## CS-E4820 Machine Learning: Advanced Probabilistic Methods (spring 2021)

Pekka Marttinen, Santosh Hiremath, Tianyu Cui, Yogesh Kumar, Zheyang Shen, Alexander Aushev, Khaoula El Mekkaoui, Shaoxiong Ji, Alexander Nikitin, Sebastiaan De Peuter, Joakim Järvinen.

### Exercise 3, due on Tuesday February 9 at 23:00.

### Problem 1: Poisson-Gamma

Suppose you have $N$ i.i.d. observations $\mathbf{x} = \{x_i\}_{i=1}^N$ from a Poisson($\lambda$) distribution with a rate parameter $\lambda$ that has a conjugate prior

$$\lambda \sim \text{Gamma}(a, b)$$

with the shape and rate hyperparameters $a$ and $b$. Derive the posterior distribution $\lambda|\mathbf{x}$.

Write your solutions in LateX or attach a picture in the answer cell provided below. You can add a picture using the command `!(imagename_in_the_folder.jpg)`. Latex in here works similarly as you would write it normally! You can use some of the definitions from the exercise description as a reference. The list of valid Latex commands in Jypyter notebook can be found here: http://www.onemathematicalcat.org/MathJaxDocumentation/TeXSyntax.htm

#### Solution

We have

$$p(x_i|\lambda) = Poisson(x_i|\lambda) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}$$
$$p(\lambda) = Gamma(\lambda|a, b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}.$$

Now using Bayes' rule to obtain the posterior distribution, we get

$$\begin{aligned} p(\lambda|x) &\propto \prod_{i=1}^{N} p(x_i|\lambda)p(\lambda) \\ &\propto \prod_{i=1}^{N} \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} \\ &\propto \lambda^{a+\sum_i x_i - 1}e^{-(b+N)\lambda} \\ &\propto Gamma(\lambda|a + \sum_{i=1}^{N} x_i, b + N). \end{aligned}$$

### Problem 2: Multivariate Gaussian

Suppose we have $N$ i.i.d. observations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ from a multivariate Gaussian distribution

$$\mathbf{x}_i \mid \boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with unknown mean parameter $\boldsymbol{\mu}$ and a known covariance matrix $\boldsymbol{\Sigma}$. As prior information on the mean parameter we have

$$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{m_0}, \mathbf{S_0}).$$

**(a)** Derive the posterior distribution $p(\boldsymbol{\mu}|\mathbf{X})$ of the mean parameter $\boldsymbol{\mu}$. Write your solution in LateX or attach a picture of the solution in the cell below.

**(b)** Compare the Bayesian estimate (posterior mean) to the maximum likelihood estimate by generating $N = 10$ observations from the bivariate Gaussian

$$\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right).$$

For this you can use the Python function numpy.random.normal, making use of the fact that the elements of the bivariate random vectors are independent. In the Bayesian case, use the prior with $\mathbf{m_0} = [0, 0]^T$ and $\mathbf{S_0} = \left[\begin{smallmatrix}0.1 & 0\\0 & 0.1\end{smallmatrix}\right]$. Report both estimates. Is the Bayesian estimate closer to the true value $\boldsymbol{\mu} = [0, 0]^T$? Use the code template given below (after the answer cell) to complete your answer.

Write your solutions to **(a)** and **(b)** in LateX or attach a picture in the answer cell provided below.

**Solution**

**(a)** The posterior up to a normalizing constant is given by

$$\begin{aligned}
p(\boldsymbol{\mu}|\mathbf{X}) &\propto \exp\left(\sum_i \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \log \mathcal{N}(\boldsymbol{\mu}|\mathbf{m_0}, \mathbf{S_0})\right)\\
&\propto \exp\left(-\frac{1}{2}\sum_i (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \mathbf{m_0})^T \mathbf{S_0}^{-1}(\boldsymbol{\mu} - \mathbf{m_0})\right)\\
&\propto \exp\left(-\frac{1}{2}\sum_i (\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu}^T \mathbf{S_0}^{-1}\boldsymbol{\mu} - 2\mathbf{m_0}^T \mathbf{S_0}^{-1}\boldsymbol{\mu})\right)\\
&= \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(N\boldsymbol{\Sigma}^{-1} + \mathbf{S_0}^{-1})\boldsymbol{\mu} + (\boldsymbol{\Sigma}^{-1}\sum_i \mathbf{x}_i + \mathbf{S_0}^{-1}\mathbf{m_0})^T\boldsymbol{\mu}\right)\\
&= \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{S}^{-1}\boldsymbol{\mu} + \mathbf{b}^T\boldsymbol{\mu}\right)\\
&= \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{Sb})^T \mathbf{S}^{-1}(\boldsymbol{\mu} - \mathbf{Sb})\right).
\end{aligned}$$

In the third line, we have omitted all terms which do not include $\boldsymbol{\mu}$, and the last line was found by completing the square (see Barber: Bayesian Reasoning and Machine Learning, ch. 8.4.1). The posterior can now be recognized as a Gaussian $\mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, \mathbf{S})$ with parameters

$$\begin{aligned}
\mathbf{m} = \mathbf{Sb} &= (N\boldsymbol{\Sigma}^{-1} + \mathbf{S_0}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\sum_i \mathbf{x}_i + \mathbf{S_0}^{-1}\mathbf{m_0}) \qquad (2.1)\\
\mathbf{S} &= (N\boldsymbol{\Sigma}^{-1} + \mathbf{S_0}^{-1})^{-1}.
\end{aligned}$$

**(b)** In this case the Bayesian estimate is roughly half of the maximum likelihood mean estimate (i.e. the simple average of the values), as shown below:

[1]:
```
# template for 2(b)
import numpy as np
from numpy.linalg import inv

S0 = np.array([[0.1, 0],[0, 0.1]])
Sigma = np.array([[1, 0],[0, 1]])
N = 10

# Sample N bivariate normal vectors
# compute MLE and also the posterior mean solution

# x = ? #EXERCISE
# mle = ? #EXERCISE
# posterior_mean = ? #EXERCISE

## BEGIN SOLUTION
x = np.random.normal(size = (N, 2))
mle = np.mean(x, axis = 0)
posterior_mean = np.dot(inv(N*inv(Sigma) + inv(S0)), np.
↪dot(inv(Sigma),sum(x)))
### END SOLUTION

print(mle)
print(posterior_mean)
```

# 1 Problem 3: Posterior of regression weights

Suppose $y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$, for $i = 1, \ldots, n$, where $\epsilon_i \sim \mathcal{N}(0, \beta^{-1})$. Assume a prior

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}).$$

Use 'completing the square' to show that the posterior of $\mathbf{w}$ is given by $p(\mathbf{w} \mid \mathbf{y}, \mathbf{x}, \alpha, \beta) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S})$, where

$$\mathbf{S} = \left( \alpha\mathbf{I} + \beta \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1},$$
$$\mathbf{m} = \beta \mathbf{S} \sum_{i=1}^{n} y_i \mathbf{x}_i.$$

Write your solution in LateX or attach a picture of the solution in the cell below.

**Solution**

Denote $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]^T$, and note that the likelihood can be written as

$$p(\mathbf{y}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I}).$$

Then, starting with Bayes' rule we get

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{y},\mathbf{x},\alpha,\beta) &\propto p(\mathbf{y}|\mathbf{w},\mathbf{x},\beta)p(\mathbf{w}|\alpha) \\
&= \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\beta^{-1}\mathcal{I})\mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) \\
&\propto \exp\left(-\frac{1}{2}\beta(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) - \frac{1}{2}\alpha\mathbf{w}^T\mathbf{w}\right) \\
&\propto \exp\left(-\frac{1}{2}\beta\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \beta\mathbf{y}^T\mathbf{X}\mathbf{w} - \frac{1}{2}\alpha\mathbf{w}^T\mathbf{w}\right) \\
&= \exp\left(-\frac{1}{2}\mathbf{w}^T\underbrace{(\beta\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I})}_{=\mathbf{S}^{-1}}\mathbf{w} + \underbrace{\beta\mathbf{y}^T\mathbf{X}}_{=(\mathbf{S}^{-1}\mathbf{m})^T}\mathbf{w}\right) \\
&\propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m})^T\mathbf{S}^{-1}(\mathbf{w}-\mathbf{m})\right).
\end{aligned}
$$