Philipp Lintl
12152498
philipp.lintl@student.uva.nl

Homework Assignment 1

Machine Learning 2, 19/20

2019-09-08

As disclosed to the professor, I work remotely for the first three weeks and therefore have not collaborated with anyone. Group E.

Problem 1. Given: $\boldsymbol{x} \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x})$ and $\boldsymbol{z} \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu_z}, \boldsymbol{\Sigma_z})$. Random vector $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}$.

- Mean of y:

$$\mathbb{E}(y) = \mathbb{E}(\boldsymbol{x} + \boldsymbol{z}) = \mathbb{E}(\boldsymbol{x}) + \mathbb{E}(\boldsymbol{z}) = \boldsymbol{\mu_x} + \boldsymbol{\mu_z}$$

- Covariance of y:

$$\begin{aligned}
\mathrm{var}(\boldsymbol{y}) &= Cov_{\boldsymbol{y}}(\boldsymbol{y}, \boldsymbol{y}) = \mathbb{E}\left[(\mathbf{y} - \mathbb{E}(\boldsymbol{y}))(\mathbf{y} - \mathbb{E}(\boldsymbol{y}))^\top\right] \\
&= \mathbb{E}\left[(\boldsymbol{x} + \boldsymbol{z} - \mathbb{E}(\boldsymbol{x} + \boldsymbol{z}))(\boldsymbol{x} + \boldsymbol{z} - \mathbb{E}(\boldsymbol{x} + \boldsymbol{z}))^\top\right] \\
&= \mathbb{E}\left[(\boldsymbol{x} + \boldsymbol{z} - \boldsymbol{\mu_x} - \boldsymbol{\mu_z}))(\boldsymbol{x} + \boldsymbol{z} - \boldsymbol{\mu_x} - \boldsymbol{\mu_z}))^\top\right] \\
\text{(multiply out)} \quad &= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{x}\boldsymbol{z}^\top - \boldsymbol{x}\boldsymbol{\mu_x}^\top - \boldsymbol{x}\boldsymbol{\mu_z}^\top + \boldsymbol{z}\boldsymbol{x}^\top + \boldsymbol{z}\boldsymbol{z}^\top - \boldsymbol{z}\boldsymbol{\mu_x}^\top - \boldsymbol{z}\boldsymbol{\mu_z}^\top \\
&- \boldsymbol{\mu_x}\boldsymbol{x}^\top - \boldsymbol{\mu_x}\boldsymbol{z}^\top + \boldsymbol{\mu_x}\boldsymbol{\mu_x}^\top + \boldsymbol{\mu_x}\boldsymbol{\mu_z}^\top - \boldsymbol{\mu_z}\boldsymbol{x}^\top - \boldsymbol{\mu_z}\boldsymbol{z}^\top + \boldsymbol{\mu_z}\boldsymbol{\mu_z}^\top + \boldsymbol{\mu_z}\boldsymbol{\mu_x}^\top \\
(\boldsymbol{x}\boldsymbol{z}^\top = \boldsymbol{z}\boldsymbol{x}^\top) \quad &= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top + 2\boldsymbol{x}\boldsymbol{z}^\top - 2\boldsymbol{x}\boldsymbol{\mu_x}^\top - 2\boldsymbol{x}\boldsymbol{\mu_z}^\top + \boldsymbol{z}\boldsymbol{z}^\top - 2\boldsymbol{z}\boldsymbol{\mu_x}^\top - 2\boldsymbol{z}\boldsymbol{\mu_z}^\top + \boldsymbol{\mu_x}\boldsymbol{\mu_x}^\top + 2\boldsymbol{\mu_x}\boldsymbol{\mu_z}^\top + \boldsymbol{\mu_z}\boldsymbol{\mu_z}^\top) \\
\text{(Resort)} \quad &= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top + \boldsymbol{\mu_x}\boldsymbol{\mu_x}^\top - 2\boldsymbol{x}\boldsymbol{\mu_x}^\top) + \mathbb{E}(\boldsymbol{z}\boldsymbol{z}^\top + \boldsymbol{\mu_z}\boldsymbol{\mu_z}^\top - 2\boldsymbol{z}\boldsymbol{\mu_z}^\top) + \mathbb{E}(2\boldsymbol{x}\boldsymbol{z}^\top + 2\boldsymbol{\mu_x}\boldsymbol{\mu_z}^\top - 2\boldsymbol{\mu_x}\boldsymbol{z}^\top \\
&- 2\boldsymbol{x}\boldsymbol{\mu_z}^\top) \\
\text{(Binomial rules)} \quad &= \mathbb{E}\left[\left(\boldsymbol{x} - \boldsymbol{\mu_x}\right)\left(\boldsymbol{x} - \boldsymbol{\mu_x}\right)^\top\right] + \mathbb{E}\left[\left(\boldsymbol{z} - \boldsymbol{\mu_z}\right)\left(\boldsymbol{z} - \boldsymbol{\mu_z}\right)^\top\right] + 2\mathbb{E}\left[\left(\boldsymbol{x} - \boldsymbol{\mu_x}\right)\left(\boldsymbol{z} - \boldsymbol{\mu_z}\right)^\top\right] \\
&= \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_z} + 2\,\mathrm{cov}(\boldsymbol{x}, \boldsymbol{z})
\end{aligned}$$

- When x and z are independent, their covariances are zero, thus

$$\mathrm{var} = \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_z}$$

Problem 2. Given: $\boldsymbol{x} \in \mathbb{R}^D \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known Σ
a, $\mathcal{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_N})$, $\boldsymbol{x_i} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$

1. Likelihood of $p(\mathcal{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\overset{iid}{=} \prod_{i=1}^{N} p\left(x_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
&= \prod_{i=1}^{N} \mathcal{N}\left(x_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\
\text{(Multivariate Gaussian)} \quad &= \prod_{i=1}^{N} (2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(x_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(x_i - \boldsymbol{\mu}\right)\right] \\
\text{(product of exp)} \quad &= (2\pi)^{\frac{-D \cdot N}{2}} \det(\boldsymbol{\Sigma})^{\frac{-N}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(x_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(x_i - \boldsymbol{\mu}\right)\right]
\end{aligned}$$

2. The posterior $p\left(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ (no need to normalize the probability distribution by calculating the evidence).

$$\begin{aligned}
p\left(\boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) &\overset{\text{Bayes}}{=} \frac{p(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot p\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)}{p\left(\boldsymbol{X}|\boldsymbol{\Sigma}, \mu_0, \boldsymbol{\Sigma}_0\right)} \\
&= \frac{\prod_{i=1}^{N} \mathcal{N}\left(x_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot \mathcal{N}\left(\boldsymbol{\mu}|\mu_0, \boldsymbol{\Sigma}_0\right)}{p\left(\boldsymbol{X}|\boldsymbol{\Sigma}, \mu_0, \boldsymbol{\Sigma}_0\right)}
\end{aligned}$$

3. Show that $p\left(\boldsymbol{\mu}|\mathcal{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)$ is a Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right)$ and find the values of $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ (hint: use "completing the square")

$$
\begin{aligned}
p\left(\boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) &= \frac{\prod_{i=1}^{N} \mathcal{N}\left(x_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot \mathcal{N}\left(\boldsymbol{\mu}|\mu_0, \boldsymbol{\Sigma}_0\right)}{p\left(\boldsymbol{X}|\boldsymbol{\Sigma}, \mu_0, \boldsymbol{\Sigma}_0\right)} \\
\text{(intractable evidence)} \quad &\propto \prod_{i=1}^{N} \mathcal{N}\left(x_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \cdot \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right) \\
\text{(multiv. Gaussian)} \quad &= (2\pi)^{-\frac{D \cdot N}{2}} \det(\boldsymbol{\Sigma})^{-\frac{N}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(x_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(x_i - \boldsymbol{\mu}\right)\right] \\
&\cdot (2\pi)^{-\frac{D}{2}} \det\left(\boldsymbol{\Sigma}_0\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\right] \\
\text{(Resorting)} \quad &= \underbrace{(2\pi)^{\frac{-D(N+1)}{2}} \det(\boldsymbol{\Sigma})^{-\frac{N}{2}} \det\left(\boldsymbol{\Sigma}_0\right)^{-\frac{1}{2}}}_{=\text{ C: Normalizing Constant indep. of } \boldsymbol{\mu}} \\
&\cdot \exp\underbrace{\left[-\frac{1}{2}\sum_{i=1}^{N}\left(x_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(x_i - \boldsymbol{\mu}\right) - \frac{1}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\right]}_{:=EXPONENT}
\end{aligned}
$$

As hinted in the problem, we draw on the *Completing the square* operation, which relies on this equation (Bishop 2.71):

$$
-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \quad \text{const}
$$

According to it, the exponent of the general Gaussian can be rewritten as seen in the right side. As the term in our exponent is quadratic in $\boldsymbol{\mu}$, we look to rewrite the exponent according to the right side of 2.71. Thus by multiplying out, we arrive at

$$
\begin{aligned}
EXPONENT &= -\frac{1}{2}\left(\boldsymbol{\mu}^T N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\sum_{i}^{N} x_i + \sum_{i}^{N} x_i^T\boldsymbol{\Sigma}^{-1}x_i\right) - \frac{1}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) \\
&= -\frac{1}{2}\left(\boldsymbol{\mu}^T\left(N\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\sum_{i}^{N}\boldsymbol{x}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \underbrace{\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \sum_{i}^{N}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i}_{:=\text{ const}}\right) \\
&= -\frac{1}{2}\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\mu} + \boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} x_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) + \text{const}
\end{aligned}
$$

By multiplying, resorting and taking the terms independent of $\boldsymbol{\mu}$ as const, we arrive with the form required in Bishops formula. According to it, the covariance matrix is rather easily detected to be the term in between the brackets on the very left:

$$
\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \rightarrow \boldsymbol{\Sigma}_N = (\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1})^{-1}
$$

As the right side of 2.71 expects $\boldsymbol{\mu_N}$ to be multiplied to $\boldsymbol{\mu\Sigma^{-1}}$, we need to multiply the inverse of $\boldsymbol{\Sigma^{-1}}$, which is simply $\boldsymbol{\Sigma_N}$:

$$
\boldsymbol{\mu}_N = \left(\boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} x_i\right) = \boldsymbol{\Sigma_N}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N} x_i\right)
$$

4. Derive the MAP solution for $\boldsymbol{\mu}$: The MAP is obtained by deriving the posterior wrt to $\boldsymbol{\mu}$. As usual, we drawn on the log posterior, as the exponent thus evaporates and the differentiation is easier. So, we end up with $\mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu_N}, \boldsymbol{\Sigma_N}\right)$.

$$
\frac{\partial\mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu_N}, \boldsymbol{\Sigma_N}\right)}{\partial\boldsymbol{\mu}} \propto \frac{\partial\ln\mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu_N}, \boldsymbol{\Sigma_N}\right)}{\partial\boldsymbol{\mu}}
$$

Instead of writing all the terms included in C (Constants outside the exponent from above) and const., I focus on the terms that are dependent on $\boldsymbol{\mu}$, as we derive for it and therefore save unnecessary writing.

$$
\begin{aligned}
&\propto \frac{\partial \ln \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu_N}\right)^T \boldsymbol{\Sigma_N}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu_N}\right)\right]}{\partial \boldsymbol{\mu}} \\
&\overset{\text{(form of 2.71)}}{=} \frac{\partial -\frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu}+\boldsymbol{\mu^T}\Sigma_N^{-1}\boldsymbol{\mu_N}}{\partial \boldsymbol{\mu}} \\
&= -\boldsymbol{\mu^T}\boldsymbol{\Sigma_N^{-1}} + \left(\boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu_N}\right)^T
\end{aligned}
$$

Now, this term is set to zero and solved for $\boldsymbol{\mu}$ to arrive at the MAP solution for this parameter:

$$
\begin{aligned}
&-\boldsymbol{\mu^T}\boldsymbol{\Sigma_N^{-1}} + \left(\boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu_N}\right)^T = 0 \\
&\boldsymbol{\mu^T}\boldsymbol{\Sigma_N^{-1}} = \left(\boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu_N}\right)^T \\
&\boldsymbol{\Sigma_N^{-1}}^T\boldsymbol{\mu} = \boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu_N} \\
(\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma} \text{ if } \boldsymbol{\Sigma} \text{ symm. matrix}) \qquad &\boldsymbol{\mu} = \boldsymbol{\Sigma_N}\boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu_N} \\
&\to \boldsymbol{\mu}_{MAP} = \boldsymbol{\mu_N}
\end{aligned}
$$

Problem 3 Tossing a biased coin with probability that it comes up heads is μ

1. We toss the coin 3 times and it all comes up with heads. How likely is that in the next toss, the coin comes up with head according to MLE?

   A single coin flip is usually modelled with a Bernoulli distribution.

$$
X_i \sim \mathrm{Ber}(x|\mu) \quad \text{with density} \quad \begin{cases} \mu & x=1 \\ 1-\mu & x=0 \end{cases}
$$

As given in the assignment sheet, three observations have been made as head. We arrive at the MLE by deriving the log likelihood for the probability parameter and setting it zero. So, $\mathcal{X} = (X_1 = 1, X_2 = 1, X_3 = 1)$ and one observation is distributed as such

$$
X_i \sim p(X_i|\mu_i) = \mu_i^{X_i}(1-\mu_i)^{(1-X_i)}
$$

As our observations are supposed to be iid, the log likelihood arises as

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\log p(\mathcal{X}|\mu) &= \frac{\partial}{\partial\mu}\log\prod_{i=1}^{N}\mu^{x_i}(1-\mu)^{(1-x_i)} \\
&= \frac{\partial}{\partial\mu}\sum_{i=1}^{N} x_i\log\mu + (1-x_i)\log(1-\mu) \\
&= \frac{\sum_{i=1}^{N} x_i}{\mu} - \frac{\sum_{i=1}^{N}(1-x_i)}{1-\mu}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\log p(\mathcal{X}|\mu) = 0 \to \sum_{i=1}^{N} x_i(1-\mu) &= \mu\sum_{i=1}^{N}(1-x_i) \\
\sum_{i=1}^{n} x_i - \mu\sum_{i=1}^{n} x_i &= \mu\sum_{i=1}^{n}(1-x_i) \\
\sum_{i=1}^{n} x_i &= \mu\sum_{i=1}^{n} 1 \\
\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{N} x_i &= \frac{1}{3}3 = 1
\end{aligned}
$$

According to the MLE, the 4th toss gets assigned probability 1. When the number of observed heads ($x = 1$) is denoted as m, Bishops 2.8 states, that the MLE can be rewritten as

$$
\mu_{MLE} = \frac{m}{N} = \frac{3}{3} = 1
$$

.

2. Given prior $\mu \sim \mathrm{Beta}(\mu|a,b)$. What is the probability that the coin comes up with head in the 4th toss? The straight forward way would be to take the MAP estimate according to the prior and data likelihood by deriving the log posterior for μ:

$$\frac{\partial}{\partial\mu}\log p(\mathcal{X}|\mu)p(\mu|a,b) = 0$$

As it clearly says we can use Bishops results, I want to direct to 2.15-2.20 and we arrive at the MAP for a binomial likelihood with a beta prior as such

$$p(x=1|D) = \frac{m+a}{m+a+l+b} = \frac{3+a}{3+a+b}$$

Here, a,b are the parameters of the beta distribution and m is the number of heads, l the number of tails and N the number of observations.

3. Suppose that we observe m times that the coin lands heads and l times that it lands tails. Show that the posterior mean $\mathbb{E}(\mu|\mathcal{D})$ lies between the prior mean and μ_{MLE}.
To show that the posterior mean lies between the prior mean and the MLE, we at first give them according to Bishop:

$$\begin{aligned}\mathbb{E}(\mu|a,b) &= \frac{a}{a+b} \quad \text{(beta prior mean: Bishop 2.15)}\\ \mu_{MLE} &= \frac{m}{N} = \frac{m}{m+l} \quad \text{(MLE: Bishop 2.7)}\\ \mathbb{E}(\mu|\mathcal{D}) = p(x=1|\mathcal{D}) &= \frac{m+a}{m+a+l+b} \quad \text{(posterior mean: Bishop 2.19/20)}\end{aligned}$$

Now, we start from the posterior mean to end up with a term, that includes the prior mean and the MLE:

$$\begin{aligned}\mathbb{E}(\mu|\mathcal{D}) &= \frac{m+a}{m+a+l+b}\\ &= \frac{m}{m+a+l+b} + \frac{a}{m+a+l+b}\\ &= \underbrace{\frac{m+l}{m+a+l+b}}_{K_1}\cdot\underbrace{\frac{m}{m+l}}_{\mu_{MLE}} + \underbrace{\frac{a+b}{m+a+l+b}}_{K_2}\cdot\underbrace{\frac{a}{a+b}}_{\mu_{prior}}\end{aligned}$$

As both K_1 and K_2 are smaller or equal than 1 and multiplied to the prior mean and the ML!, we can say that the posterior mean, has to lie in betweem them.

Problem 4. Are the following distributions part of the exponential family, if yes cast them in the exponential form with minimum nuber of parameters and show their sufficient statistic.
A Exponential Family?
Reminder: Exponential Family form:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x})\right) \tag{1}$$

, where $\boldsymbol{\mu}$ are called natural parameters of the distribution and $\boldsymbol{u}(\boldsymbol{x})$ is some function of x. An equivalent form, that explicitly involves the sufficient statistic T(x) is given as

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\left(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{A}(\boldsymbol{\eta})\right) \tag{2}$$

1. Poisson distribution: $p(k|\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}$

   -

$$\begin{aligned}&= \frac{1}{k!}\cdot\exp(\ln(\lambda^k))\exp(-\lambda)\\ &= \underbrace{\frac{1}{k!}}_{h(k)}\exp(\underbrace{\ln(\lambda)}_{\eta}\cdot\underbrace{k}_{T(k)} - \underbrace{\lambda}_{A(\lambda)})\\ &\to \eta = \ln(\lambda) \Rightarrow \lambda = \exp(\eta)\\ &\to A(\eta) = \exp(\eta)\end{aligned}$$

2. Gamma distribution:

$$\begin{aligned} Gam(\tau|a,b)) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau) \\ \text{(exp(log()) on everything} \quad &= \exp\left[-b\tau + a\ln(b) + (a-1)\ln(\tau) + a\ln(b) - \ln(\Gamma(a))\right] \\ \text{(refactoring to exp.fam. form)} \quad &= \tau^{-1}\exp\left[-b\tau + a\ln(b) + a\ln(\tau) + a\ln(b) - \ln(\Gamma(a))\right] \\ &= \tau^{-1}\exp\left[\begin{bmatrix} a \\ -b \end{bmatrix}^\top \begin{bmatrix} \ln(\tau) \\ \tau \end{bmatrix} + a\ln(b) - \ln(\Gamma(a))\right] \end{aligned}$$

$$\begin{aligned} h(\tau) &= \tau^{-1} \\ \boldsymbol{\eta} &= \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} a \\ -b \end{bmatrix} \\ \boldsymbol{T(x)} &= \begin{bmatrix} \ln(\tau) \\ \tau \end{bmatrix} \\ A(\boldsymbol{\eta}) &= -a\ln(b) + \ln(\Gamma(a)) \quad \overset{\text{in terms of } \eta}{=} \quad -\eta_1 \ln(-\eta_2) + \ln(\Gamma(\eta_1)) \end{aligned}$$

3. Cauchy distribution.
   Without looking at the distribution formula, it is known that the Cauchy distribution does not have a moment generating function. Therefore, mean and variance (moment of order 1 and 2) are undefined. However, members of the exponential family must have finite moments for any order. Thus the Cauchy distribution is not part of the exponential family.

4. Von Mises distribution

$$\begin{aligned} \text{VonMises}(x|k,\mu) &= \frac{1}{2\pi I_0(k)} \exp(k \cdot \cos(x-\mu)) \\ &= \frac{1}{2\pi I_0(k)} \exp(k\cos(x)\cos(\mu) + k\sin(x)\sin(\mu))) \\ &= \exp(k\cos(x)\cos(\mu) + k\sin(x)\sin(\mu)) - \ln(2\pi I_0(k))) \end{aligned}$$

In the second row, we draw on the difference rule property of trigonomic functions : $\cos(x-y) = \cos x \cos y + \sin x \sin y$. Thus, the exponential family form is achieved as

$$\begin{aligned} h(x) &= 1 \\ \boldsymbol{\eta} &= \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} ksin(\mu) \\ k\cos(\mu) \end{bmatrix} \\ &\to \sqrt{\cos^2(\mu) + \sin^2(\mu)} = 1 \\ &\to \sqrt{\left(\frac{\eta_1}{k}\right)^2 + \left(\frac{\eta_2}{k}\right)^2} = 1 \\ &\to \sqrt{\eta_1^2 + \eta_2^2} = k \\ T(x) &= \begin{bmatrix} sin(x) \\ \cos(x) \end{bmatrix} \\ A(\boldsymbol{\eta}) &= \ln(2\pi I_0(k)) = \ln\left(2\pi I_0\left(\sqrt{\eta_1^2 + \eta_2^2}\right)\right) \end{aligned}$$

For distributions 1,2 derive first order and second order moment
The moments of distributions that belong to the exponential family can be obtained by deriving $A(\eta)$ for η. Whereas, the first order derivative yields the mean and the second order derivative yields the variance. Following this rule, we get:

1. Poisson distribution:

$$\begin{aligned} \mathbb{E}(p(k|\eta) &= \frac{\partial A}{\partial \eta} = \frac{\partial \exp(\eta)}{\partial \eta} = \exp(\eta) = \lambda \\ \text{Var}(p(k|\eta) &= \frac{\partial^2 A}{\partial^2 \eta} = \frac{\partial^2 \exp(\eta)}{\partial^2 \eta} = \exp(\eta) = \lambda \end{aligned}$$

2. Gamma distribution: Reminder:

$$\boldsymbol{\eta} = \begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}a\\ -b\end{bmatrix}$$

$$A(\boldsymbol{\eta}) = -a\ln(b) + \ln(\Gamma(a)) \overset{\text{in terms of } \eta}{=} -\eta_1 \ln(-\eta_2) + \ln(\Gamma(\eta_1))$$

$$\mathbb{E}(p(\tau|a,b) = \frac{\partial A}{\partial \eta_2} = \frac{\partial - \eta_1 \ln(-\eta_2) + \ln(\Gamma(\eta_1))}{\partial \eta_2} = \frac{-\eta_1}{\eta_2} = \frac{a}{b}$$
$$\mathrm{Var}(p(\tau|a,b) = \frac{\partial^2 A}{\partial^2 \eta_2} = \frac{\partial^2 - \eta_1 \ln(-\eta_2) + \ln(\Gamma(\eta_1))}{\partial^2 \eta_2} = \frac{\eta_1}{\eta_2^2} = \frac{a}{b^2}$$

Does the Poisson distribuion have a conjugate prior? If yes, derive it
Every distribution, that is part of the exponential family does have a conjugate prior. Therefore, the Poisson distribution has a conjugate prior which can be derived in several ways. In general, the conjugate prior means that that the posterior is of the same distribution family as the prior. One way is to calculate the posterior assuming that a certain prior is in fact the conjugate and then show that the posterior is again of the that distribution family. Another, more elegant way is to draw on the properties of exponential families. According to the alternative formulation shown in equation 2, the conjugate prior arises as the general form:

$$\begin{aligned} p_\pi(\eta|\tau,\nu) &= p(\tau,\nu)\exp\left(\eta^{\mathrm{T}}\tau - \nu A(\eta)\right) \\ &\propto \exp(\eta^\top \tau - \nu A(\eta)) \end{aligned}$$

If we now take into account the results from above and take $\tau = a, \nu = b$ as the parameters from the Gamma distribution and parameters of the Poisson distribution $A(\eta) = \exp(\eta)$ and therefore $\lambda = \exp(\eta)$.

$$\begin{aligned} \text{(apply A)} \qquad &= \exp\left[\eta^T \tau - \nu \exp(\eta)\right] \\ \text{(apply a,b)} \qquad &= \exp[\ln(\lambda)a - b\lambda] \\ \text{(gamma form)} \qquad &= \lambda^a \exp[-b\lambda] \end{aligned}$$

The last row corresponds to the Gamma distribution with a missing normalizing factor, as I used the propto sign. However, this corresponds to a $\mathrm{Gam}(\lambda|(a+1), b)$ Gamma distribution. Thus, the posterior for a Gamma prior and a Poisson likelihood function yields another Gamma distribution, which makes it a conjugate to the Gamma prior.