Machine Learning (Homework #2)

A014587 羅羿牧

1. Information Theory

(a) Please show that the maximum entropy distribution for a continuous variable with three constrains

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$$

is a Gaussian distribution.

Ans.

$$L = -\int_{-\infty}^{\infty} p(x)\ln p(x)\,dx + \lambda_1\left[\int_{-\infty}^{\infty} p(x)dx - 1\right] + \lambda_2\left[\int_{-\infty}^{\infty} xp(x)dx - \mu\right] + \lambda_3\left[\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx - \sigma^2\right]$$

$$\frac{\partial L}{\partial p(x)} = -\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

$$\Rightarrow p(x) = \exp\left\{-1+\lambda_1+\lambda_2 x+\lambda_3 (x-\mu)^2\right\}$$

代入
$$\begin{cases} \int_{-\infty}^{\infty} p(x)dx = 1 \\ \int_{-\infty}^{\infty} xp(x)dx = \mu \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2 \end{cases} \Rightarrow \begin{cases} \lambda_1 = 1-\frac{1}{2}\ln(2\pi\sigma^2) \\ \lambda_2 = 0 \\ \lambda_3 = \frac{1}{2\sigma^2} \end{cases}$$

$$\therefore\ p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

(b) Gaussian distribution is given by

$$p(\mathrm{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Please derive the corresponding entropy.

Ans.

(b)

$$H[x] = -\int p(x) \ln p(x)\, dx = -\int p(x)\left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right]dx$$

$$= \frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\int p(x)(x-\mu)^2 dx\right]$$

$$= \frac{1}{2}\left[\ln(2\pi\sigma^2) + 1\right]$$

2. Bayesian Inference for the Gaussian

We develop a Bayesian learning by introducing prior distributions to estimate Gaussian parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Traditionally, batch learning is performed by using the whole training set where high computational complexity is caused. If training data is sufficiently large, it is suitable to use sequential learning (on-line learning) algorithm. Please solve the following question. The file **r2.mat** contains a 1000-point sequence, which is generated by the following multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = [1, \; -1]^{\mathrm{T}}$ and $\boldsymbol{\Sigma}$ **(** $\boldsymbol{\Sigma}$ is unknown**)**. The sequential learning of the posterior distribution of $\boldsymbol{\Lambda}$ ($\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$) with the contribution from the final data $\mathbf{x}_N$ can be expressed as follows:

$$p(\boldsymbol{\Lambda}|\mathbf{X}) \propto \left[p(\boldsymbol{\Lambda}) \prod_{\mathrm{n}=1}^{N-1} \mathrm{p}(\mathbf{x}_{\mathrm{n}}|\boldsymbol{\Lambda})\right] \mathrm{p}(\mathbf{x}_N|\boldsymbol{\Lambda})$$

(a) Please derive the posterior distribution of precision matrix $\boldsymbol{\Lambda}$, $p(\boldsymbol{\Lambda}|\mathbf{X}) = \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_{\boldsymbol{\Lambda}}, \mathcal{V}_{\boldsymbol{\Lambda}})$, in details where $\mathcal{V}_{\boldsymbol{\Lambda}}$ is called the *degrees of freedom* of the distribution and $\mathbf{W}_{\boldsymbol{\Lambda}}$ is a $D \times D$ symmetric matrix. Here, we apply the conjugate prior of $\boldsymbol{\Lambda}$ which is a *Wishart* distribution $p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W_0}, \mathcal{V}_0)$.

Ans.

(a)

$$p(\mu | m_1, m_2, a, b) \propto \mu^{m_1+a-1}(1-\mu)^{m_2+b-1}$$

$$= \frac{\Gamma(m_1+m_2+a+b)}{\Gamma(m_1+a)\Gamma(m_2+b)} \mu^{m_1+a-1}(1-\mu)^{m_2+b-1}$$

$$\mu_{MAP} = \frac{\partial \ln p(\mu | m_1, m_2, a, b)}{\partial \mu} = \frac{\partial}{\partial \mu} \ln\left[K \cdot \mu^{m_1+a-1}(1-\mu)^{m_2+b-1}\right]$$

$$\Rightarrow \frac{\partial}{\partial \mu}\left[(m_1+a-1)\ln \mu + (m_2+b-1)\ln(1-\mu)\right] = 0$$

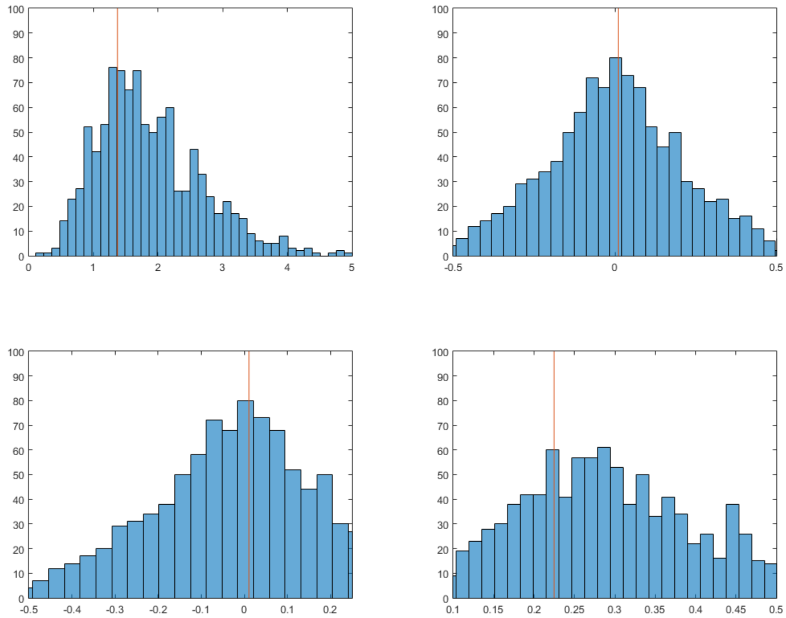
$$\Rightarrow \frac{m_1+a-1}{\mu} = \frac{m_2+b-1}{1-\mu} \Rightarrow (m_1+a-1)(1-\mu) = \mu(m_2+b-1)$$

$$\Rightarrow \mu = \frac{m_1+a-1}{m_1+m_2+a+b-2}$$

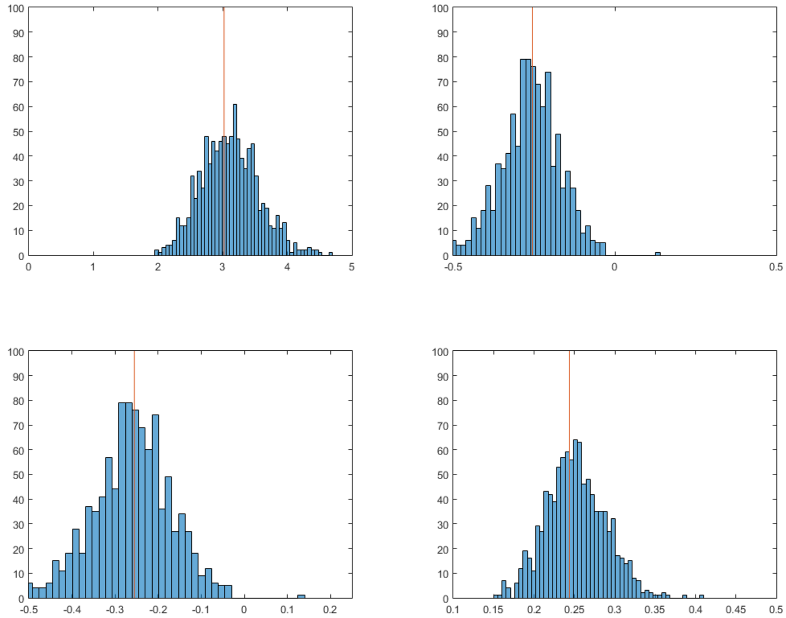
(b) Please consider the *Wishart* prior $p_1(\mathbf{\Lambda}) = \mathcal{W}(\mathbf{\Lambda}|\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, 1)$ and find the MAP solution of $\mathbf{\Lambda}$ (or $\mathbf{\Sigma}$) for N =10, 100, and 500. ($\mathbf{\Lambda_{MAP}} = \mathrm{argmax}_{\mathbf{\Lambda}}\, p(\mathbf{\Lambda}|\mathbf{X})$) You may also directly use the Matlab command 'wishrnd' to generate many samples of $\mathbf{\Lambda}$ and compare their corresponding $p(\mathbf{\Lambda})$ to obtain the approximate MAP solution.

Ans.

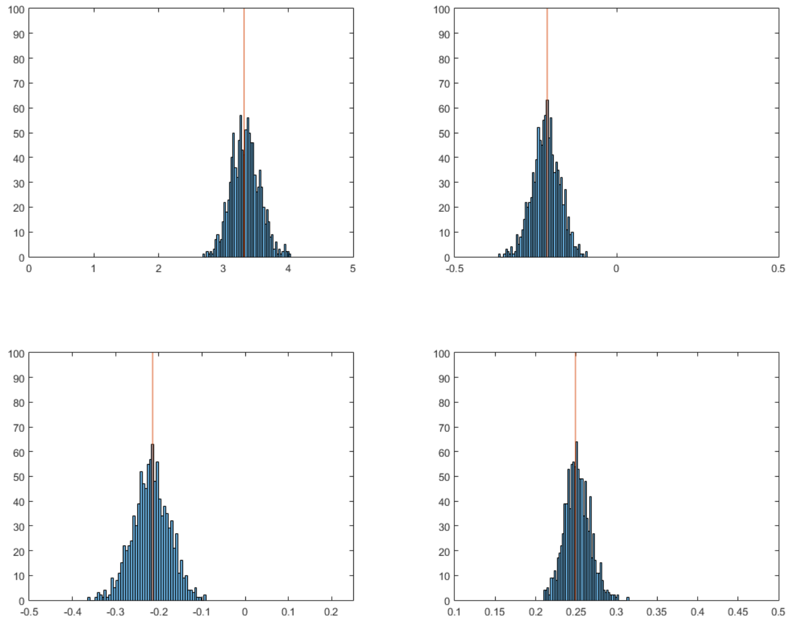
- $N = 10$



- $N = 100$



- $N = 500$



3. Bayesian Inference for the Binomial

A discrete variable is given with two possible states. Suppose we draw this variable N times, the outcomes of the N trials are recorded as **O.mat**. Let $D = (m_1, m_2)$ denote the numbers of occurrences of two states from the draws. These draws can be represented by a binomial distribution $\mathrm{Bin}(m|N, \mu)$ where μ denotes the probability or parameter of the first state which satisfies $\mu \geq 0$. Please solve the following problems.

(a) Please apply the conjugate prior of $\boldsymbol{\mu}$, which is a Beta distribution, $\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, derive the posterior distribution $p(\mu|D,a,b)$, and show the derivation of MAP solution $\boldsymbol{\mu}_{\text{MAP}}$ in details.

Ans.

$$p(\Lambda|X) \propto \left[p(\Lambda) \prod_{n=1}^{N-1} p(x_n|\Lambda) \right] p(x_N|\Lambda) \propto p(\Lambda) \prod_{n=1}^{N} p(x_n|\Lambda)$$

$$= \mathcal{W}(\Lambda|W_0, \nu_0) \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \Lambda^{-1})$$

$$\propto |\Lambda|^{(\nu_0-D-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}(W_0^{-1}\Lambda)\right\} \cdot |\Lambda|^{N/2} \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Lambda(x_n-\mu)\right\}$$

$$= |\Lambda|^{[(\nu_0+N)-D-1]/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}(W_0^{-1}\Lambda)\right\} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\Lambda\sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T\right)\right\}$$

$$= |\Lambda|^{(\nu_\Lambda-D-1)/2} \exp\left\{-\frac{1}{2}\mathrm{Tr}(W_\Lambda^{-1}\Lambda)\right\} = \mathcal{W}(\Lambda|W_\Lambda, \nu_\Lambda)$$

where $\nu_\Lambda = \nu_0 + N$, $W_\Lambda^{-1} = W_0^{-1} + \sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T$

(b) **Programming**:

You can use Beta random variable for parameter ¹ . Please use the recorded data **O.mat** and plot the prior and posterior distributions from 50 data samples and from the whole data samples. The parameters of the prior distribution are given as $a = b = 0.1$.

Ans.