Algorithms for Learning and Inference: Final Exam

Instructor: Morteza H. Chehreghani

Due: See Canvas

- NOTE 1. You explain your solutions, for any question that the calculations are needed you must show the steps (for anything more advanced than +-*/ which you can do with a calculator).
- NOTE 2. You must submit your solution to Canvas, in the same way as the assignments.
- NOTE 3. The exam must be done individually. You may not receive help from anyone else.
- NOTE 4. Your submission must be in pdf format. You may either type your solutions in latex/word and submit a pdf file, or take the photo/scanning of the handwritten solutions and upload the pdf file. If you take photos, make sure that it is easy to read and that you combine photos into a single pdf file such that each page appears in the right order. There are both command line and online tools to do this.
- NOTE 5. Read the questions carefully such that you do not miss any question and ensure you clearly give the answer required for each (sub)question.
- NOTE 6. You do not need to write (Python) code for any question. Your submitted solution should not include any (Python) code.
- The maximum score of this exam is 75. Your score will be normalized to be between 0 and 60. Then your grade will be computed according to the formula in Canvas.
- For questions contact: Morteza Haghir Chehreghani and Arman Rahbar.

1. (20 points) We are given the dataset $\mathbf{D}$ with N data points. Each data point i has two variables: x_i and t_i where both of them are real numbers. Thus, $\mathbf{D} = \{(x_1, t_1), ..., (x_N, t_N)\}$. We use the following Gaussian model to fit to the data.

$$y_i \sim \mathcal{N}(\exp(\theta x_i), 1). \tag{1}$$

Its unknown parameter is θ and the variance is set in advance to 1. Assume the data points are i.i.d.

   (a) (5 points) Write down the full log-likelihood for the entire dataset. Your final answer should be in the following form.

$$N \times .... + \sum_{i=1}^{N} .... \tag{2}$$

   (b) (2 points) Why may one prefer log-likelihood instead of likelihood?

   (c) (5 points) Complete the following equation for the maximum log-likelihood solution. Write down your calculations.

$$\sum_{i=1}^{N} (\exp(2\theta x_i) \times ....) = \sum_{i=1}^{N} .... \tag{3}$$

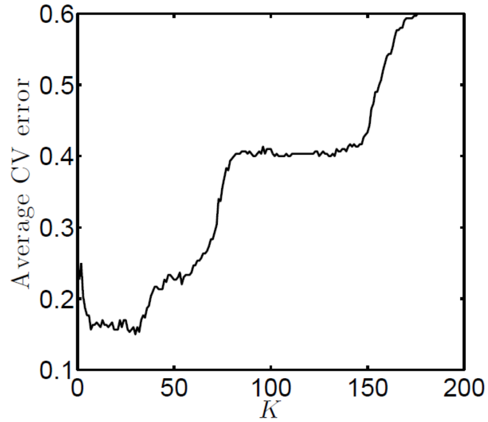

Figure 1: K-nearest neighbor classification with large steps.

(d) (4 points) Does changing the model variance from 1 to a positive real number such as a affect the equation 3 for optimal θ? Explain your answer.

(e) (4 points) Assume we want to obtain the same optimal θ but via minimizing a loss function. Then, what would be the model you use? What loss function would you use?

2. (15 points) Consider SVM and K-nearest neighbor classification methods and answer the following questions.

(a) (4 points) What is the *minimum* number of support vectors in a dataset of two classes and N data points when we apply hard SVM to classify it? Draw a picture to show such a dataset. Assume the two classes are linearly separable.

(b) (4 points) What is the *maximum* number of support vectors in a dataset of two classes and N data points when we apply hard SVM to classify it? Draw a picture to show such a dataset. Assume the two classes are linearly separable.

(c) (3 points) What is the impact of choosing $K = 2$ in K-nearest neighbor classification?

(d) (4 points) When we look at the K-nearest neighbor cross validation error (or the test error) as a function of K, we may observe large steps. See for example Figure 1. Explain the reason for such large steps.

3. (10 points) Remember the MAP estimate used in Bayesian Logistic Regression and assume the number of classes is 2.

(a) (4 points) For the case of a uniform prior distribution, can we minimize the cross entropy to obtain the MAP estimate of the parameters? Explain your answer.

(b) (3 points) What is a main limitation of the MAP estimate solution that makes us use other methods such as Laplace approximation?

(c) (3 points) In both Laplace approximation and Metropolis-Hastings methods we use sampling to predict the class label for a new data point. What is the difference between the sampling used in these two methods?

4. (13 points) Consider the neural network shown in Figure 2 which acts on the input data $\mathbf{X}$ of size N and produces the output $\hat{\mathbf{y}}$. Assume that the input is three dimensional and x_1, x_2 and x_3 represent the three features of an input data point x.

In this model, w_1 , w_2 , w_3 , w_4 , w_6 , w_7 and w_8 are the (unknown) parameters of the model that should be estimated using the data. The activation function is defined as $f(z) = \sin(z)$.

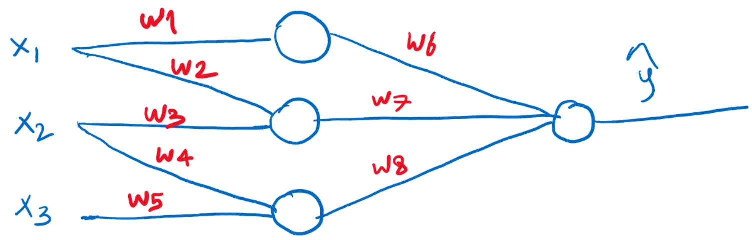

Figure 2: The neural network model.

The error of the network is measured by

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{4}$$

where y_i and $\hat{y}_i$ respectively correspond to the true and predicted outputs for the i-th data point.

(a) (5 points) Use backpropagation and write down the gradients of the error $\mathcal{E}$ with respect to all the different unknown parameters. Show an outline for your derivations. You do not need to compute the exact derivatives, but sufficiently describe the outline.

(b) (4 points) Describe an optimization procedure using the gradients to estimate the parameters.

(c) (4 points) Now consider a more complex neural network model composed of CNN and RNN in some way. Briefly explain how such a model can be used for image captioning.

5. (17 points) Consider a Gaussian Mixture Model (GMM) with K components applied to a dataset of N d-dimensional data points.

(a) (3 points) Write down the likelihood for a single data point and extend it to full log-likelihood for the entire dataset.

(b) (5 points) Except the covariance matrix, compute the free (unknown) parameters of the model for $K = 1$. Here assume $d = 1$ and write down the detail of your calculations.

(c) We apply AIC and BIC to obtain the correct number of clusters. Identify the number of free parameters (i.e., c_K) in each of the following settings (d can be any natural number).

- (3 points) The covariance matrices are known and given in advance.
- (3 points) We only know that the covariance matrices are diagonal.
- (3 points) We only know that the covariance matrices are positive semidefinite.