ELL784: Minor Test

Sumeet Agarwal

September 27, 2022

Maximum marks: 20

Instructions:

- **Please clearly indicate the question number, and part number if applicable, at the start of each response.**
- **Please read all questions carefully.**
- **Please ensure that your responses are to-the-point and that you write only what is asked for on the answer script you submit.**
- **Please try to be clear and careful with all mathematical notation, so that there is no ambiguity in the expressions/formulae you write down. Try to stick to the notation used in class,** *e.g.*, **using an underbar to denote vector variables.**

Questions

1. Suppose you are seeking to model the *connection density* on Facebook between any two districts of India: i.e., out of all possible friendships that could exist between those two districts, what fraction actually exist? This can be thought of as a regression problem: let each data point represent a pair of districts, and consist of one feature, denoted for the n^{th} data point

   x_n – the distance between the centres of the two districts (in km);

   and one label

   t_n – the Facebook connection density between the n^{th} pair of districts.

   I would like to model the relationship between the label and the features probabilistically, just like we did for curve fitting in class. For the deterministic part of the model, I assume that the *expected* connection density between a pair of districts is inversely proportional to the distance between them. So

   $$y(x_n; w) = \frac{w}{x_n}.$$

   Note that w is scalar, as there is only one parameter here.

   For the probabilistic part, I assume that the variation or *noise* around the expected value follows a zero-mean Gaussian distribution with precision β. This leads to the following overall model:

   $$p(t_n|x_n, w, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta(t_n - w/x_n)^2}{2}\right),$$

   Given the above modelling setup, please answer the following questions, showing all working clearly and precisely.

1.1 Given a data set $X = \{x_1, ..., x_N\}, \mathbf{t} = (t_1; ...; t_N)$, which represents a set of district pairs for which you know the feature and label values, write down the expression for the likelihood as a function of the model parameter, *i.e.*, $\mathcal{L}(w)$. **[1.5]**

1.2 How will you convert this likelihood into a convenient error function, $E(w)$? Write down an expression for this $E(w)$. **[2]**

1.3 Use the error function you have just obtained to derive the maximum likelihood estimate for the model parameter, *i.e.*, $\hat{w}_{ML}$. **[2.5]**

1.4 Try to interpret the estimate just obtained – explain, in words, what it is capturing about the data and why it makes sense. (Hint: what does it capture if you just have one data point?) **[1]**

1.5 Now suppose I wish to carry out Bayesian inference of w, and for this purpose use a zero-mean Gaussian prior with precision α:

$$p(w|\alpha) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha w^2}{2}\right),$$

Using this prior and for the above given data set and probabilistic model, write down an expression for the posterior over w. **[2]**

1.6 Convert the above expression for the posterior into a convenient error function, $\tilde{E}(w)$. Write down the expression for this $\tilde{E}(w)$. **[2]**

1.7 Use the error function just obtained to derive the maximum a posteriori estimate for the model parameter, *i.e.*, $\hat{w}_{MAP}$. **[2.5]**

1.8 How can you control the strength of the prior? **[1]**

2. Suppose you are seeking to fit a regression function of the form

$$y(x; \mathbf{w}) = w_0 + w_1 x + w_1^2 x^2$$

to a data set consisting of feature-label pairs (x_n, t_n), using sum-of-squares error with quadratic or L2 regularisation.

2.1 Obtain the *stochastic gradient* vector of the regularised error function with respect to $\mathbf{w}$, using a single data point (x_n, t_n). Show your working clearly. **[2]**

2.2 Suppose you want to learn $\mathbf{w}$ via stochastic gradient descent. Write down the update rules for each of the weights from iteration τ to iteration $\tau + 1$; *e.g.*,

$$w_0^{(\tau+1)} = w_0^{(\tau)} + \_\_\_\_\_\_,$$

where you need to fill in the blank. Similarly for the other weights. **[2]**

3. Which of the following will generally have the effect of lowering the *variance* of a polynomial regression model, which is trained via gradient descent (assume a sufficiently small learning rate that the algorithm will not jump over minima)? Specify all that apply. **[1.5]**

(a) Reducing the degree of the polynomial; (b) Increasing the degree of the polynomial; (c) Reducing the regularisation strength; (d) Increasing the regularisation strength; (e) Reducing the number of training iterations; (f) Increasing the number of training iterations.