathbf{w}|\mathbf{X},\mathbf{t}\right) = -\frac{1}{2\sigma^2}||\mathbf{t}-\mathbf{X}\mathbf{w}||^2 - \frac{\alpha}{2}||\mathbf{w}-\mu_{\mathbf{w}}||^2 + \text{const} \tag{6}$$

where “const” refers to a constant term that does not depend on the parameters $\mathbf{w}$. Point out which terms in the resulting equation arise from the log-likelihood and the log-prior respectively (the log-prior is defined here as the log of the pdf of the prior distribution).

b) Show that the maximum a posteriori are given by the expression:

$$\mathbf{w}^{MAP} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\lambda\mu_{\mathbf{w}} + \mathbf{X}^T\mathbf{t}\right), \tag{7}$$

where λ is a positive constant, and justify why one can invert the matrix $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$.

c) Show independently that the posterior distribution $p\left(\mathbf{w}|\mathbf{X},\mathbf{t}\right)$ is a Gaussian distribution. Compute its mean and co-variance matrix. Compare the mean of the posterior distribution and the maximum a posteriori, is it expected to find a relationship between them ?

3. **Implementations** Generate a one-dimensional dataset where $x_1, \dots x_N$ are distributed uniformly in $[-1,1]$ and the targets are given by the formula $t^i = w_0 + x_i w_1 + \epsilon_i$ where ϵ_i is sampled from a Gaussian of mean 0 and standard deviation $\sigma = 1$ and the true parameters are given by $w_0 = 1.26$ and $w_1 = 1.77$. We consider that the true parameter vector $\mathbf{w} = [w_0, w_1]$ is unknown, but we assume the prior knowledge that is drawn with a Gaussian of mean $\mu_{\mathbf{w}} = [1.3, 1.7]$ and and co-variance $\frac{1}{\alpha}\mathbf{I}$ with α=10.

Using your results in section 1 and 2, write python functions to compute:

- the maximum a posteriori $\mathbf{w}^{MAP}$,
- the maximum likelihood estimator $\mathbf{w}^{ML}$,
- the mean and the co-variance of the posterior distribution $p(w_0, w_1|\mathbf{X},\mathbf{t})$,

Display a scatter plot of the data-points (x_i, t_i) with the true line generating the data in green, the line obtained from the maximum likelihood in blue, and the line obtained from the maximum a posteriori in red. Do not forget to put a color legend and annotate all axis. Show 3 subplots of this kind for 3, 20 and 100 data points respectively.

Display a grid plot with 9 subplots where each column represents from left to right: the prior $p(w_0, w_1)$, the likelihood $p(\mathbf{X},\mathbf{t}|w_0,w_1)$ and the posterior $p(w_0,w_1|\mathbf{X},\mathbf{t})$ and from top to bottom the results obtained with 3, 20 and 100 data points. Each plot should represent w_0 and w_1 on the x and y-axis respectively. Use the same range for the w_0 and w_1 axes in all plots and draw on each plot a cross at the position of the true parameters, the maximum likelihood and the maximum a posteriori. To represent distributions of w_0 and w_1 we recommend to use contours of iso probability or colormaps showing the pdf itself.