

## CS 4210 – Assignment #3

### Maximum Points: 100 pts.

Bronco ID:

Last Name: \_\_\_\_\_

First Name: \_\_\_\_\_

**Note 1:** Your submission header must have the format as shown in the above-enclosed rounded rectangle.

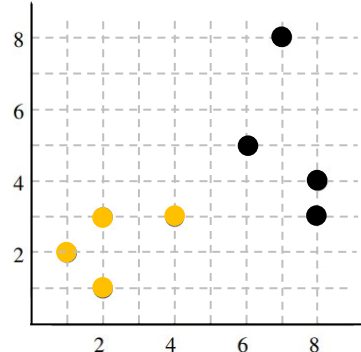
**Note 2:** Homework is to be done individually. You may discuss the homework problems with your fellow students, but you are NOT allowed to copy – either in part or in whole – anyone else's answers.

**Note 3:** Your deliverable should be a .pdf file submitted through Gradescope until the deadline. Do not forget to assign a page to each of your answers when making a submission. In addition, source code (.py files) should be added to an online repository (e.g., github) to be downloaded and executed later.

**Note 4:** All submitted materials must be legible. Figures/diagrams must have good quality.

**Note 5:** Please use and check the Canvas discussion for further instructions, questions, answers, and hints. The bold words/sentences provide information for a complete or accurate answer.

1. [20 points] Say you are given the training dataset shown below. This is a binary classification task in which the instances are described by two integer-valued attributes.



Dataset

- a. [2 points] Draw the decision boundary and its parallel hyperplanes for a linear SVM with maximum margin (hard margin formulation) and identify the support vectors.
- b. [2 points] If a black circle is added as a training sample in the position (7,5), does this affect the previously learned decision boundary? Explain why.
- c. [2 points] If a yellow circle is added as a training sample in the position (4,2), does this affect the previously learned decision boundary? Explain why.
- d. [2 points] If a black circle is added as a test sample in the position (7,5), will this sample be classified correctly according to the previously learned decision boundary? Explain why.
- e. [2 points] If a black circle is added as a test sample in the position (6,4), will this sample be classified correctly according to the previously learned decision boundary? Explain why.
- f. [2 points] If a yellow circle is added as a test sample in the position (4,2), will this sample be classified correctly according to the previously learned decision boundary? Explain why.
- g. [2 points] If a yellow circle is added as a test sample in the position (5,3), will this sample be classified correctly according to the previously learned decision boundary? Explain why.

- h. [2 points] If a black circle is added as a test sample in the position (5,3), will this sample be classified correctly according to the previously learned decision boundary? Explain why.
- i. [2 points] If a yellow circle is added as a test sample in the position (6,4), will this sample be classified correctly according to the previously learned decision boundary? Explain why.
- j. [2 points] If a black circle is added as a training sample in the position (4,4), how this will affect the decision boundary if  $C = 1$  and  $C = \infty$ ? Consider the soft margin formulation.

2. [20 points] Consider the following 1-dimensional data with two classes:

$x$	-3	0	1	2	3	4	5
Class	-	-	+	+	+	+	+

- a. [5 points] Find the decision boundary of a linear SVM on this data (hard-margin formulation) and identify the support vectors (write the  $x$  coordinate to provide your answer).
  - b. [5 points] Find the solution parameters  $w$  and  $b$  for this linear SVM and the width of the margin. Hint: place the identified support vectors (positive and negative) into the formula  $y_i(w \cdot x_i + b) = 1$  since you know this formula holds for them.
  - c. [5 points] Show mathematically that the SVM classifications for the test data  $\{-1.5, 1.5\}$  are negative and positive respectively.
  - d. [5 points] Suppose we remove the point (1,+) from this training set and train the SVM again. Find the new values of the solution parameters  $w$  and  $b$  and the width of the margin.
3. [20 points] The quadratic kernel  $K(x, y) = (x \cdot y + 1)^2$  should be equivalent to mapping each  $x$  into a six-dimensional space where

$$\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

for the case where  $x = (x_1, x_2)$ . Demonstrate this equivalence by answering the following questions while using the data points:  $A = (1, 2)$ ,  $B = (2, 4)$ .

- a. [5 points]  $\Phi(A)$
  - b. [5 points]  $\Phi(B)$
  - c. [5 points]  $\Phi(A)\Phi(B)$
  - d. [5 points]  $K(A, B)$ . Hint: your answers for (c) and (d) should be the same. By using the kernel function, SVM “cheats” and performs significantly fewer calculations (kernel trick).
4. [15 points] Complete the Python program (svm.py) that will read the file optdigits.tra (3,823 samples) that includes training instances of handwritten digits (optically recognized). Read the file optdigits.names to get detailed information about this dataset. Also, check the file optdigits-orig.tra and optdigits-orig.names to see the original format of this data, and how it was transformed to speed-up the learning process (pre-processing phase). Your goal is to build multiple SVM classifiers using this data. You will simulate a grid search, trying to find which combination of four SVM hyperparameters (c, degree, kernel, and decision\_function\_shape) leads you to the best prediction performance. To test the accuracy of those distinct models, you will use the file optdigits.tes (1,797 samples).

5. [25 points] Consider the dataset below.

Outlook	Temperature	PlayTennis
Sunny	Hot	No
Overcast	Cool	Yes
Overcast	Hot	Yes
Rain	Cool	No
Overcast	Mild	Yes

We will apply *steady state* Genetic Algorithms ( $r = 0.5$ ) to solve this classification problem. If the instance matches the rule pre-condition of the chromosome, you predict according to its post-condition, otherwise you predict the opposite class defined by the chromosome (there must be a prediction for each instance then). Follow the planning below to build your solution.

- Representation: single if-then rule by using bit strings (binary encoding).
  - Outlook  $\langle \text{Sunny}, \text{Overcast}, \text{Rain} \rangle$
  - Temperature  $\langle \text{Hot}, \text{Mild}, \text{Cool} \rangle$
  - Examples:
    - Outlook = *Overcast*  $\rightarrow$  010
    - Outlook = *Overcast*  $\vee$  Rain  $\rightarrow$  011
    - Outlook = *Sunny*  $\vee$  *Overcast*  $\vee$  Rain  $\rightarrow$  111
    - Outlook = (*Overcast*  $\vee$  *Rain*)  $\wedge$  (Temperature = *Hot*)  $\rightarrow$  011100
    - Outlook = *Sunny*  $\wedge$  Temperature = *Hot* then PlayTennis = *yes*  $\rightarrow$  1001001
- Initial population (Chromosomes) : ( $C_1=1001001$ ,  $C_2=0100101$ ,  $C_3=1011000$ ,  $C_4=1101100$ ). Population size should remain the same (4 individuals) over time.
- Fitness function: accuracy
- Penalty criterion: no penalty.
- Selection method: The best two chromosomes are carried over to the next generation. The other two are selected for crossover by using the roulette wheel (simulation).
- Crossover strategy: single-point crossover with mask 1110000 in the 1<sup>st</sup> generation, two-point crossover with mask 0001100 in the 2<sup>nd</sup> generation. Use the following chromosomes to perform crossover (simulating the process of spinning the roulette wheel) according to the relative fitness (sectors of a roulette wheel), generating two offspring for each crossover:
  - (1<sup>st</sup> and 3<sup>rd</sup>) in the 1<sup>st</sup> generation
  - (1<sup>st</sup> and 2<sup>nd</sup>) in the 2<sup>nd</sup> generation
- Mutation: on the 6<sup>th</sup> bit of the chromosome(s) 1011000 selected/produced during the 2<sup>nd</sup> generation.
- Termination criteria: accuracy = 1.0. **Return the corresponding chromosome(s) – this will be your model.**

Solution format:

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

1<sup>st</sup> generation ( $C_?$ ,  $C_?$ ,  $C_?$ ,  $C_?$ ):

Pr( $C_?$ ) = ? (4<sup>th</sup>)

Pr( $C_?$ ) = ? (3<sup>rd</sup>)

Pr( $C_?$ ) = ? (2<sup>nd</sup>)

Pr( $C_?$ ) = ? (1<sup>st</sup>)

$C_?$  = ???????  $\rightarrow$   $C_?$  = ???????

$C_?$  = ???????  $\rightarrow$   $C_?$  = ???????

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

2<sup>nd</sup> generation (?, ?, ?, ?)

Pr( $C_?$ ) = ? (4<sup>th</sup>)

Pr( $C_?$ ) = ? (3<sup>rd</sup>)

Pr( $C_?$ ) = ? (2<sup>nd</sup>)

Pr( $C_?$ ) = ? (1<sup>st</sup>)

$C_?$  = ???????  $\rightarrow$   $C_?$  = ???????

$C_?$  = ???????  $\rightarrow$   $C_?$  = ???????

Applying mutation on 1011000

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Fitness( $C_?$ ) = ?

Final answer:

**Important Note:** Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may resubmit multiple times until the deadline (the last submission will be considered).

**NO LATE ASSIGNMENTS WILL BE ACCEPTED. ALWAYS SUBMIT WHATEVER YOU HAVE COMPLETED FOR PARTIAL CREDIT BEFORE THE DEADLINE!**