Machine Learning (Homework #2)

Due date: 11/27

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.

(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.

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 $p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}_{\mathbf{0}}, \mathcal{V}_0)$.

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

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, $\mathrm{Beta}(\mu|a,b)=\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, derive the posterior distribution $p(\mu|D,a,b)$, and show the derivation of MAP solution $\boldsymbol{\mu}_{\mathrm{MAP}}$ in details.

(b) **Programming**:

You can use Beta random variable for parameter μ. 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 $a=b=0.1$.