CSC 735 Machine Learning and Data Mining

Review I

1. What is the difference between the traditional approach versus the machine learning approach?

2. What is unsupervised learning? List some unsupervised learning algorithms.

3. What is a pipeline in machine learning?

4. What are the main challenges of machine learning to "bad data"?

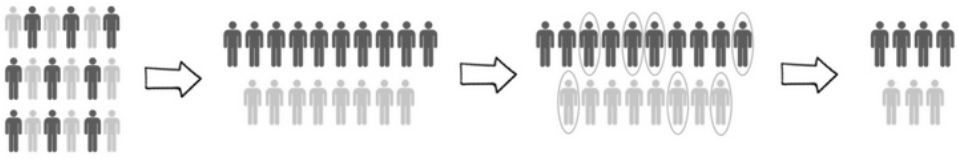
5. How to do a 5-fold cross-validation?

6. What is online learning?

7. When a model is trained, we evaluate it on the training set finding no errors at all. What happens?

8. Assume that you have some promising models. You now need to fine-tune them. How to fine-tune these models?

9. What is the story behind the figure below?



10. Suppose the data set is as follows:

```
+--------+------+----+-------+-----+-------+------+
| Toyota | Ford | VW | Tesla | BMW | Tesla | Ford |
+--------+------+----+-------+-----+-------+------+
```

    Convert these categories from text to numbers using:

    (a) Ordinal encoding

    (b) One-hot encoding

11. What is correlation coefficient between the column `x` and `y`? `x` and `z`? `x` and `w`? (roughly estimate)

x	y	z	w
0.9	1.8	-1.6	9.7
2.4	4.9	-4.7	6.4
2.7	5.4	-5.2	5.0
3.2	6.5	-6.2	4.7
4.7	9.7	-9.3	3.2
5.0	10.0	-9.8	2.7
6.4	12.8	-12.6	2.4
9.7	19.6	-19.2	0.9

12. What is the difference between multiclass classification and multilabel classification?

13. An ROC curve plots TPR versus FPR at different classification thresholds. Draw
    a) a normal classifier curve,
    b) a better classifier curve,
    c) a random classifier curve,
    together within one graph with TPR and FPR annotated.

14. What will happen if a learning rate is set too low or too high?

15. What is sigmoid and its role in Logistic Regression?

16. What is the role of the C in SVM?

17. Suppose the features in your training set have very different scales. Which algorithms might suffer from this:
    a) Linear Regressions Using Normal Equation
    b) Linear Regressions Using Gradient Descent
    c) Ridge Regressions
    d) Support Vector Machines

18. Can we use SVM for multiclass classification? The answer is both yes and no.
    Why no and how yes?

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

20. What is the story behind the Figure 1?

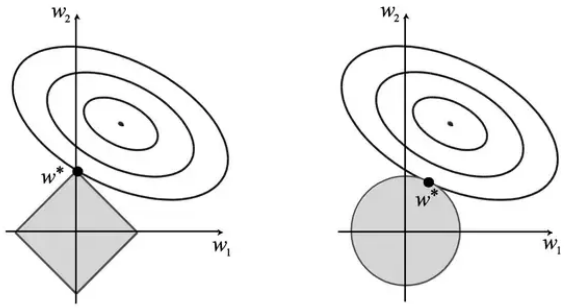


Figure 1

21. What is the story behind the Figure 2?

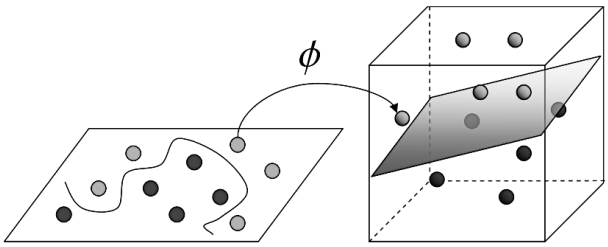


Figure 2

22. **Confusion Matrix**

    An electronics company is planning to introduce a new phone. The company commissions a marketing report for each new product that predicts either the success or the failure of the product. Of new products introduced by the company, 60 % have been successes. Furthermore, 70 % of their successful products were predicted to be successes, while 40 % of failed products were predicted to be successes.

(a) Let the positive mean success. What do FP and FN mean?

(b) Calculate F_1 score

(c) Find the probability that this new phone will actually be a success if its failure has been predicted.

23. Linear Regression

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

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

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:

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

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 $\alpha = 10$)

24. Lagrange Multiplier

Use the method of Lagrange Multipliers to find the maximum value of xy subject to the constraint $x + y = 4$.

25. Support Vector Machine

Figure 3 shows the instances for two different classes:

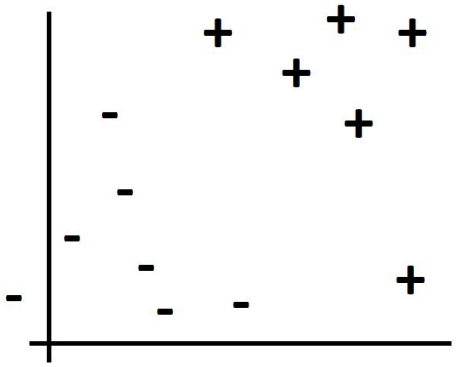


Figure 3

(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?

26. Kernel Trick

Suppose you want to apply a third-degree polynomial transformation to a one-dimensional training set, and then train a linear SVM classifier on the transformed training set. The mapping function:

$$\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\phi}(\,[x]\,) = \begin{bmatrix} 1 \\ \sqrt{3}\,x \\ \sqrt{3}\,x^2 \\ x^3 \end{bmatrix}$$

Can we perform a kernel trick with the function:

$$\boldsymbol{\kappa}(\boldsymbol{a}, \boldsymbol{b}) = \left(\boldsymbol{a}^\intercal \boldsymbol{b} + 1\right)^3$$

? Why or why not?

27. What is the approximate depth of a Decision Tree trained (without restrictions) on a training set with one billion (10^9) instances?

28. If a Decision Tree is overfitting the training set, is it a good idea to try
    (a) Increasing `max_depth`?
    (b) Decreasing `min_samples_leaf`?
    (c) Scaling the input features?

29. Why does the voting classifier often achieve a higher accuracy than the best classifier in the ensemble?

30. What is the difference between hard and soft voting classifiers?

31. Tell me something you know about Random Forests.

32. What is the story behind the Figure 4?

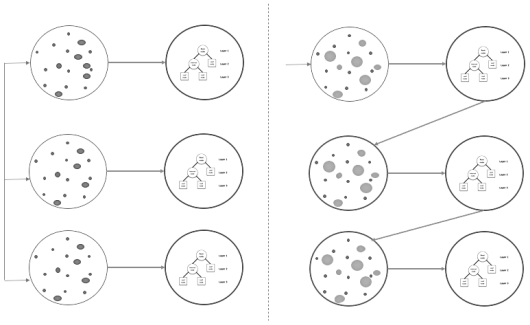


Figure 4

33. What is the idea of Stacking?

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

35. **Decision Tree**

Imagine we want to predict whether Mike will go grocery shopping on any given day. We can look at previous factors that led Mike to go to the store. See Figure 5. Here we can see the amount of grocery supplies Mike had, the weather, and whether Mike worked each day.

Build a decision tree on this data set. You only need to find the first branch.

(You can choose either Gini Index or Information Gain.)

	Supplies	**Weather**	**Worked?**	**Shopped?**
D1	Low	Sunny	Yes	Yes
D2	High	Sunny	Yes	No
D3	Med	Cloudy	Yes	No
D4	Low	Raining	Yes	No
D5	Low	Cloudy	No	Yes
D6	High	Sunny	No	No
D7	High	Raining	No	No
D8	Med	Cloudy	Yes	No
D9	Low	Raining	Yes	No
D10	Low	Raining	No	Yes
D11	Med	Sunny	No	Yes
D12	High	Sunny	Yes	No

Figure 5