UNIVERSITY of PENNSYLVANIA

CIS 520: Machine Learning
Sample Midterm, based on clicker questions

Exam policy: This exam allows one one-page, two-sided cheat sheet; No other materials.

Time: 80 minutes. Be sure to write your name and Penn student ID (the 8 bigger digits on your ID card) on the scantron form and fill in the associated bubbles *in pencil*.

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 scantron 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.

these are the clicker questions from class, which are typical of the midterm you will get

1. [0 points] This is version **A** of the exam. Please fill in the “bubble” for that letter.

   ‘

2. [2 points] If an event is certain, the entropy is

   (a) 0
   (b) between 0 and 1/2
   (c) 1/2
   (d) between 1/2 and 1
   (e) 1

3. [2 points] If two events are equally likely, the entropy is

   (a) 0
   (b) between 0 and 1/2
   (c) 1/2
   (d) between 1/2 and 1
   (e) 1

4. [2 points] Linear regression is

   (a) Parametric
   (b) Non-parametric

5. [2 points] K-NN is

   (a) Parametric
   (b) Non-parametric

6. [2 points] When, if ever does $E[X+Y] = E[X] + E[Y]$

   (a) All the time?
   (b) Only when X and Y are independent?
   (c) It can fail even if X and Y are independent?

7. [2 points] 1-nearest neighbors is a consistent estimation algorithm.

   (a) True
   (b) False

8. [2 points] Which is usually unbiased
   (a) MLE
   (b) MAP

9. [2 points] The conjugate prior to a Bernoulli is
   (a) Bernoulli
   (b) Gaussian
   (c) Beta
   (d) none of the above

10. [2 points] The conjugate prior to a Gaussian is
    (a) Bernoulli
    (b) Gaussian
    (c) Beta
    (d) none of the above

11. [2 points] KL Divergence is a metric (distance)
    (a) True
    (b) False

12. [2 points] KL divergence can be used in k-nn instead of a distance
    (a) True
    (b) False

13. [2 points] If you are dividing up a data set that someone gives you into a training and test set
    (a) It is better to randomly select the observations into the two subsets
    (b) It is better to divide the data so that the first half is the training set and the second half is the testing set
    (c) It is unlikely to matter which one you do
    (d) It depends upon what sort of data and what you're doing with it

14. [2 points] Ordinary least squares (OLS) and logistic regressiond are MLE estimators that minimize

(a) bias

(b) variance

(c) bias + variance

15. [2 points] Ridge regression is an MAP estimator that minimizes

    (a) bias

    (b) variance

    (c) bias + variance

16. [2 points] Minimizing the first term in $|y - w.x|_2^2 + \lambda |w|_2^2$, reduces

    (a) bias

    (b) variance

    (c) neither

17. [2 points] Minimizing the second term in $|y - w.x|_2^2 + \lambda |w|_2^2$, which can be viewed as the amount that the test error is expected to be bigger than the training error reduces

    (a) bias

    (b) variance

    (c) neither

18. [2 points] Which norm most heavily shrinks large weights?

    (a) L_0

    (b) L_1

    (c) L_2

19. [2 points] Which norm, when used as a penalty for linear regression, most strongly encourages weights to be set to zero?

    (a) L_0

    (b) L_1

    (c) L_2

20. [2 points] Which norm, when used as a penalty for linear regression, is scale invariant?

    (a) L_0

(b) L_1

(c) L_2

21. [2 points] Which norm, when used as a penalty for linear regression, is called "LASSO"

(a) L_0

(b) L_1

(c) L_2

22. [2 points] Which norm, when used as a penalty for linear regression, does **not** lead to convex optimization problems?

(a) L_0

(b) L_1

(c) L_2

23. [2 points] Ridge regression (Tikhonov regularization) minimizes $Err + \lambda|w|_2^2$ Is Err here

(a) $\sum_i (y_i - \hat{y}_i)^2$

(b) $(1/n)\sum_i (y_i - \hat{y}_i)^2$

(c) $sqrt((1/n)\sum_i (y_i - \hat{y}_i)^2)$

(d) $sqrt(\sum_i (y_i - \hat{y}_i)^2)$

24. [2 points] Elastic net regularization minimizes $Err + \lambda_1|w|_1 + \lambda_2|w|_2^2$

(a) True

(b) False

25. [2 points] Will $Err + \lambda_1|w|_1 + \lambda_2|w|_2^2$ sometimes zero out some features?

(a) yes

(b) no

26. [2 points] AIC, BIC and RIC Minimize $Err/2\sigma + \lambda|w|_0$ Is this error

(a) $\sum_i (y_i - \hat{y}_i)^2$

(b) $(1/n)\sum_i (y_i - \hat{y}_i)^2$

(c) $sqrt((1/n)\sum_i (y_i - \hat{y}_i)^2)$

(d) $sqrt(\sum_i (y_i - \hat{y}_i)^2)$

27. [2 points] Which penalty should you use if you expect 10 out of 100,000 features , n = 100

    (a) AIC

    (b) BIC

    (c) RIC

28. [2 points] Which penalty should you use if you expect 200 out of 1,000 features, n = 1,000,000

    (a) AIC

    (b) BIC

    (c) RIC

29. [2 points] Which penalty should you use if you expect 500 out of 1,000 features, n = 1,000

    (a) AIC

    (b) BIC

    (c) RIC

30. [2 points] You think maybe 10 out of 100,000 features will be significant. Use

    (a) L_2 with CV

    (b) L_1 with CV

    (c) L_0 with AIC

    (d) L_0 with BIC

    (e) L_0 with RIC

31. [2 points] You think maybe 500 out of 1,000 features will be significant. Do not use

    (a) L_2 with CV

    (b) L_1 with CV

    (c) L_0 with AIC

    (d) L_0 with BIC

(e) L_0 with RIC

32. [2 points] Which estimator is consistent?

(a) AIC

(b) BIC

(c) RIC

(d) none of them

(e) all of the above

33. [2 points] Does LOOCV systematically _____ test error

(a) Overestimate

(b) Underestimate

(c) sometimes overestimate and somtimes underestimate