Machine learning for data science I

7 September 2022

Surname, name (all caps) _____

Student ID: _____

This is a closed book exam.

Write clearly and justify your answers.

Time limit: 90 min.

Question:	1	2	3	4	5	Total
Points:	20	20	20	20	20	100
Score:						

1. Somebody was preparing a data set that consists of several features with real values and a real target variable. He accidentally duplicated all the features. How does such mistake affect the following models (besides the increased time complexity due to a larger data set)? Explain your answers.

[3] (a) k-nearest neighbors with Euclidean distance

[3] (b) regression tree

[4] (c) random forest

[3] (d) linear regression

[4] (e) L2-regularized linear regression

[3] (f) linear support-vector machine

2. We have n samples x_i drawn from a Geometric distribution $f(x; p) = p(1-p)^x$

[15] (a) Estimate p using maximum likelihood estimation if the samples are iid from Geometric distribution.

[5] (b) Suppose that the samples are still independent but not identically distributed. We will assume that each sample x_i comes from its own Geometric distribution parameterized by p_i. What are the estimates of p_i using MLE in this case?

3. Answer the following questions about the SVM method.

[4] (a) How does SVM deal with non-separable data points?

[6] (b) How many points can a *single side* of the margin touch: 0, 1, 2? Explain your answer for each number (consider also non-separable cases).

[5] (c) What is the purpose of hyperparameter C? What is the effect of very small C and very large C?

[5] (d) SVM does not directly provide probability estimates. How would you adapt the method to output some estimate of the probability besides the predicted class?

4. Dendrograms rely on increasing distances between clusters in the process of hierarchical clustering. Describe each linkage and explain why the distances between clusters don't decrease in the process of hierarchical clustering or find a counter-example:

[6] (a) Single linkage

[6] (b) Complete linkage

[8] (c) Average linkage

5. We want to model a problem with L2-normalized linear regression (ridge regression). However, the individual data instances are know to be of different importance. Their importance is given by weights w_i. The loss function that we want to minimize is therefore

$$L(\beta) = \sum_{i=1}^{n} w_i (\sum_{j=1}^{m} \beta_j x_{i,j} - y_i)^2 + \lambda \sum_{j=1}^{m} \beta_j^2$$

[7] (a) Write the loss function in a matrix form. Use the following:

- X is a matrix of size $n \times m$, which describes n data instances with m features
- β and y are a column vector of m model coefficients and n target values, respectively
- W is a diagonal matrix with instance weights w_i on the diagonal

[13] (b) Derive a closed-form solution for β.

Help: $\frac{\partial(v^T A)}{\partial v} = A, \quad \frac{\partial(v^T A v)}{\partial v} = 2Av, \quad v^T v = v^T I v.$