TDA 231 Machine Learning 2018: Final Exam

Instructors: Morteza Chehreghani and Devdatt Dubhashi

Due: 4 PM, Room 6446, May 30, 2018

Instructions

1- Hand in the exam to Morteza Chahreghani (office no. 6446) at 1600 on May 30th.

2- Direct any questions/clarifications on exam by email to Morteza Chehreghani: morteza.chehreghani@chalmers.se

1. (16 points) Consider a dataset of N items, in which the i^{th} item has two elements: one real-valued input x_i and the respective output y_i, i.e., the set $\{(x_1, y_1), ..., (x_i, y_i), ..., (x_N, y_N)\}$.
   We use the following normal (Gaussian) model to fit the data, which has an unknown parameter w (the variance is known in advance and is set to 1).

$$y_i \sim \mathcal{N}(\log(wx_i), 1).$$

   (a) Describe a maximum likelihood approach to infer w and write down the log-likelihood objective for this problem. [5 points]

   (b) Complete the right side of the following equation for the maximum log-likelihood solution. Write down your calculation. [4 points]

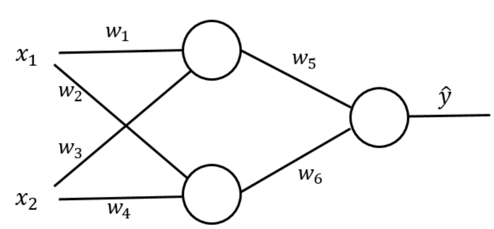
$$\sum_{i=1}^{N} y_i = \text{...}$$

   (c) We add $-\alpha||w||_2^2$ to the full log-likelihood objective, where the hyperparameter α is fixed in advance ($\alpha \geq 0$). Complete the equation for the maximum log-likelihood solution of this new objective. [4 points]

$$\sum_{i=1}^{N} y_i = \text{...}$$

   (d) What does happen if $\alpha \to \infty$ in the new objective of part (c)? [3 points]

2. (19 points) Given a set of N triplets $\{(x_{i1}, x_{i2}, y_i)\}, 1 \leq i \leq N$, the goal is to design a model to predict y_i based on the input attributes x_{i1} and x_{i2}. For this, we use the following neural network.



   In this model, $w_1, ..., w_6$ are the free parameters that should be learned. The (nonlinear) activation is defined as (q is a hyperparameter which is fixed in advance):

$$f(z) = \begin{cases} (z + |z|)^q & \text{if } z \geq 0 \\ (z - |z|)^q & \text{otherwise} \end{cases}$$

The error of the netwrok is measured by

$$\mathcal{E} = \frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2.$$

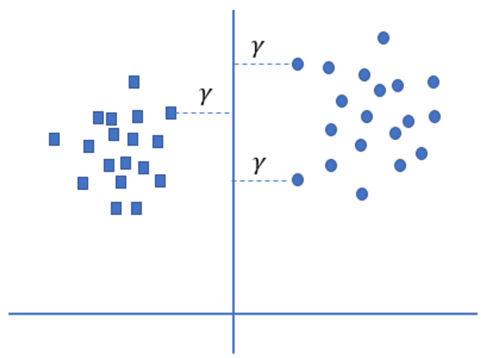
(a) Write down the gradients of the error $\mathcal{E}$ with respect to all the parameters. Show an outline of your derivation (you do not need to compute the exact derivatives, but sufficiently describe the outline). [5 points]

(b) Describe a gradient descent algorithm to estimate the parameters. [5 points]

(c) For $q = 1$, can you derive an equivalent but simpler neural network (i.e., a network without a hidden layer)? Prove your answer. [5 points]

(d) For $q = 1$, is the model equivalent to a linear regression model? Explain your answer. [4 points]

3. (12 points) Consider the dataset shown in the following figure, where the goal is to classify the squares and circles using a (hard margin) linear SVM model.



(a) What would be the training error of the optimal linear SVM? Explain your answer. [3 points]

(b) Due to computational bottlenecks, we pick a subset of the items which would yield exactly the same solution as the SVM on the original data. Which items would you choose (mark them)? Explain your answer. [5 points]

(c) Assume the data dimensionality is d and the number of training data points is N. Then, what is the computational complexity for predicting the class label of a new test data point? [4 points]

4. (13 points) Clustering/unsupervised learning

(a) Consider the K-means cost function for clustering N $d-$dimensional data points into K clusters. Compute the optimal parameters when $K = 1$. [4 points]

(b) Derive the optimal parameters of the model when $K = N$. [4 points]

Now, we use Gaussian mixture models (GMM) to cluster the data, wherein the covariance matrices are fixed and given in advance. In addition, we aim to obtain the correct number of clusters via Akaike information criterion (AIC) and Bayesian information criterion (BIC).

(c) Assume we know that all the estimated clusters should have the same size. Describe how you would apply AIC and BIC to compute the correct number of clusters. Discard the steps for calculating the log-likelihoods. [5 points]