CS 539 Machine Learning Homework 2

Conceptual and Theoretical Questions (6 questions, 50 pts)

1. Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

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

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 (i) data dependent noise variance and (ii) replicated data points (This is question 3.3, from Bishop textbook). **(6 pts)**

2. We showed in the class that 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 $p(t|\mathrm{x}, \mathrm{w}, \beta)$ of the linear regression model. If we consider the likelihood function (3.10),

$$p(\mathbf{t}|\mathbf{X}, \mathrm{w}, \beta) = \prod_{n=1}^{N} N(t_n|\mathrm{w}^{\mathrm{T}}\phi(\mathrm{x}_n), \beta^{-1})$$

then the conjugate prior for w and β is given by

$$p(\mathbf{w}, \beta) = N(w|m_o, \beta^{-1} * \mathrm{S}_0) Gam(\beta|a_0, b_0)$$

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

$$p(\mathbf{w}, \beta|\mathbf{t}) = N(w|m_N, \beta^{-1} * \mathrm{S}_\mathrm{N}) Gam(\beta|a_N, b_N)$$

and find expressions for the posterior parameters m_N, S_N, a_N, and b_N. (This is question 3.12, from Bishop textbook). **(10 pts)**

3. Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector **w** whose decision boundary $\mathrm{w}^{\mathrm{T}}\, \phi(x) = 0$ separates the classes and then taking the magnitude of **w** to infinity (This is problem 4.14, from Bishop textbook). **(4 pts)**

4. Show that the Hessian matrix H for the logistic regression model, given by (4.97), is positive definite. Here R is a diagonal matrix with elements $y_n(1 - y_n)$, and y_n is the output of the logistic regression model for input vector x_n. Hence show that the error function is a concave function of w and that it has a unique minimum. (This is problem 4.15, from Bishop textbook) **(10 pts)**

5. **Likelihood Estimate for Gamma regression**. Gamma distribution is defined by

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

a) Write down the probability in general form of the exponential distribution family, and find natural parameter (η), u(x), $h(x)$, and $g(\eta)$.

b) Let's assume we have a set of data points (t_i, x_i) i = 1 ... N, and we assume t_i follows a Gamma distribution where its mean is defined by

$$y_k = exp(w_0 + w_1 * x_k)$$

and the conditional distribution is

$$f(t_k|x_k) = \frac{1}{\Gamma(\nu)} * \left(\frac{\nu t_k}{y_k}\right)^\nu * \frac{1}{y} * e^{-\frac{\nu * t_k}{y_k}}$$

discuss how you can find maximum likelihood estimates of w_0 and w_1 using a gradient ascent algorithm. Derive the gradient and discuss whether the likelihood function is a concave function of the w_0 and w_1 or not. **(10 pts)**

6. **Laplacian Prior.** Laplacian prior for the weights of a linear (or logistic) regression will turn into Lasso regularization. Laplacian distribution on w is defined by

$$p(w) = \frac{1}{2b}\exp\left(-\frac{|w|}{b}\right)$$

which can be defined for weights of the model (except the intercept), where we assume different weigths are independent. b is a hyperparameter.

a) Let's assume we have $D = \{(x_i, t_i); i = 1 \dots N\}$ and we want to build a linear regression model with the Laplacian prior on the model weights. Define the likelihood and prior term here, and show it turns to a lasso regression. You can assume weights share the same hyperparameter.

b) Lasso regression is defined by

$$E_D(W) = -\frac{1}{2}\sum_{i=1}^{N}\left(t_i - W^T * \phi(\mathrm{x_i})\right)^2 + \lambda \sum_{j=1}^{M}|w_j|$$

We can use a gradient descent algorithm to find the model parameters, but the issue is that derivative of $|w|$ has a discontinuity at zero. A remedy is to rewrite the optimization by

$$E_D(W) = -\frac{1}{2}\sum_{i=1}^{N}\left(t_i - W^T * \phi(\mathrm{x_i})\right)^2 + \lambda \sum_{j=1}^{M}\frac{w_j^2}{|w_j|}$$

where, you replace the term in denominator of the regularization term by a known value. Let's assume, you are in the r^{th} iteration of a gradient descent algorithm (r represents the iteration), and your partial derivative for j^{th} weight is defined by

$$\frac{\partial E_D^{(r)}(W)}{\partial w_j} \approx \sum_{i=1}^{N}\phi(\mathrm{x_i})\left(t_i - W^{(r),T} * \phi(\mathrm{x_i})\right) + \lambda\, d_j^{(r)}\, w_j^{(r)}$$

$$d_j^{(r)} = \frac{1}{\max\{\epsilon, \left|w_j^{(r-1)}\right|\}}$$

where, ϵ has a small vale, like 0.0001. Complete the update rule for all other weights in the model and show its result in a simulated data.

Create 100 sample data points for $t_i = 1 + 0.001 * x_i - 2 * x_i^2 + \varepsilon_k$ where ε_k has normal distribution with a mean zero and variance of 0.1. Show how the estimated weights will change as a function of λ. For x, you can draw 100 random values from a normal distribution with mean 0 and variance 1. To find the model parameters you can use the gradient descent algorithm we discussed here. **(10 pts)**

Application Questions (2 questions, 50 pts)
Please note that you do not need to write the codes for different functions; you can use Linear Regression, GLM, QDA, and LDA tools in Matlab, Python, and R

Linear Regression Problem (25 points) We will use the Real estate valuation data set.xlsx dataset to build different regression models. The dataset includes 7 predictors, and one response variable. In the dataset, X1 represents the transaction date; find the earliest transaction date and set that as a reference time – ie, subtract that from all transaction dates.

1. Visualize the dataset; visualization step includes (not limited) to: a) histogram of individual predictors and response, and b) scatter plots for pairs of predictors and also pairs of predictor and response variable. Discuss you observation on possible outlier data points, and possible predictive power (correlation).
2. Build a linear regression model using all of predictors (no basis function) and discuss your model outcome (predicted weights, price as a function of time, and RMSE).
3. Build a Bayesian regression model using all of predictors (no basis function) and discuss your model outcomes (predicted weights and their confidence interval, prediction of response variable for the training set and their confidence interval). Build the model for different values of $\lambda = \frac{\alpha}{\beta}$ – here, you can pick: $\lambda_1 = 1$, $\lambda_2 = 10$, and $\lambda_3 = 100$.
4. Pick two predictors (X2 and X3) and standardize these features (Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation). Use 7 Gaussian basis with their means on $\mu = (-3, -2, -1, 0, 1, 2, 3)$ and $s = 1$ per each predictor. Build Bayesian regression model and discuss your prediction. Here, you can check the model performance for different values of $\lambda = \frac{\alpha}{\beta}$.
5. We want compare evidence of two different models. The one which is built using all of predictors (part 2) and the second one which is built based on 7 Gaussian basis for two predictors (part 4). We compare:
$$ln\ p(D) \cong ln\ p(D|\theta_{MAP}) - \frac{1}{2}\ M \ln N$$
which is defined in Bishop textbook chapter 4, equation (4.139). Discuss how you find θ_{MAP}, and which model has a higher evidence (note, the prior on θ is broad).
6. Cross validation technique is widely used in ML. In fact, cross validation is a sampling solution to what we are doing in part 5 using a Bayesian viewpoint. Using 10-fold cross validation, compare prediction accuracy (RMSE) of models we built in part 3 and 4. Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.

Dataset link: https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set

Classification Problem (25 points) We will use ENB2012_data.xlsx dataset to build different classifier. The dataset includes 8 predictors, and two response variables (y1, and y2). We create 1 and 0 label based on the difference between y1 and y2 (ie, y1-y2). If y1-y2 is larger than zero, we label that as 1 and otherwise 0 (this defines our class label).

1. Visualize the dataset; here, you plot the histogram of each predictor given two labels. Discuss which predictor have the highest predictive power. For discrete predictors, you can plot their pmf conditioned on the label.
2. Build a logistic regression model using all of predictors (no basis function) and discuss your model outcome (predicted weights, and classifier accuracy).
3. Using estimated weights in part 2, we have probability of two classes for each data point. Plot ROC curve for different threshold values (probability of class 1 larger than threshold, you will assign to class 1). Plot the accuracy as a function of threshold too.
4. Build a Bayesian regression model using all of predictors (no basis function) and discuss your model outcome (predicted weights and their confidence interval). Assume the prior is defined by $N(0, \alpha \mathrm{I})$, where α is the hyperparameter. Discuss the result for $\alpha_1 = 0.1$, $\alpha_2 = 1$, $\alpha_3 = 10$, and $\alpha_4 = 100$.
5. Do 10-fold cross validation for four different models you examined in part 4 and discuss which one has the highest decision accuracy. In the prediction step, use W_{MAP}.
6. We want to calculate

$$p(y = 1|\phi, \mathbf{t}) = \int p(y = 1|\phi, \mathrm{w})\, p(\mathrm{w}|\mathbf{t})\mathrm{dw}$$

using sampling technique (this is equation 4.145 from the Bishop text book). We derived a Gaussian approximation solution for $p(\mathrm{w}|\mathbf{t})$, where we can draw samples from it. We can then take these samples to calculate the probability. This is defined by:

$$p(y = 1|\phi, \mathbf{t}) \cong \frac{1}{S}\sum_{s=1}^{S} p(y = 1|\phi, \mathrm{w}^{\mathrm{s}})$$

where, w^{s} is the sample drawn from the posterior. For

$$X_1 = (0.8,600.0,286.0,138.1,5,4,0.25)$$

and

$$X_2 = (0.67,630.0,296.0,238.1,2,6,0.5)$$

Calculate probability for the models you built in part 4 for different αs.

Dataset link: https://archive.ics.uci.edu/ml/datasets/Energy+efficiency