UNIVERSITY OF PENNSYLVANIA
CIS 5200: Machine Learning
Midterm 2022

Exam policy: You are allowed one two-sided cheat sheet. No calculators.

Time: 90 minutes.

Please write your name and Penn ID on the bubble sheet.

If you think a question is ambiguous, mark what you think is the best answer. As always, we will consider written regrade requests if your interpretation of a question differed from what we intended. *We will only grade the answer forms.*

For the "TRUE or FALSE" questions, note that "TRUE" is (a) and "FALSE" is (b). For the multiple choice questions, select exactly one answer.

Questions follow the course convention that n or N represents the number of observations and p the number of features in the model.

There are **46** questions totaling **58** points.

1. [1 points] *True or False?* Ordinary least squares (with L2 loss) is more commonly used than regression with L1 loss, because the L2 loss, unlike the L1 loss, is convex.

2. [2 points] Which of the following is true of ordinary least squares regression?

   (a) It is prone to underfitting.
   (b) It always has a closed-form Solution.
   (c) The weights are usually calculated using gradient descent.
   (d) It has a closed-form solution, except when the columns of $\mathbf{X}$ are linearly dependent
   (e) none of the above

3. [2 points] In a regression task, suppose one feature is in inches. If we convert that feature to centimeters, which model would in general have the same predictions as before?

   (a) OLS regression
   (b) Ridge regression
   (c) K-NN
   (d) two of the above
   (e) none of the above

4. [2 points] If you had to choose whether to use the MLE or MAP estimate for a machine learning model, which is generally better to use and why?

   (a) MAP, because it generalizes MLE.
   (b) MLE, because it is unbiased.
   (c) MLE, because it avoids overfitting better than MAP.
   (d) MAP, because it is unbiased.

5. [1 points] *True or False?* Using ridge regression will always result in all of the weights being smaller than using ordinary least squares regression on the same dataset.

6. [1 points] *True or False?* When using the MAP estimate, collecting more data increases the effect of the prior because the prior can affect more data than before.

7. [1 points] *True or False?* Suppose θ is the parameter we are interested in, D is the dataset that we have observed. Then θ_{MLE} maximizes $P(D|\theta)$, and θ_{MAP} maximizes $P(\theta|D)$.

8. [1 points] *True or False?* If we double all features (i.e. multiply all x's by 2), Decision Trees will give the same prediction.

9. [1 points] *True or False?* If we double all features (i.e. multiply all x's by 2), K-NN will give the same prediction.

10. [1 points] *True or False?* If we double all features (i.e. multiply all x's by 2), Ridge regression will give the same prediction.

11. [2 points] Which of the following regression methods perform feature selection?

    (a) L0

    (b) L1

    (c) L2

    (d) L-inf

    (e) Two of the above

12. [1 points] *True or False?* We regularize a Bernoulli model for estimating the probability of getting a "heads" (vs. a "tails"), by 'pretending' we have already seen a certain number of heads and tails. The effect of this prior vanishes in the limit of infinite observations.

13. [1 points] Which of the following is TRUE about the bias-variance decomposition of test error?

    (a) Expected test error is equal to $bias + var^2 + noise$

    (b) Noise can be reduced by using regularization

    (c) Low variance corresponds to high complexity

    (d) None of the above

14. [1 points] In machine learning, we tend to favor estimators that are **pick the best–or least bad–answer**:

    (a) Unbiased, because we want predictions as close to their true values as possible

(b) Biased, because we don't want to treat the training data as definitive

(c) Unbiased, because at high training set size n we will converge to the optimal predictor

(d) Biased, because we want to push model parameters to as close to zero as we can

15. [1 points] What best defines the relationship between a model's fit and its bias and variance

(a) A model with low bias and high variance is underfitting

(b) A model with high bias and high variance is underfitting

(c) A model with low bias and low variance is underfitting

(d) A model with high bias and low variance is underfitting

16. [2 points] If Z is ""High Entropy"", this suggests that

(a) Z is from a relatively uniform distribution

(b) Z is from a highly varied distribution

(c) The value of Z is certain and known

(d) Z is conditional on other variables

17. [1 points] *True or False?* A Gaussian distribution with $\sigma = 1$ has higher entropy than one with $\sigma = 3$.

18. [1 points] *True or False?* Standard SVMs, as covered in class (linear, so no Gaussian kernel), have no hyperparameters that need to be set by cross-validation.

19. [2 points] If we assume that the data is being generated from a source with no noise $y = w^T x$, choosing a classifier $f^*(x) = {w^*}^T x$ using OLS will make which of the following quantities zero?

(a) Only Bias

(b) Only Variance

(c) Both Bias and Variance

(d) None of the above

20. [1 points] *True or False?* Increasing the kernel width in a RBF model will in general increase the training error.

21. [1 points] *True or False?* Increasing the kernel width in a RBF model will in general increase the testing error.

22. [1 points] *True or False?* Increasing k in a K-NN model will in general increase the error on the 'training set'.

23. [1 points] *True or False?* If OLS gives all non-zero values for the weights, using Ridge Regression will also result in all non-zero weight values.

24. [1 points] *True or False?* Leave-one-out cross-validation is often preferred over k-fold cross-validation for smaller datasets.

25. [1 points] *True or False?* Link functions transform the feature space nonlinearly before fitting a linear model.

26. [2 points] A discrete probability distribution has (x, probability(x)) tuples $\left(1, \frac{1}{4}\right), \left(2, \frac{1}{2}\right), \left(4, \frac{1}{4}\right)$. What is the expected value of x?

    (a) 1

    (b) 1.5

    (c) 2.25

    (d) 7

    (e) none of the above

27. [2 points] A discrete probability distribution has (x, probability(x)) tuples $\left(1, \frac{1}{4}\right), \left(2, \frac{1}{2}\right), \left(4, \frac{1}{4}\right)$. What is the entropy of this distribution?

    (a) 1 bit

    (b) 1.5 bits

    (c) 2 bits

    (d) 3 bits

    (e) none of the above

28. [1 points] *True or False?* The L2 penalty in ridge regression comes from the assumption of Gaussian noise in the linear regression model $y = w^T x + \epsilon$.

29. [1 points] *True or False?* The MAP with a uniform prior (the same probability everywhere) will be different from the MLE because the prior's shrinkage is applied everywhere.

30. [2 points] Suppose you are given training and testing datasets for a regression problem, and you train a ridge regression model on the training dataset. If you observe low training error but high testing error, what is likely to be the best thing to do to improve the testing error?

    (a) Increase the number of features used

    (b) Increase the regularization parameter λ

    (c) Increase the number of training iterations

    (d) None of the above

31. [1 points] *True or False?* AdaGrad uses a faster learning rate for features that have been changed more in the past than for ones that have changed less.

32. [2 points] You train an SVM on a dataset with data of x_i and labels $y_i = \pm 1$ This produces a vector of weights w and a bias b. For the first 6 training examples, the values of $w^T x_i + b$ are listed below. How many support vectors are there?

i	y_i	$w^T x_i + b$
1	-1	-1.5
2	-1	-1
3	-1	-0.5
4	$+1$	1
5	$+1$	3
6	-1	4

    (a) 2

    (b) 3

    (c) 4

    (d) 5

    (e) none of the above

33. [1 points] *True or False?* SVM solutions can be expressed purely in terms of the vectors on the ”margin”.

34. [1 points] *True or False?* The hinge loss used in SVMs generally gives less weight than logistic regression to points that are **misclassified** with a high probability or score.

35. [1 points] *True or False?* The hinge loss used in SVMs generally gives less weight than logistic regression to points that are **correctly classified** with a high probability or score.

36. [1 points] Your random forest is overfitting. What should you do to the number of trees in the model?

    (a) Increase

    (b) Decrease

    (c) the number of trees won't effect overfitting

37. [1 points] *True or False?* $k(x, x') = exp(\frac{||x-x'||_2^2}{2\sigma^2})$ is a valid kernel function.

38. [1 points] *True or False?* An ensemble will usually have lower variance than a single model that is part of the ensemble.

39. [1 points] *True or False?* Boosting methods tend to give higher test accuracy when they use more accurate 'weak models'.

40. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) bagging.

41. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) boosting.

42. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) randomly selecting subsets of features.

43. [1 points] *True or False?* Random forests (as we used them in class) try to make their weak learning less accurate by (among other methods) adding noise to the labels.

44. [2 points] Which of the following statements about KL-Divergence is true? Note that P refers to the true distribution and Q refers to the alternative distribution.

    (a) It measures how well (or more precisely, how badly) Q approximates P

(b) It measures the difference between two distributions, and thus can be seen as a distance metric

(c) It can be written as cross-entropy of P, Q minus the entropy of Q. i.e. $D_{KL}(P||Q) = H(P, Q) - H(Q)$

(d) Two of the above

(e) None of the above

45. [1 points] *True or False?* For a model with more than 1 feature, streamwise regression always tests fewer models than stepwise regression.

46. [2 points] Consider classification using k-NN. Suppose we try different norms and get different decision boundaries. Which norms best match figures A, B, and C, respectively?

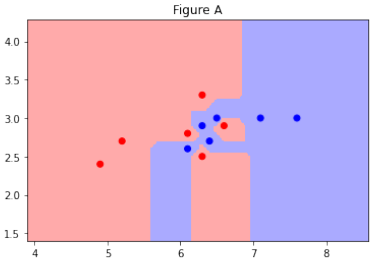
(a) $L1, L2, L_{\text{inf}}$

(b) $L2, L1, L_{\text{inf}}$

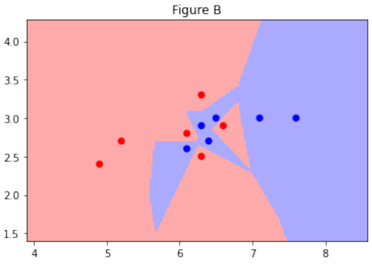
(c) $L1, L_{\text{inf}}, L2$

(d) $L_{\text{inf}}, L2, L1$

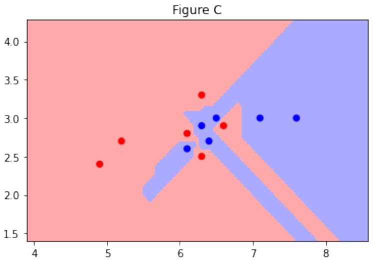
(e) none of the above



(a)



(b)



(c)

Total Points: 58