University Jean Monnet MLDM Master Program

Exam of Introduction to Machine Learning

December 2019

Guidelines:
- In the *Answer Sheet*, circle the letter corresponding to the correct answer (**only one is correct**).
- You can leave questions unanswered. Each correct answer **adds one point.**
- Each incorrect answer **subtracts half a point.**

Questions

1. Among the following algorithms, which one is not a classification method?
   a. Logistic regression.
   b. k-nearest neighbors.
   c. Polynomial regression.

2. A training set of labeled examples is a sample drawn over a space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the set of classes. What does $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ mean?
   a. $\mathcal{Z}$ is a joint space over $\mathcal{X}$ and $\mathcal{Y}$.
   b. $\mathcal{Z}$ corresponds to the product of the features by the class.
   c. $\mathcal{Z}$ is a finite set of examples jointly drawn from $\mathcal{X}$ and $\mathcal{Y}$.

3. As the number of features grows, the amount of data we need to generalize accurately:
   a. Grows exponentially.
   b. Grows linearly.
   c. Grows quadratically.

4. Which of the following statements about the regularization parameter λ is correct?
   a. Using too large a value of λ can cause your hypothesis to underfit the data.
   b. Using too large a value of λ can cause your hypothesis to overfit the data.
   c. Using a very large value of lambda cannot hurt the performance of your hypothesis.

5. $\mathbb{E}_{z \sim \mathcal{D}_Z} \ell(h, z)$ stands for:
   a. The empirical risk.
   b. The bayesian error.
   c. The true risk.

6. An indicator function $[a]$ is defined as follows:

   a. $[a] = 1$ if a is true and 0 otherwise.
   b. $[a] = 0$ if a is true and 0 otherwise.
   c. $[a] > 0$ if a is false and 0 otherwise.

7. The hinge loss suffered by h at x when $h(x) = 0.6$ and $y = +1$ is

   a. 0
   b. 0.4
   c. 1.6

$\max(0, 1 - y h(x))$
$1 - 1(0.6) = 0.4$

8. The use of the ℓ_1-norm in a regularization term boils down to:

   a. selecting features only.
   b. reducing the risk of overfitting only.
   c. both selecting features and reducing the risk of overfitting.

lasso - L1
L2 - Ridge

9. Regarding bias and variance, which of the following statements is right?

   a. Models which overfit have a high bias.
   b. Models which overfit have a low bias.
   c. Models which underfit have a high variance.

10. Which ℓ_p-norm is differentiable everywhere?

    a. $\ell_{0.5}$-norm
    b. ℓ_1-norm
    c. ℓ_2-norm

11. Let's suppose that the Bayesian error $\epsilon_B = 3\%$. A classifier h has a training error=3% and a validation error=10%. What would you suggest?

    a. Increase the complexity of the classifier.
    b. Increase the number of training examples.
    c. Increase both.

over fitting

12. The error obtained using K-cross-validation is an estimate of the generalization error. Is this estimate unbiased?

    a. Yes
    b. No
    c. It depends on K

13. K-fold cross-validation is

    a. linear in K
    b. quadratic in K
    c. cubic in K

14. k-NN algorithm does more computation at test time rather than at training time.

    a. True
    b. False
    c. It depends on the number of samples

15. Which of the following statements are true about k-NN algorithm?

    1. k-NN performs much better if all of the data have the same scale
    2. k-NN struggles when the number of features is very large.
    3. k-NN makes no assumptions about the underlying distribution of the problem being solved

    a. 1 and 2
    b. 1 and 3
    c. All of the above

16. When you find noise in data which of the following option would you consider in k-NN?

    a. Increase the value of k
    b. Decrease the value of k
    c. Noise can not be dependent on value of k

17. A 1-NN classifier has higher variance than a 3-NN classifier.

    a. Yes
    b. No
    c. It depends on the training data

18. Which of the following statements is true?

    1. k-NN immediately adapts as we collect new training data.
    2. The computational complexity for classifying new samples grows linearly with the number of samples in the training dataset in the worst-case scenario.

    a. 1
    b. 2
    c. both 1 and 2

19. Which of the following statements is true for k-NN classifiers?

a. The classification accuracy is always better with larger values of k

b. The decision boundary is linear

c. None of these

20. Which of the following is true about Naive Bayes?

a. It assumes that all the features are equally important

b. It assumes that all the features are independent

c. Both a and b

21. Which of the following machine learning algorithms is based on the idea of bagging?

a. Decision Tree

b. Random Forest

c. Adaboost

22. Which of the following algorithms is not an example of an ensemble method?

a. Random Forest

b. Gradient Boosting

c. Decision Tree

23. Ensembles will yield bad results when there is significant diversity among the weak hypotheses.

a. True

b. False

c. It depends on the problem

24. Which of the following is true about weak learners used in an ensemble model?

a. They have low variance and they don't usually overfit

b. They have high variance and small bias

c. They have high variance and they don't usually overfit

25. Suppose, you are working on a binary classification problem, and there are 3 weak models each with 70% accuracy. If you want to ensemble these models using a majority voting method. What will be the maximum accuracy you can get?

a. 100%

b. 90.1%

c. 70%

$P(y_i = H_i(x_i)) = P(y_i = h_1(x_i), y_i = h_2(x_i), y_i = h_3(x_i)) + 3 \cdot P(y_i = h_1(x_i), y_i = h_2(x_i), y_i \neq h_3(x_i))$

$P(yH > 0) = P(y > 0, H > 0) + P(y < 0, H < 0)$

$0.7 \times 0.7 \times 0.7$

$+ 3 \times 0.7 \times 0.7 \times 0.3$

$= 0.343$

$+ 0.441$

$= 0.784$

26. Which of the following is true about bagging?

    1. Bagging can be parallelized ← ind. iteration.
    2. The aim of bagging is to reduce bias not variance
    3. Bagging helps in reducing overfitting

    a. 1 and 2

    b. 2 and 3

    c. 1 and 3

27. In boosting, individual weak learners can be parallelized.

    a. True

    b. False

    c. It depends on the number of iterations

28. Let S be a training set of m examples lying in a p-dimensional space. Let's assume that the number of iterations of Adaboost is T.

    a. Adaboost learns a linear separator in $\mathbb{R}^T$.

    b. Adaboost learns a linear separator in $\mathbb{R}^p$.

    c. Adaboost learns a linear separator in $\mathbb{R}^m$.

29. Adaboost usually works well with:

    a. A decision stump algorithm.

    b. A k-nearest-neighbor algorithm.

    c. A decision tree without pruning.

30. In Adaboost, the error of each hypothesis is calculated by the ratio of misclassified examples to the total number of examples.

    a. Yes, always

    b. No, never.

    c. Yes, but only for the first iteration.