ELL409: Minor Test

Sumeet Agarwal

September 21, 2021

Maximum marks: 24

Instructions:

- **Please clearly indicate the question number, and part number if applicable, at the start of each response; and correspondingly, indicate the page numbers corresponding to each question when uploading your answer script onto Gradescope.**
- **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. Please write out the below pledge on your answer script, and add your signature underneath to indicate your assent. This will be needed for your responses to be graded.

   I pledge to attempt all questions in this test on my own, without seeking assistance from anyone or copying from anywhere.

2. Suppose you are seeking to model the number of Facebook connections between any two districts of India. This can be thought of as a regression problem: let each data point represent a pair of districts, and consist of two features, denoted for the n^{th} data point

   x_{n1} – population of the first district, and
   x_{n2} – population of the second district;

and one label

t_n – the number of Facebook connections 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* number of connections between a pair of districts is proportional to the product of their populations. For the probabilistic part, I assume that the variation or *noise* around the expected value follows a Poisson distribution. This leads to the following overall model:

$$p(t_n|\mathbf{x}_n; w) = \frac{(wx_{n1}x_{n2})^{t_n} e^{-wx_{n1}x_{n2}}}{t_n!},$$

where $\mathbf{x}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix}$. Note that w is scalar, as there is only one parameter here.

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

2.1 Given a data set $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}, \mathbf{t} = (t_1; ...; t_N)$, 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]**

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

2.3 Use the error function you have just obtained to derive the maximum likelihood estimate for the model parameter, *i.e.*, $\hat{w}_{ML}$. **[2]**

2.4 Try to interpret the estimate just obtained – explain, in words, what it is capturing about the data and why it makes sense. **[1.5]**

2.5 Now suppose I wish to carry out Bayesian inference of w, and for this purpose use a Gamma prior:

$$p(w|\alpha, \beta) = \frac{\beta^\alpha w^{\alpha-1} e^{-\beta w}}{(\alpha - 1)!},$$

where α, β are hyperparameters which we take to be positive integers. Using this prior and for the above given data set and probabilistic model, write down an expression for the posterior over w. **[2]**

2.6 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]**

2.7 Use the error function just obtained to derive the maximum a posteriori estimate for the model parameter, *i.e.*, $\hat{w}_{MAP}$. **[2]**

2.8 Suppose you want your prior to encode the belief that about 0.1% of all possible connections/friendships between the populations of any 2 districts actually exist on Facebook. In this case, what should be the relation between α and β? **[1]**

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

3. Explain, in your own words, what the Bayesian perspective is on the bias-variance trade-off in machine learning. How do Bayesian models seek to control this trade-off, and what approach might you use in a Bayesian setup to try and find the 'sweet spot' in between high bias and high variance? **[3]**

4. Suppose you are seeking to fit a second-order polynomial of the form

$$y(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2$$

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

4.1 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]**

4.2 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.*,

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

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

4.3 Based on the above update rules, what role does L2 regularisation play in the gradient descent updates? Why is this useful? **[2]**