1 Error *(8 points)*

Discuss whether the following statements are true or false.

- If the bias is high, increasing the training data will not help reduce the bias.
  (2 points)
- Reducing training error leads to reducing test error.
  (2 points)
- Increasing model complexity in regression always reduces the training error and increases the test error.
  (2 points)
- In a regression problem, When 6th degree polynomial regression results in a significant training error, linear regression should be used instead.
  (2 points)

2 Logistic Sigmoid Function *(10 points)*

We know $\sigma(a)$ is the logistic sigmoid function defined by

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

2.1 Show that the $\tanh$ function and the logistic sigmoid function are related by

$$\tanh(a) = 2\sigma(2a) - 1$$

(5 points)
2.2 Hence show that a general linear combination of logistic sigmoid functions of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x-\mu_j}{s})$$

is equivalent to a linear combination of $\tanh$ functions of the form

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh(\frac{x-\mu_j}{s})$$

and find expressions to relate the new parameters $\{u_1, ..., u_M\}$ to the original parameters $\{w_1, ..., w_M\}$.
(5 points)

3 Priors and Regularization *(15 points)*

Consider a model of Bayesian linear regression. Define the prior on the parameters as

$$p(w) = N(w|0, \alpha^{-1}\mathbf{I})$$

where α is as scalar hyperparameter that controls the variance of the Gaussian prior. Define the likelihood as

$$p(y|w) = \prod_{i=1}^{n} N(y_i|w^T x_i, \beta^{-1})$$

where β^{-1} is another fixed scalar defining the variance.
Using the fact that the posterior is the product of the prior and the likelihood (up to a normalization constant)

$$\arg\max_w \ln p(w|y) = \arg\max_w (\ln p(w) + \ln p(y|w))$$

Show that maximizing the log posterior is equivalent to minimizing a regularized loss function given by $L(w) + \lambda R(w)$, where

$$L(w) = \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

$$R(w) = \frac{1}{2} w^T w$$

4 Regression and Gradient Descent *(20 points)*

Suppose you have following model :

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 {x_1}^2 + w_4 {x_2}^2 + \epsilon \;\; where \;\; \epsilon \sim N(0, \sigma^2)$$

4.1 Write down an expression for $P(y|x_1, x_2)$
(4 points)
4.2 Assume you are given a set of training observations $({x_1}^{(i)}, {x_2}^{(i)}, y^{(i)}) for \;\; i = 1, 2, ..., n$ write down the conditional log likelihood of this training data. Drop any constants that do not depend on the parameters $\{w_0, ..., w_4\}$.
(4 points)
4.3 Write down a function $f(w_0, w_1, w_2, w_3)$ that can be minimized to find the desired parameter estimates.
(4 points)
4.4 Calculate the gradient of $f(w)$ with respect to the parameter vector w.
(4 points)

4.5 Write down a gradient descent update rule for w in terms of $\nabla_w f(w)$.
(4 points)

5 Linear Regression and SSE *(14 points)*

Assume n training data as $D = (x^{(1)}, y^{(1)}), ...(x^{(n)}, y^{(n)})$ which each x has a dimension of d. Consider a Linear Regression model and SSE as its cost function like below.

$$J(w) = \sum_{i=1}^{n} \left(y^{(i)} - w^T x^{(i)}\right)^2$$

5.1 Prove that :

$$w_{opt} = (X^T X)^{-1} X^T y$$

(5 points)
5.2 If we add $L2$ regularization term to SSE function , find the closed form solution for w.
(4 points)
5.3 Weighted Linear Regression is a generalization of ordinary least squares and linear regression in which knowledge of the variance of observations is incorporated into the regression. Find the closed form solution for w_{opt} for the cost functions below.

$$J(w) = \sum_{i=1}^{n} F_i \left(y^{(i)} - w^T x^{(i)}\right)^2$$

(5 points)

6 Decision Theory *(8 points)*

The squared loss is not the only possible choice of loss function for regression. Consider a situation in which the conditional distribution $p(t|x)$ is multimodal. In this case we use another loss function which expectation is given by :

$$E[L_q] = \iint |t - y(x)|^q p(x, t)\, dx\, dt$$

Write down the condition that $y(x)$ must satisfy in order to minimize $E[L_q]$. Show that for $q = 1$ this solution represents the conditional median. Then show that the minimum expected L_q loss for $q \to 0$ is given by the function $y(x)$ equal to the value of t that maximizes $p(t|x)$ for each x.

7 Practical *(25 points)*

Consider the $y = w_1 x + w_0$ linear regression problem. Assume we've got n train data , We are looking to minimize the following cost function:

$$\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - w^{(1)} x^{(i)} - w^{(0)}\right)^2$$

7.1 Write a python code that compute closed form solution for w_0 and w_1 on the first dataset by using gradient descent and stochastic gradient descent , then compare these two together.
7.2 Split second dataset into train and test data , Then Train a regression model on train data that satisfy following problems. (first 3 columns are features and the last one is the target).

- Train a 1st order regression model with SSE as cost function. Then report the w vector and error on train and test data.
- Train a 3rd order regression model with SSE as cost function. Then report the w vector and error on train and test data.
- Train a 3rd order regression model with SSE as cost function and $||w||_2$ as regularization term. Then report the w vector and plot the error for train and test data based on $ln(\lambda)$.

7.3 For determine the best λ implement k-fold cross validation. Then by using 10-fold cross validation plot the error on train and test data based on different λ values and then report the best λ.