CUB Spring 2024. Machine Learning.
Mid-term exam. Test variant.

Exam rules:

(a) The exam duration time: 120 minutes.

(b) The exam is written on student's own paper. On the first page student's name in printed letters and matriculation number should be given.

(c) One A4 page cheat sheet (two-sided) is allowed.

(d) The total number of points is 27. The exam grade is computed as a sum of points for all tasks divided by 22. So for getting 100% it is enough to get in total 22 points out of 27. In order to get the minimum grade 45% it is needed to get in total 10 points.

(e) In tasks with given set of answers the number of correct answers may vary (it could be one answer or multiple answers). Only giving correct answers here without any explanation is not enough for getting positive grades for this tasks.

Task 1 (1 pts) Give your own example of practical problem (not from the lectures) that can be solved as a supervised machine learning problem. Describe what are input objects, target variables and speculate on possible features.

Task 2 (1 pts) Give your own examples (not from the lectures) of nominal feature and set-valued feature.

Task 3 (2 pts) Consider the following function:

$$F(X) = \frac{1}{2}\|X - X_0\|_F^2 + \boldsymbol{\lambda}^T(X\boldsymbol{a} - \boldsymbol{b}).$$

Here $X, X_0 \in \mathbb{R}^{n\times m}$, $\boldsymbol{\lambda}, \boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{a} \in \mathbb{R}^m$. Using the technique of differentials find the gradient of $F(X)$ w.r.t. matrix X and find the minimum point X_{min} of the function F by solving the equation $\nabla F(X) = 0$.

Task 4 (1 pts) Suppose that x is one feature in the dataset, m – its mean, s – its variance, M – its median, S – its median of absolute deviations of feature values from its median, min and max – minimal and maximal feature values. Choose feature normalization method that is stable to the presence of outliers:

(a) $(x - m)/s$;

(b) $(x - \min)/(\max - \min)$;

(c) $(x - m)/(\max - \min)$;

(d) $(x - M)/(\max - \min)$;

(e) $(x - M)/s$;

(f) $(x - M)/S$.

Explain your answer.

Task 5 (1 pts) What are the reasons for data normalization (transforming of features to the same scale) prior to running gradient descent optimizer for training of linear ML model?

(a) to provide convergence with fewer number of iterations;

(b) to provide convergence to a local minima with lower value of loss function;

(c) to provide convergence to a local minima with wider vicinity of low loss function values;

(d) to accelerate computation of one optimization iteration.

Explain your answer.

Task 6 (1 pts) Consider the situation when two features have exactly the same values in training set. Linear regression with what regularization would give the same weights for both features?

(a) L_1-regularization;

(b) L_2-regularization;

(c) Elastic Net regularization.

Explain your answer.

Task 7 (1 pts) Suppose that we are finding a constant prediction for regression problem with MSE loss function. Show that the optimal constant prediction is a mean of target values.

Task 8 (1 pts) Suppose that we minimize some function using stochastic gradient descent. Let's take some constant value for learning rate. Is it enough to get the optimal point with number of iterations going to infinity?

(a) yes, if this constant is small enough;

(b) yes, if this constant is big enough;

(c) yes, regardless of constant value;

(d) no, it is required to dynamically reduce learning rate.

Explain your answer.

Task 9 (1 pts) Suppose we build a classifier for some disease identification (gives positive class in case of disease's presence and negative class for healthy person). For some dataset our classifier correctly identifies 4 persons as sick, 2 persons as healthy, one healthy person is identified as sick one and three sick persons are identified as healthy ones. Compute accuracy and F_1 measure.

Task 10 (1 pts) Suppose you develop a test for identification of some dangerous disease (positive class). The number of healthy people is sufficiently larger that the number of ill people. In case of test's false alarm there exist additional more precise identification procedures. Choose the most adequate quality metric for your test:

(a) F_1-measure

(b) Precision

(c) Recall

Explain your answer.

Task 11 (2 pts) Suppose that for some two-class classifier its prediction scores for some dataset are given by

$$[0.9, 0.7, 0.7, 0.7, 0.5, 0.6, 0.1, 0.9, 0].$$

The corresponding class labels are

$$[-1, -1, 1, 1, -1, -1, 1, 1, -1].$$

Plot ROC-curve.

Task 12 (2 pts) Suppose that your have trained an SVM model with some hyperparameters and want to reduce the number of errors on the training set. Choose the options that may help you in achieving this goal:

(a) increase coefficient C;

(b) decrease coefficient C;

(c) increase polynomial degree d in polynomial kernel function $K(\boldsymbol{x}, \boldsymbol{x}) = (\boldsymbol{x}^T \boldsymbol{y})^d$;

(d) increase polynomial degree d in polynomial kernel function;

(e) increase σ parameter in RBF kernel function $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/(2\sigma^2))$;

(f) decrease σ parameter in RBF kernel function.

Explain your answer.

Task 13 (2 pts) Suppose that we estimate the performance of SVM model using leave-one-out procedure. Suppose that the model has M support vectors when trained on the full dataset. Then the number of errors in leave-one-out

(a) doesn't exceed M;

(b) not lower than M;

(c) may be both more than M and less than M.

Explain your answer.

Task 14 (2 pts) Let's consider the following constrained optimization problem:

$$\boldsymbol{c}^T \boldsymbol{x} \to \min_{\boldsymbol{x}},$$
$$\|\boldsymbol{x}\|_2^2 \leq b.$$

Here $\boldsymbol{x}, \boldsymbol{c} \in \mathbb{R}^n$, $b > 0$. Construct the dual optimization problem.

Task 15 (1 pts) Suppose we consider a two-class SVM model with RBF kernel function $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2/(2\sigma^2))$. Write down the decision rule for determining class label for testing object $\boldsymbol{x}$. Estimate computational complexity of this rule if number of features is D and number of support vectors is M.

Task 16 (1 pts) Let's consider training a linear classifier for two classes with hinge loss function:

$$L(M) = \max(0, 1 - M), \; M = y\boldsymbol{w}^T \boldsymbol{x}$$

Write down one iteration of stochastic gradient descent with one object in a mini-batch and learning rate α.

Task 17 (1 pts) Suppose that in two-class classification problem we need to construct an algorithm that outputs $a_i \in [0, 1]$ for i-th input object. Suppose that the error is measured as follows:

$$(y_i - a_i)^2 0.9, \text{ if } y_i = 1,$$
$$(y_i - a_i)^2 0.1, \text{ if } y_i = 0.$$

What would be the optimal constant prediction here if probability of class $y_i = 1$ is p?

Task 18 (2 pts) Consider the *geometric* distribution. This a discrete distribution where random variable x can take values $1, 2, 3, \ldots$ with probabilities

$$p(x|q) = (1 - q)^{x-1} q.$$

Here $q \in (0, 1)$ – parameter of the distribution. Suppose we have an independent and identically distributed samples from this distribution:

$$x_1, x_2, \ldots, x_N \sim p(x|q)$$

Find maximal likelihood estimate for the parameter q.

Task 19 (1 pts) Suppose we solve a multi-class classification problem using all-vs-all approach with linear classifiers. How many parameters should we train if number of classes is K and number of features is D? Explain your answer.

Task 20 (2 pts) Consider construction of Decision Tree using Gini index as impurity criterion. Suppose that in some node we have 8 training objects and we sort them according to some feature value. Suppose that class labels after the sorting are the following:

$$y_1 = y_2 = y_3 = +1, \; y_4 = -1, \; y_5 = y_6 = y_7 = y_8 = 1.$$

Which threshold would be chosen for splitting in this case? Justify your answer by computing impurity reduction criterion for different splits.