Machine Learning 2
Homework Assignment 1

Volodymyr Medentsiy
id: 12179078
volodymyr.medentsiy@student.uva.nl

September 2019

1 Problem 1

Consider two random vectors $x \in R^n$ and $z \in R^n$ having Gaussian distribution $p(x) = N(x|\mu_x, \sum_x)$ and $p(z) = N(z|\mu_z, \sum_z)$. Consider random vector $y = x + z$. Derive mean and covariance of $p(y)$. What is the covariance of y if you assume that x and z are independent?

Solution
1)

$$\mu_y = E(x + z) = E(x) + E(z) = \mu_x + \mu_z$$

2)

$$\Sigma_y = E((y - Ey)(y - Ey)^T) = E((x + z - \mu_x - \mu_z)(x + z - \mu_x - \mu_z)^T) =$$
$$= E(xx^T + xz^T - x\mu_x^T - x\mu_z^T + zz^T + zx^T - z\mu_x^T - z\mu_z^T - \mu_x x^T - \mu_x z^T + \mu_x\mu_x^T + \mu_x\mu_z^T - \mu_z x^T - \mu_z z^T + \mu_z\mu_x^T + \mu_z\mu_z^T) =$$
$$= E(xx^T - x\mu_x^T - \mu_x x^T + \mu_x\mu_x^T) + E(zz^T - z\mu_z^T - \mu_z z^T + \mu_z\mu_z^T) + E(xz^T + zx^T - x\mu_z^T - z\mu_x^T - \mu_x z^T + \mu_x\mu_z^T - \mu_z x^T + \mu_z\mu_x^T) =$$
$$= \Sigma_x + \Sigma_z + E(xz^T) + E(zx^T) - \mu_x\mu_z^T - \mu_z\mu_x^T - \mu_x\mu_z^T + \mu_x\mu_z^T - \mu_z\mu_x^T + \mu_z\mu_x^T = \Sigma_x + \Sigma_z + E(xz^T) + E(zx^T) - 2\mu_x\mu_z^T$$

3) Assuming that x and z are independent, $E(xz^T) = ExEz^T = \mu_x\mu_z^T = E(zx^T)$:

$$\Sigma_y = \Sigma_x + \Sigma_z + E(xz^T) + E(zx^T) - 2\mu_x\mu_z^T = \Sigma_x + \Sigma_z + \mu_x\mu_z^T + \mu_x\mu_z^T - 2 * \mu_x\mu_z^T = \Sigma_x + \Sigma_z$$

2 Problem 2

Consider a D-dimensional Gaussian random variable x with distribution $N(x|\mu, \Sigma)$ in which the covariance Σ is known. Given a set of N i.i.d. observations X = x_1 , . . . , x_N . Assume that $x_i \sim N(\mu, \Sigma)$ and $\mu \sim N(\mu_0, \Sigma_0)$. [Hint: you may directly use results from Bishop]

1. Write down the likelihood of the data $p(X|\mu,\Sigma)$;
2. Write down the form of the posterior $p(\mu|X,\Sigma,\mu_0,\Sigma_0)$ (you do not need to normalize the probability distribution by calculating the evidence).
3. Show that $p(\mu|X,\Sigma,\mu_0,\Sigma_0)$ is a Gaussian distribution $N(|N,N)$ and find the values of μ_N and Σ_N (hint: use "completing the square")
4. Derive the maximum a posteriori solution for μ;

Solution

1

$$p(X|\mu,\Sigma) = [X_1,..X_N - i.i.d.] = \prod_{i=1}^{N} p(x_i|\mu,\Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{D/2}(det\Sigma)^{\frac{1}{2}}} e^{\frac{-1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)} =$$
$$= \frac{1}{(2\pi)^{ND/2}(det\Sigma)^{\frac{N}{2}}} e^{\frac{-1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$$

2

$$p(\mu|X,\Sigma,\mu_0,\Sigma_0) = \frac{p(X|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0)}{\int p(X|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0)d\mu} \propto p(X|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0) = \prod_{i=1}^{N} p(x_i|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0) \propto$$
$$\propto e^{-\frac{1}{2}((\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)+\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu))}$$

3

First let us rewrite the power of exponent (I will use equation 2.71 from Bishop):

$$-\frac{1}{2}((\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)+\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)) = -\frac{1}{2}(\mu^T\Sigma_0^{-1}\mu-\mu_o^T\Sigma_0^{-1}\mu-\mu^T\Sigma_0^{-1}\mu_0+\mu_0^T\Sigma_0^{-1}\mu_0+$$
$$+\sum_{n=1}^{N} x_n^T\Sigma_{-1}x_n - \sum_{n=1}^{N}\mu^T\Sigma^{-1}x_n - \sum_{n=1}^{N}x_n^T\Sigma^{-1}\mu + \sum_{n=1}^{N}\mu^T\Sigma^{-1}\mu) =$$
$$= -\frac{1}{2}\mu^T(\Sigma_0^{-1}+N\Sigma^{-1})\mu + \mu^T(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N}x_n) + const =$$
$$\left[\text{ with } const = -\frac{1}{2}(-\mu_0^T\Sigma_0^{-1}\mu+\mu_0^T\Sigma_0^{-1}\mu_0+\sum_{n=1}^{N}x_n^T\Sigma^{-1}x_n-\sum_{n=1}^{N}x_n^T\Sigma^{-1}\mu-\mu^T(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N}x_n))\right]$$

$$= -\frac{1}{2}\mu^T(\Sigma_0^{-1}+N\Sigma^{-1})\mu + \mu^T(\Sigma_0^{-1}+N\Sigma^{-1})(\Sigma_0^{-1}+N\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N}x_n) + const =$$
$$= [\text{Using equation 2.71}] = -\frac{1}{2}(\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N),$$

With $\mu_N = (\Sigma_0^{-1}+N\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N} x_n) = \Sigma_N(\Sigma_0^{-1}\mu_0+\Sigma^{-1}\sum_{n=1}^{N} x_n), \Sigma_N^{-1} = \Sigma_0^{-1}+N\Sigma^{-1}$. We showed that posterior $p(\mu|X,\Sigma,\mu_0,\Sigma_0) \propto e^{-\frac{1}{2}(\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N)}$, thus we can conclude that posterior has Normal distribution with parameters μ_N and Σ_N
4

$$\mu_{MAP} = argmax_{\mu} p(\mu|X,\Sigma,\mu_0,\Sigma_0) = argmax_{\mu} lnp(\mu|X,\Sigma,\mu_0,\Sigma_0) = argmax_{\mu} lnp(X|\mu,\Sigma)p(\mu|\mu_0,\Sigma_0) =$$
$$argmax_{\mu} lnp(X|\mu,\Sigma) + lnp(\mu|\mu_0,\Sigma_0) = argmax_{\mu}(-\frac{ND}{2}ln(2\pi) - \frac{N}{2}ln(det(\Sigma)) -$$
$$-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu) - \frac{D}{2}ln2\pi - \frac{1}{2}ln(det(\Sigma_0)) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)) =$$
$$= argmax_{\mu}(-\frac{1}{2}\sum_{n=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)) \Rightarrow$$
$$\Rightarrow \frac{d}{d\mu}\bigg(-\frac{1}{2}\sum_{n=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) - \frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)\bigg) = 0$$
$$\Rightarrow -\Sigma^{-1}\sum_{n=1}^{N}(x_n-\mu) + \Sigma_0^{-1}(\mu-\mu_0) = 0 \Rightarrow (\Sigma^{-1}N + \Sigma_0^{-1})\mu - (\Sigma^{-1}\sum_{n=1}^{N} x_n + \Sigma_0^{-1}\mu_0) = 0$$
$$\Rightarrow \mu_{MAP} = \frac{\Sigma^{-1}N + \Sigma_0^{-1}}{\Sigma^{-1}\sum_{n=1}^{N} x_n + \Sigma_0^{-1}\mu_0}$$

3 Problem 3

Tossing a biased coin with probability that it comes up heads is μ [Hint: you may use results from Bishop]

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?
2. Suppose that the prior $\mu \sim Beta(\mu|a,b)$. What is the probability that the coin comes up with head in the 4th toss?
3. Suppose that we observe m times that the coin lands heads and l times that it lands tails. Show that the posterior mean $E[\mu|D]$ (see Bishop 2.19) lies between the prior mean and μ_{MLE} .

Solution
1) x $\sim Bernoulli(\mu)$, so $\mu_{MLE} = \frac{number\ of\ successes}{number\ of\ trials} = \frac{3}{3} = 1$
2) Beta prior is a conjugate prior for Bernoulli likelihood, thus the posterior distribution of μ, which corresponds to the probability that coin will come up heads in the 4-toss, could be easily computed (Bishop's equation 2.20):

$$p(\mu|Data = \{heads\ =\ 3,\ tails\ =\ 0\}) = Beta(\mu|a+3,b)$$

3) $Data = \{heads = m, \ tails = l\}$.
Let us show that $E(\mu|Data) = \lambda Ep(\mu) + (1-\lambda)\mu_{MLE}$ with $\lambda \in [0;1]$, which proves that $E(\mu|Data)$ lies between $Ep(\mu)$ and μ_{MLE}. Using that: $E(\mu|Data) = \frac{a+m}{a+m+b+l}$, $\mu_{MLE} = \frac{m}{m+l}$ and $Ep(\mu) = \frac{a}{a+b}$:

$$\frac{a+m}{a+m+b+l} = \lambda\frac{a}{a+b} + (1-\lambda)\frac{m}{m+l} \Rightarrow \lambda(\frac{a}{a+b} - \frac{m}{m+l}) + \frac{m}{m+l} = \frac{a+m}{a+m+b+l} \Rightarrow$$

$$\lambda = (\frac{a+m}{a+m+b+l} - \frac{m}{m+l})\frac{(a+b)(m+l)}{a(m+l)-m(a+b)} = \frac{((a+m)(b+l)-m(a+m+b+l))(a+b)(m+l)}{(a+m+b+l)(m+l)(a(m+l)-m(a+b))} =$$

$$= \frac{(am+m^2+al+ml-am-m^2-mb-ml)(a+b)}{(a+m+b+l)(al-mb)} = \frac{a+b}{a+b+m+l} \in [0;1]$$

4 Problem 4

Consider the following distributions:

- $Poiss(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $Gam(\tau|a,b) = \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}$
- $Cauchy(x|\gamma,\mu) = \frac{1}{\pi\gamma(1+(\frac{x-\mu}{\gamma})^2)}$
- $vonMises(x|k,\mu) = \frac{1}{2\pi I_0(k)}e^{kcos(x-\mu)}$

Answer the following questions:

1. Are the above distributions members of an exponential family. If yes, then (a) cast them in exponential form (Bishop eq. 2.194) with a minimum numbers of parameters: $p(x|\eta) = h(x)g(\eta)exp(\eta^T u(x))$ (b) derive their sufficient statistics.

2. Derive the first moment about zero (i.e. the mean) and the second moment about the mean (i.e. the variance) of the distributions $Poiss(k|\lambda)$ and $Gam(\theta|a,b)$.

3. Does the Poisson distribution have a conjugate prior? Derive the conjugate prior, if the answer is "yes".

Solution
1.1.
$Poiss(k|\lambda) = \frac{1}{k!}e^{ln\lambda^k}e^{-\lambda} = \frac{1}{k!}e^{-\lambda}e^{k*ln\lambda}$. So $h(k) = \frac{1}{k!}$, $u(k) = k$, $\eta(\lambda) = ln(\lambda)$ and $g(\eta) = e^{-\lambda}$, which shows that $Poiss(k|\lambda)$ is part of an exponential family. Thus $u(k) = k$ is its sufficient statistic.
1.2.
$Gam(\tau|a,b) = \frac{b^a}{\Gamma(a)}\frac{1}{\tau}e^{-b\tau+(a-1)ln\tau}$. So $u(\tau) = (ln(\tau);\tau)^T$, $h(\tau) = 1$, $\eta(a,b) = (a-1;-b)^T$ and $g(\eta) = \frac{b^a}{\Gamma(a)}$, which shows that $Gam(\tau|a,b)$ is part of an exponential family. Thus $u(\tau) = (ln(\tau);\tau)^T$ are its sufficient statistics.

1.3.
Cauchy distribution does not have finite moments of any order, which proves that it is not part of the exponential family.
1.4.
$vonMises(x|k,\mu) = \frac{1}{2_0(k)} e^{k(cos(x)cos(\mu)-sin(x)sin(\mu))}$. So $u(\tau) = (cos(x); sin(x))^T$, $h(\tau) = 1$, $\eta(k,\mu) = (kcos(\mu); -ksin(\mu))^T$ and $g(\eta) = \frac{1}{2_0(k)}$, which shows that $vonMises(x|k,\mu)$ is part of an exponential family. Thus $u(\tau) = (cos(x); sin(x))^T$ are its sufficient statistics.
2.1
Let us use probability generating function properties to derive mean and variance of the Poisson distribution:

$$PGF_{poiss}(z) = \sum_{k=0}^{\infty} \frac{a^k}{k!} e^{-a} z^k = e^{-a} e^{az} = e^{-a(1-z)}$$

$$E(poiss(k|\lambda)) = \frac{dPGF_{poiss}(z)}{dz}\bigg|_{z=1} = ae^{-a(1-z)}\bigg|_{z=1} = a$$

$$E(poiss(k|\lambda)^2) = (\frac{d^2 PGF_{poiss}(z)}{dz^2} + \frac{dPGF_{poiss}(z)}{dz})\bigg|_{z=1}$$

$$Var(poiss(k|\lambda)) = (\frac{d^2 PGF_{poiss}(z)}{dz^2} + \frac{dPGF_{poiss}(z)}{dz} + (\frac{dPGF_{poiss}(z)}{dz})^2)\bigg|_{z=1} = a^2 + a - a^2 = a.$$

2.2

$$E(Gam(\tau|a,b)) = \int_0^{\infty} \tau \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau} = \frac{b^a}{\Gamma(a)} \int_0^{\infty} \tau^a e^{-b\tau} d\tau = \frac{b^a}{\Gamma(a) b^{a+1}} \int_0^{\infty} (b\tau)^a e^{-b\tau} d(b\tau) =$$
$$= \frac{b^a}{\Gamma(a) b^{a+1}} \Gamma(a+1) = \frac{b^a}{\Gamma(a) b^{a+1}} a\Gamma(a) = \frac{a}{b}$$

$$E(Gam(\tau|a,b)^2) = \int_0^{\infty} \tau^2 \frac{b^a}{\Gamma(a)} \tau^{a-1} e^{-b\tau} = \frac{b^a}{\Gamma(a)} \int_0^{\infty} \tau^{a+1} e^{-b\tau} d\tau = \frac{b^a}{\Gamma(a) b^{a+2}} \int_0^{\infty} (b\tau)^{a+1} e^{-b\tau} d(b\tau) =$$
$$= \frac{b^a}{\Gamma(a) b^{a+1}} \Gamma(a+2) = \frac{b^a}{\Gamma(a) b^{a+1}} a(a+1)\Gamma(a) = \frac{a}{b} = \frac{a(a+1)}{b^2}$$

$$Var(Gam(\tau|a,b)) = E(Gam(\tau|a,b)^2) - E(Gam(\tau|a,b))^2 = \frac{a^2}{b^2} + \frac{a}{b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}$$

3
Let us show that Gamma prior is conjugate to Poisson likelihood, so we will show that posterior $p(\tau|k)$ has Gamma distribution:

$$p(\tau|k) \quad = \quad \frac{Poiss(k|\tau) * Gam(\tau|a,b)}{\int_0^\infty Poiss(k|\tau) * Gam(\tau|a,b)} \quad = \quad \frac{\frac{1}{k!}e^{-\tau}\tau^k * \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}}{\int_0^\infty \frac{1}{k!}e^{-\tau}\tau^k * \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}d\tau}$$

Let us first find the denominator:

$$\int_0^\infty \frac{1}{k!}e^{-\tau}\tau^k * \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}d\tau = \frac{b^a}{k!\Gamma(a)}\int_0^\infty e^{-\tau(b+1)}\tau^{a+k-1}d\tau =$$
$$= \frac{b^a}{k!\Gamma(a)(b+1)^{a+k}}\int_0^\infty e^{-\tau(b+1)}((b+1)\tau)^{a+k-1}d((b+1)\tau) = \frac{b^a\Gamma(a+k)}{k!\Gamma(a)(b+1)^{a+k}}$$

Thus

$$p(\tau|k) = \frac{\frac{1}{k!}e^{-\tau}\tau^k * \frac{1}{\Gamma(a)}b^a\tau^{a-1}e^{-b\tau}}{\frac{b^a\Gamma(a+k)}{k!\Gamma(a)(b+1)^{a+k}}} = \frac{\tau^{k+a-1}e^{\tau(b-1)}(b+1^{a+k})}{\Gamma(a+k)} = Gam(\tau|a+k,b+1)$$

So posterior is in the same distribution family as prior, which proves that Gamma prior is conjugate to Poisson likelihood.