CSC 735 Machine Learning and Data Mining

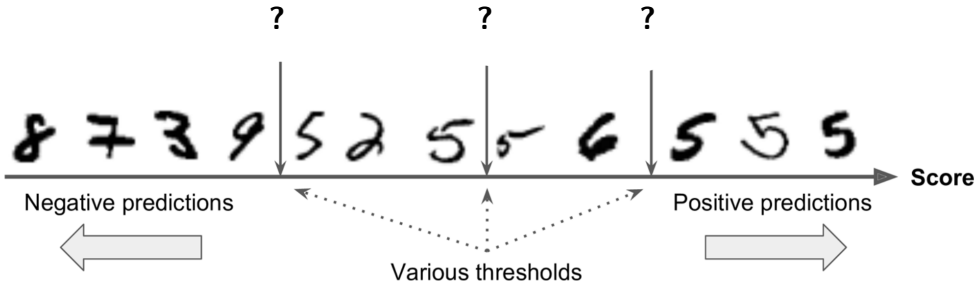
Midterm Exam

Name: _____

1. (1 point) **"We may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus (data) development."** What does this comment mean?

2. (1 point) **"You must resist the temptation to tweak the hyperparameters to make the numbers look good on the test set."** Why?

3. (2 points) What are the "precision" and "recall" at the thresholds in the figure below?



4. (2 points) What are OvR and OvO?

5. (2 points) Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization hyperparameter α or reduce it?

6. (1 point) What is the trick to use SVMs for regression instead of classification?

7. (2 points) Suppose the data set is as follows:

   ```
   social_platforms = ['X', 'Instagram', 'TikTok', 'X', 'Facebook', 'Instagram', 'LinkedIn']
   ```

   Convert these categories from text to numbers using:

   (a) Ordinal encoding

   (b) One-hot encoding

8. (1 point) What will happen if a learning rate is set too low or too high?

9. (1 point) What is sigmoid and its role in Logistic Regression?

10. (1 point) What is the role of the C in SVM?

11. (2 points) Do scaling the features generally improves:
    a) Linear Regressions Using Gradient Descent
    b) Support Vector Machines
    c) Decision Trees

12. (1 point) Can we use SVM for multiclass classification? The answer is both yes and no.
    Why no and how yes?

13. (1 point) Say we have four input numbers -1, 0, 3, and 4 of the Softmax function, what is the probability of the last number?

14. (1 point) What is the story behind the Figure 1?

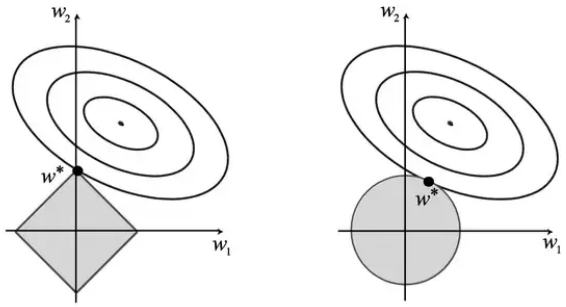


Figure 1

15. (6 points) **Confusion Matrix**

Suppose that 10 % of all bicycle racers use steroids, that a bicyclist who uses steroids tests positive for steroids 95 % of the time, and that a bicyclist who does not use steroids tests positive for steroids 10 % of the time.

(a) Draw the confusion matrix for this steroids test (with TP, TN, FP, FN tagged)

(b) Calculate the precision, recall, and F_1 score

(c) What is the probability that a randomly selected bicyclist who tests positive for steroids actually uses steroids?

16. (4 points) **Lagrange Multiplier**

Use the method of Lagrange Multipliers to find the maximum value of $x^2 - 2y^2 + z^2$ subject to the constraint $x^2 + y^2 + z^2 = 1$

17. (5 points) **Support Vector Machine**

Figure 2 shows the instances for two different classes:

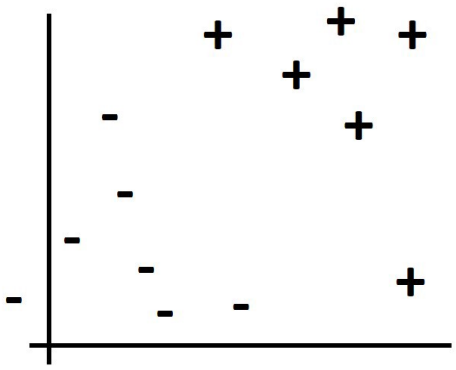


Figure 2

(a) Draw the best decision boundary by a solid line.

(b) Draw the margins by the dashed lines.

(c) Circle the support vectors.

(d) Annotate which lines are $\boldsymbol{w}^\intercal \boldsymbol{x} + b = +1$ and $\boldsymbol{w}^\intercal \boldsymbol{x} + b = -1$.

(e) What is the hard margin linear SVM classifier objective?

18. (1 point) What is the difference between hard and soft voting classifiers?

19. (1 point) Tell me something you know about Random Forests.

20. (1 point) What is the story behind $\lim_{m \to \infty} (1 - \frac{1}{m})^m = \frac{1}{e} = 0.368\ldots$?

21. (9 points) **Linear Regression**

After we trained a Linear Regression model, we got a parameter vector:

$$\hat{\boldsymbol{\theta}}^{\mathsf{T}} = \begin{bmatrix} 4 & 3 & 5 & 2 \end{bmatrix}$$

Now we have a test set:

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{y}$
7	8	9	78
4	5	6	53
1	2	3	42

(Please include $\boldsymbol{x}_0$ for the following questions)

(a) What is test's $\boldsymbol{x}^{(3)}$?

(b) What is test's $\boldsymbol{y}$?

(c) What is $\hat{\boldsymbol{y}}$?

(d) What is the Mean Absolute Error of this test set?

(e) What is the new n (#features), if we transform the data to fit a Second-degree Polynomial Regression?

(f) As with Linear Regression, we can perform Ridge Regression by computing a closed-form equation:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \alpha\boldsymbol{A})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$$

where $\boldsymbol{A}$ is the $(n+1) \times (n+1)$ identity matrix, except with a 0 in the top-left cell.

If we regard this test set as a training set, how do we perform Ridge Regression using this closed-form equation? (Just plug in the numbers, and let $\alpha = 10$)

22. (4 points) **Decision Tree**

Calculate the Information Gain using **Cholesterol**:

Patient ID	**Age**	**Sex**	**BP**	**Cholesterol**	**Drug**
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

Table 1: Patient Data for Drug Classification