UNIVERSITY COLLEGE LONDON
Faculty of Engineering Sciences

Department of Computer Science

Problem Set: Regression

Dr. Dariush Hosseini (dariush.hosseini@ucl.ac.uk)

Notation

Inputs:

$\mathbf{x} = [1, x_1, x_2, ..., x_m]^T \in \mathbb{R}^{m+1}$

Outputs:

$y \in \mathbb{R}$ for regression problems

$y \in \{0, 1\}$ for binary classification problems

Training Data:

$\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$

Input Training Data:

The design matrix, $\mathbf{X}$, is defined as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)T} \\ \mathbf{x}^{(2)T} \\ \cdot \\ \cdot \\ \mathbf{x}^{(n)T} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdot & \cdot & x_m^{(1)} \\ 1 & x_1^{(2)} & \cdot & \cdot & x_m^{(2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_1^{(n)} & \cdot & \cdot & x_m^{(n)} \end{bmatrix}$$

Output Training Data:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdot \\ \cdot \\ y^{(n)} \end{bmatrix}$$

Data-Generating Distribution:

Members of $\mathcal{S}$ are drawn i.i.d. from a data-generating distribution, $\mathcal{D}$

1. *This question is about polynomial regression, overfitting, and regularisation. To begin you are asked to spot that the data scenario with which you have been presented will result in an underdetermination of the weight vector which you seek. You are then asked to observe that this problem disappears if you restrict your feature space somewhat, and then to observe that the problem can also be remedied by particular forms of regularisation. Finally you are asked to consider the LASSO and how it gives rise to feature selection.*

   You are given a raw training data set consisting of 50 input/output sample pairs. Each output data point is real-valued, while each input data point consists of a vector of real-valued attributes, $\mathbf{x} = [1, x_1, x_2, x_3, x_4, x_5]^T$.
   You are asked to investigate the relationship between inputs and outputs.
   Your domain expertise leads you to believe that the class of cubic functions will be most appropriate for the modelling of this relationship.

   (a) ***[7 marks]***
   What is the problem with using the ordinary least squares approach to find such a model? Explain.

   (b) ***[3 marks]***
   If you discarded one of the attributes in your model would this problem remain? Explain.

   (c) ***[5 marks]***
   Briefly describe an approach that can help to remedy the problem in part (a). Explain.

   (d) ***[5 marks]***
   You are asked to select only those features which are most important for use in your model. Explain how you would achieve this in an efficient fashion.

2. *This question is about Bayesian linear regression. After being asked to derive the likelihood associated with a training set under a Gaussian additive noise model assumption, you are then asked to consider learning the weight vector (which is 1-dimensional in this case) in a Bayesian context. The weight vector is first elevated to being a random variable, after which you are asked to derive the posterior distribution of the weight vector (using Bayes' rule), given a Gaussian form for the prior distribution. Having done this you are asked to generate the MAP parameter estimate (this is the 'halfway house' approach to Bayesian analysis that we discussed in the Statistics lecture). At the close of the question you are asked to consider what other machine learning approach that MAP estimation in the context of Gaussian prior noise gives rise to. This was mentioned in lectures, but performing the analysis yourself should cement for you the connection between Bayesian approaches and regularisation in linear regression.*

Recall that in linear regression in general we seek to learn a linear mapping, $f_{\mathbf{w}}$, characterised by a weight vector, $\mathbf{w} \in \mathbb{R}^{m+1}$, and drawn from a function class, $\mathcal{F}$:

$$\mathcal{F} = \{f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} | \mathbf{w} = [w_0, w_1, ..., w_m]^T \in \mathbb{R}^{m+1}\}$$

Now, consider a data-generating distribution described by an i.i.d. Gaussian additive noise model:

$$y = \mathbf{w} \cdot \mathbf{x} + \epsilon \quad \text{where: } \epsilon \sim \mathcal{N}(0, \alpha), \ \alpha > 0$$

Assume that $\mathbf{x}$ is one dimensional, and that there is no bias term to learn, i.e. $\mathbf{x} = x \in \mathbb{R}$ and $\mathbf{w} = w \in \mathbb{R}$.

(a) *[4 **marks**]*
Given a training set, $\mathcal{S} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, find an expression for the likelihood, $\mathbb{P}(\mathcal{S}; w)$ in the form: $A \exp(B)$. Characterise A and B.

(b) *[4 **marks**]*
Assume a prior distribution, $p_{\mathcal{W}}(w)$, over w, such that each instance of w is an outcome of a Gaussian random variable, $\mathcal{W}$, where:

$$w \sim \mathcal{N}(0, \beta) \quad \text{where: } \beta > 0$$

Provide a detailed characterisation of the posterior distribution over w.

(c) *[4 **marks**]*
Hence succintly derive the Maximum A Posteriori (MAP) estimate of w.

(d) *[4 **marks**]*
Is this solution unique? Explain.

(e) *[4 **marks**]*
This approach should remind you of another machine learning algorithm. State the algorithm, and express the connection between the two.

3. *This question is about motivations for various linear regression models. You are asked to reproduce the Gaussian and Laplacian additive noise model motivations for 'ordinary least squares' (OLS) and 'least absolute deviations' (LAD) approaches to linear regression, before being asked to consider a key difference in behaviour between these two models pertaining to outliers. Finally, you are asked to consider the nature of the optimality associated with the LAD model. For this you should use the various concepts of convexity that we have discussed many times in lectures.*

In linear regression we seek to learn a linear mapping, $f_{\mathbf{w}}$, characterised by a weight vector, $\mathbf{w} \in \mathbb{R}^{m+1}$, and drawn from a function class, $\mathcal{F}$:

$$\mathcal{F} = \{f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} | \mathbf{w} = [w_0, w_1, ..., w_m]^T \in \mathbb{R}^{m+1}\}$$

One approach to learning $\mathbf{w}$ is to seek the optimal weight vector that minimises the empirical mean squared error loss across the training set.

(a) ***[1 mark]***
What is the name usually given to this approach?

(b) ***[1 mark]***
Write down the optimisation problem that this approach defines.

(c) ***[4 marks]***
Provide a motivation for this problem by considering an i.i.d. Gaussian additive noise model:

$$y^{(i)} = \mathbf{w} \cdot \mathbf{x}^{(i)} + \varepsilon^{(i)} \quad \text{where:} \quad \varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2), \sigma \in \mathbb{R}$$

(d) ***[6 marks]***
Let us alter the additive noise model which we have just considered so that the noise is now Laplace distributed, i.e.:

$$\varepsilon^{(i)} \sim \text{Laplace}(\mu, b) \qquad \text{where: } \mu = 0, b \in \mathbb{R}^+$$

The Laplace distribution is characterised by the following probability distribution function:

$$p_\epsilon(\epsilon = \varepsilon) = \frac{1}{2b} \exp\left(-\frac{|\varepsilon - \mu|}{b}\right)$$

What optimisation problem does this model imply? Explain.

(e) ***[2 marks]***
How will the treatment of outliers in the training data differ between these two models?

(f) ***[2 marks]***
Explain how you might solve this optimisation problem?

(g) ***[4 marks]***
Discuss the nature of this optimality.