Homework 2

Handed out: Sunday, February 3, 2019

Due: Friday, February 15, 2019 11:55 pm

Notes:

- We *highly* encourage typed (Latex or Word) homework. Compile as single report containing solutions, derivations, figures, etc.
- Submit all files including report pdf, report source files (e.g. .tex or .docx files), data, figures produced by computer codes and programs files (e.g. .py or .m files) in a **.zip** folder. Programs should include a Readme file with instructions on how to run your computer programs.
- Zipped folder should be turned in on Sakai with the following naming scheme: **HW2_LastName_FirstName.zip**
- Collaboration is encouraged however all submitted reports, programs, figures, etc. should be an individual student's writeup. Direct copying could be considered cheating.
- Homework problems that simply provide computer outputs with no technical discussion, Algorithms, etc. will receive no credit.
- Software resources (if any) for this Homework set can be downloaded from this link.

1 Linear Regression

Consider a data set in which each data point y_n is associated with a weighting factor $r_n > 0$. Therefore, the sum of square error function becomes

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left\{y_n - \boldsymbol{w}^T \phi(\mathbf{x}_n)\right\}^2. \tag{1}$$

Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (a) data dependent noise variance (b) replicated data points.

2 Evidence

A. In linear regression, the marginal likelihood function is given as

$$p(\boldsymbol{t}|\alpha,\beta) = \int p(\boldsymbol{t}|\boldsymbol{w},\beta)\, p(\boldsymbol{w}|\alpha) d\boldsymbol{w}. \tag{2}$$

where

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}\left(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right) \tag{3a}$$

$$\begin{aligned}\log p\left(\boldsymbol{t}|\boldsymbol{w},\beta\right) &= \sum_{n=1}^{N}\log\mathcal{N}\left(t_n|\boldsymbol{w}^T\phi\left(\mathbf{x}_n\right),\beta^{-1}\right)\\ &= \frac{N}{2}\log\beta-\frac{N}{2}\log\left(2\pi\right)-\beta E_D\left(\boldsymbol{w}\right)\end{aligned} \tag{3b}$$

$$E_D\left(\boldsymbol{w}\right)=\frac{1}{2}\sum_{n=}^{N}\left\{t_n-\boldsymbol{w}^T\phi\left(\mathbf{x}_n\right)\right\}^2 \tag{3c}$$

Using the above equations, derive an expression for the marginal likelihood.

B. The conjugate prior for a Gaussian distribution with unknown mean and unknown precision (inverse variance) is a normal-gamma distribution. This property also holds for the case of the conditional Gaussian distribution $p\left(t|\mathbf{x},\boldsymbol{w},\beta\right)=\prod_{n=1}^{N}\mathcal{N}\left(t_n|\boldsymbol{w}^T\phi\left(\mathbf{x}_n\right),\beta^{-1}\right)$ of the linear regression model. If we consider the likelihood function

$$p\left(\boldsymbol{t}|\mathbf{x},\boldsymbol{w},\beta\right)=\prod_{n=1}^{N}\mathcal{N}\left(t_n|\boldsymbol{w}^T\phi\left(\mathbf{x}_n\right),\beta^{-1}\right) \tag{4}$$

then the conjugate prior for $\boldsymbol{w}$ and β is given by

$$p\left(\boldsymbol{w},\beta\right)=\mathcal{N}\left(\boldsymbol{w}|\boldsymbol{m}_0,\beta^{-1}\mathbf{S}_0\right)\mathrm{Gam}\left(\beta|a_0,b_0\right) \tag{5}$$

(a) Show that the corresponding posterior distribution takes the same functional form, so that

$$p\left(\boldsymbol{w},\beta|\boldsymbol{t}\right)=\mathcal{N}\left(\boldsymbol{w}|\boldsymbol{m}_N,\beta^{-1}\mathbf{S}_N\right)\mathrm{Gam}\left(\beta|a_N,b_N\right) \tag{6}$$

and find expressions for the posterior parameters $\boldsymbol{m}_N$, $\mathbf{S}_N$, a_N and b_N.

(b) Show that the marginal probability of the data (model evidence) is given by

$$p\left(\boldsymbol{t}\right)=\frac{1}{2\pi^{N/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{\left|\mathbf{S}_N\right|^{1/2}}{\left|\mathbf{S}_0\right|^{1/2}}\frac{\Gamma\left(a_N\right)}{\Gamma\left(a_0\right)} \tag{7}$$

3 Mixture of conjugate priors

A. Show that a mixture of conjugate priors is indeed a conjugate prior.

B. Suppose we use the mixture prior $p\left(\theta\right)=0.5\mathrm{Beta}\left(\theta|a_1,b_1\right)+0.5\mathrm{Beta}\left(\theta|a_2,b_2\right)$, where $a_1=b_1=20$, $a_2=b_2=10$ and we observe N_1 head and N_0 tails. Derive an expression and write a computer code for evaluating the posterior. Consider $N_1=20$ heads and $N_0=10$ tails. Show with a plot a comparison between the prior and the posterior.

Table 1: Loss matrix for problem 4

predicted label $\hat{y}$	true label y	
	0	1
0	0	λ_{01}
1	λ_{10}	0

4 Optimal threshold on classification probability

Consider a case where we have learned a conditional probability distribution $p\left(y|\boldsymbol{x}\right)$. Suppose there are only two classes, and let $p_0 = p\left(y=0|\mathbf{x}\right)$ and $p_1 = p\left(y=1|\mathbf{x}\right)$. Consider the loss matrix shown in Table 1

A. Show that the decision $\hat{y}$ that minimizes the expected loss is equivalent to setting a probability threshold θ and predicting $\hat{y}=0$ if $p_1 < \theta$ and $\hat{y}=1$ if $p_1 \geq \theta$. Derive θ as a function of λ_{01} and λ_{10}.

B. Derive the loss function where the threshold is 0.1.

5 Bayes Factor

A. Suppose we toss a coin N times and observe N_1 heads. Let $N_1 \sim \mathrm{Bin}\left(N,\theta\right)$ and $\theta \sim \mathrm{Beta}\left(1,1\right)$. Show that the marginal likelihood is $p\left(N_1|N\right) = \frac{1}{N+1}$.

B. Suppose we toss a coin $N=10$ times and observe $N_1=9$ heads. Let the null hypothesis be that the coin is fair, and the alternative be that the coin can have any bias, so $p\left(\theta\right) = \mathcal{U}\left(0,1\right)$. Derive the Bayes factor $BF_{1,0}$ in favor of the biased coin hypothesis. What if $N=100$ and $N_1=90$?

6 Behavior of training set error with increasing sample size, Multi-output regression and Ridge regression

A. The error on the test will always decrease as we get more training data, since the model will be better estimated. However, for sufficiently complex models, the error on the training set can increase as we get more training data, until we reach some plateau. Explain why.

B. When we have multiple independent outputs in linear regression, the model becomes

$$p\left(\mathbf{y}|\mathbf{x},\mathbf{W}\right) = \prod_{j=1}^{M} \mathcal{N}\left(\boldsymbol{y}_j|\mathbf{w}_j^T\mathbf{x},\sigma_j^2\right) \tag{8}$$

Since the likelihood factorizes across dimensions, so does the MLE. Thus,

$$\hat{\mathbf{W}} = \left[\hat{\mathbf{w}}_1,\ldots,\hat{\mathbf{w}}_M\right] \tag{9}$$

where $\hat{\mathbf{w}}_j = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}_{:,j}$.

In this exercise, we apply this result to a model with 2 dimensional response vector $\boldsymbol{y}_i \in \mathbb{R}^2$. Suppose we have some binary input data $x_i \in \{0,1\}$. The training data is as follows Let us embed each x_i into 2d using the following basis function

x	y
0	$(-1,-1)^T$
0	$(-1,-2)^T$
0	$(-2,-1)^T$
1	$(1,1)^T$
1	$(1,2)^T$
1	$(2,1)^T$

$$\phi(0) = (1,0)^T, \quad \phi(1) = (0,1)^T . \tag{10}$$

The model becomes

$$\hat{\mathbf{y}} = \mathbf{W}^T\phi(\mathbf{x}) \tag{11}$$

where $\mathbf{W}$ is a 2×2 matrix. Compute the MLE for $\mathbf{W}$ from the above data.

C. Assume that $\hat{x} = 0$, so that the input data is centered. Show that the optimizer (in case of ridge regression) of

$$J(\boldsymbol{w}, w_0) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{w} - w_0\mathbf{1})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w} - w_0\mathbf{1}) + \lambda\boldsymbol{w}^T\boldsymbol{w} \tag{12}$$

is

$$\hat{w_0} = \bar{y} \tag{13}$$

$$\boldsymbol{w} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T y \tag{14}$$

D. For the data set provided in this link (file data_problem5d.mat), we want to fit a linear regression model with polynomial order M. In this regard, perform the following tasks:

(a) Compute the unknown coefficients based on MLE with $M = 2, 4, 10, 14$. Compute and plot the mean square error for the training and the test set (provided with the data set) corresponding to various polynomial orders .

(b) Compute the unknown coefficients based on ridge regression and plot the fitted function. Report the mean square error corresponding to the training and the test set (provided with the data set)

The data set contains `n` (number of training samples), `sigma2` (noise variance), `xtest` (test input), `ytest` (test output), `xtrain` (training input), `ytrain` (training output), `ytestNoisefree` (Noise free test output) and `lambdas` (regularizer for ridge regression).

7 Bayesian linear regression

We consider a Bayesian linear regression model to fit a set of points $x^i, y^i, i = 1, \dots, N$ data points with up to order $M = 5$ polynomials. For your implementation consider the data provided in this link (file data_problem7.mat). The data is generated using the following data generation algorithm:

x = 10*rand(N, 1);   generate input points
noise = 3*randn(N, 1); generate Gaussian noise
$y = (x - 4).^2 +$ noise;   actual truth function

The particulars of the regression model are given below:

$$p\left(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2\right) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2 \mathbf{I}_N\right), \tag{15}$$

where $\boldsymbol{\Phi}$ is the design matrix. The prior to be considered is as follows:

$$p\left(\boldsymbol{w}|\sigma^2, \boldsymbol{\Phi}\right) = \mathcal{N}\left(\mathbf{0}, \gamma\sigma^2\mathbf{I}\right), \tag{16}$$

where $p(\sigma^2) = InvGamma(a, b)$. The particular parameters that you should use in your implementation are $a = 0.1$, $b = 0.00001$ and $\gamma = 0.001/N$.

Most of the derivations needed for this problem can be found in the lecture notes but you are being asked to repeat them here.

A. Derive an expression for the posterior of $p(\boldsymbol{w}, \sigma^2|\mathcal{D})$, the marginal posterior $p(\boldsymbol{w}|\mathcal{D})$, predictive distribution $p(y|x, \mathcal{D})$ and the model evidence $p(\mathcal{D})$.

B. For polynomial orders $M = 1, 2, 3, 4, 5$ plot the predictive mean and the predictive error bars. Your plots should also indicate the exact function as well as the training data points.

C. Draw samples of $\boldsymbol{w}$ and for each of them show the predictive mean. What additional information this graph provides that is not given in your plots in B above?

D. Show a plot comparing the model evidence for polynomials $M = 1, \dots, 5$ and select the model that best represents the training data.

E. In Bayesian regression we should not regularize the bias term. This can be easily accomplished by using centered input ($\boldsymbol{\Phi}$) and output data ($\boldsymbol{y}$). The bias term can be computed and added to your solution through a post-processing operation. Using such a procedure, repeat the plot in B above for the optimal model selected in B.