Machine Learning (Homework 2)

Due date: 11/24

1 Bayesian Inference for Gaussian (30%)

Bayesian learning is performed by introducing prior distribution to estimate Gaussian parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from training samples $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Traditionally, batch learning is run by using the whole training set where high computational complexity is caused. If the training data are sufficiently large, it is suitable to develop sequential learning algorithm (also called on-line learning). Please solve the following questions.

Case: $\boldsymbol{\mu}$ is known but $\boldsymbol{\Sigma}$ is unknown
The file 1_data.mat contains a 1000-point sequence, which is generated by the following multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} = [1, -1]^\top$ and covariance matrix $\boldsymbol{\Sigma}$. The sequential learning of the posterior distribution of precision matrix $\boldsymbol{\Lambda}(\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1})$ with the contribution from the additional sample $\mathbf{x}_N$ can be expressed as follow:

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

1. Please derive the posterior distribution of precision matrix $\boldsymbol{\Lambda}$, $p(\boldsymbol{\Lambda}|\mathbf{X}) = \mathcal{W}(\boldsymbol{\Lambda}|W_\Lambda, \nu_\Lambda)$, in details where ν_Λ is called the degrees of freedom of the distribution and W_Λ is a $D \times D$ symmetric matrix. Here, the conjugate prior of $\boldsymbol{\Lambda}$ based on a Wishart distribution $p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda}|W_0, \nu_0)$ is applied.

2. Using the Wishart prior: $p(\Lambda) = \mathcal{W}\left(\boldsymbol{\Lambda}\middle|\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, 1\right)$. Please find the MAP solution to $\boldsymbol{\Lambda}$ (or $\boldsymbol{\Sigma}$) for $N = 10$, 100 and 500. ($\boldsymbol{\Lambda}_{\text{MAP}} = \arg\max_\Lambda p(\boldsymbol{\Lambda}|\mathbf{X})$).

You may 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.

2 Bayesian Linear Regression (30%)

In this exercise, you will implement an example of Bayesian linear regression and discuss the issues on parameter distribution and predictive distribution.

Dataset:
The file 2_data.mat contains two sequences $\mathbf{x} = \{x_1, x_2, \ldots, x_{100}|0 \leq x_i \leq 2\}$ and $\mathbf{t} = \{t_1, t_2, \ldots, t_{100}\}$ which represent the input sequence and the corresponding target sequence, respectively.

Basis Function:
Please apply the sigmoidal basis functions $\boldsymbol{\phi} = [\phi_0, \ldots, \phi_{M-1}]^\top$ of the form $\phi_j(x) = \sigma(\frac{x-\mu_j}{s})$ where $\sigma(a)$ is the logistic sigmoid function defined in (3.6). In this exercise, please take the

following parameter settings for your basis functions: $M = 7$, $s = 0.1$ and $\mu_j = \frac{2j}{M}$ with $j = 0, 1, \ldots, (M-1)$. In order to discuss how the amount of training data affects the regression process, please take the data size to be N=10, 15, 30 and 80, for each of the following questions:

1. Please compute the mean vector $\mathbf{m}_N$ and the covariance matrix S_N for the posterior distribution $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ with the given prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0 = \mathbf{0}, \mathbf{S}_0^{-1} = 10^{-6}\mathbf{I})$. The precision of likelihood function $p(\mathbf{t}|\mathbf{w}, \beta)$ or $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ is chosen to be β =1.
2. Similar to Fig. 3.9, please generate five curve samples from the parameter posterior distribution.
3. Similar to Fig. 3.8, please plot the predictive distribution of target value t and show the mean curve and the region of variance with one standard deviation on either side of the mean curve.

3 Logistic Regression (40%)

You are given the Wine data set (train.csv and test.csv). This data set contains 3 classes. The first 3 dimensions in train.csv are the values of 1-of-K coding for a target vector, the other dimensions are the values of data. In this exercise, you will implement the Newton-Raphson algorithm to construct a multiclass logistic regression model with the softmax transformation as $p(C_k|\phi) = y_k(\phi) = \exp(a_k)/\sum_j \exp(a_j)$. The error function is formed by using the cross-entropy function as $E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$. Note: You have to set a stopping criterion $E(\mathbf{w}) < \epsilon$.

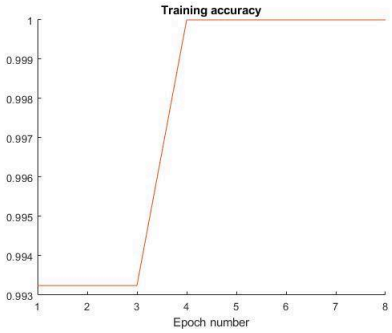
1. Set the initial $\mathbf{w}$ to be zero, and show the learning curve of $E(\mathbf{w})$ and the accuracy of classification versus number of epochs until convergence of training data.
2. Show the classification result of test data.
3. Please plot the distribution (or histogram) of each variable in training data and map different colors to each class.
4. Explain that how do you know the model you trained is on the way to global minimum.

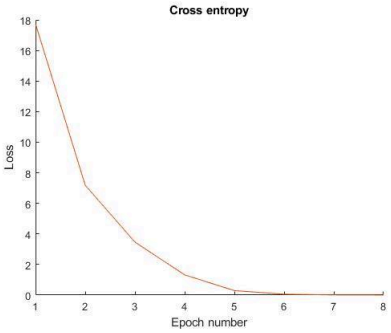
Bonus questions (15%)

5. Please choose a pair of the most contributive variables and plot the samples in training data via 2D graph.
6. Use the variables you choose in (5) redo (1) and (2).

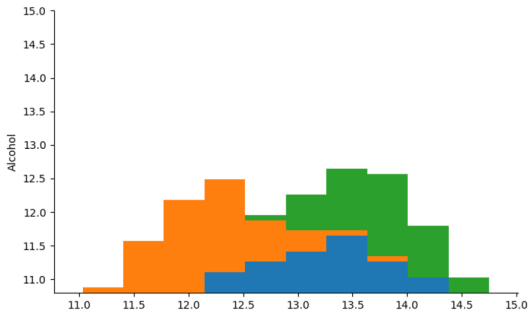
Hint:

1. To avoid the overflow, you are suggested to use the alternative of softmax function $p(C_k|\phi) = y_k(\phi) = \frac{1}{\sum_j \exp(a_j - a_k)}$.
2. Parameter of stopping criterion ϵ is suggested to be between 0.001 and 0.01.
3. An example of (1) is





4. An example of (3) is



5. An example of (5) is