UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Midterm, 2016

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 answer 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.*

There are **60** questions.

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

2. [2 points] Suppose we have a regularized linear regression model: $argmin_w \|Y - Xw\|_2^2 + \lambda \|w\|_1$. What is the effect of increasing λ on bias and variance?

   (a) Increases bias, increases variance

   (b) Increases bias, decreases variance

   (c) Decreases bias, increases variance

   (d) Decreases bias, decreases variance

   (e) Not enough information to tell

3. [2 points] Suppose we have a regularized linear regression model: $argmin_w \|Y - Xw\|_2^2 + \|w\|_p^p$. What is the effect of increasing p on bias and variance ($p \geq 1$) if the weights are all larger than 1?

   (a) Increases bias, increases variance

   (b) Increases bias, decreases variance

   (c) Decreases bias, increases variance

   (d) Decreases bias, decreases variance

   (e) Not enough information to tell

4. [2 points] Compared to the variance of the MLE estimate $\hat{w}_{MLE}$, do you expect that the variance of the MAP estimate $\hat{w}_{MAP}$ is:

   (a) higher

   (b) same

   (c) lower

   (d) it could be any of the above

5. [1 points] *True or False?* Given a distribution $p(X, y)$, it is always (in theory) possible to compute $p(y|X)$.

6. [1 points] *True or False?* It is generally more important to use consistent estimators when one has smaller numbers of training examples.

7. [1 points] *True or False?* It is generally more important to used unbiased estimators when one has smaller numbers of training examples.

8. [1 points] *True or False?* LASSO is a parametric method.

9. [2 points] Ridge regression uses what penalty on the regression weights?

   (a) L_0
   (b) L_1
   (c) L_2
   (d) L_2^2
   (e) none of the above

10. [1 points] *True or False?* Linear regression gives the MLE for data from a model $y_i \sim N(\sum_j w_j x_{ij}, \sigma^2)$.

11. [2 points] Suppose you are given the following training inputs $X = [-3, 5, 4]$ and $Y = [-10, 20, 20]$. Which of the following is **closest** to the MLE estimate $\hat{w}_{MLE}$?

    (a) 2.0
    (b) 4.1
    (c) 4.2
    (d) 8.2
    (e) 10.0

12. [2 points] Using the same data as above $X = [-3, 5, 4]$ and $Y = [-10, 20, 20]$, assuming a ridge penalty $\lambda = 50$, what ratio versus the MLE estimate $\hat{w}_{MLE}$ do you think the ridge regression L_2 estimate estimate $\hat{w}_{ridge}$ will be?

    (a) 2
    (b) 1
    (c) 0.666
    (d) 0.5
    (e) 0.25

13. [1 points] *True or False?* A probability density function (PDF) cannot be less than 0 or bigger than 1.

14. [1 points] *True or False?* A cumulative distribution function (CDF) cannot be less than 0 or bigger than 1.

15. [2 points] A and B are two events. If $P(A, B)$ decreases while $P(A)$ increases, what must be true:

    (a) $P(A|B)$ decreases

    (b) $P(B|A)$ decreases

    (c) $P(B)$ decreases

    (d) All of above

16. [2 points] After applying a regularization penalty in linear regression, you find that some of the coefficients of w are zeroed out. Which of the following penalties might have been used?

    (a) $L0$ norm

    (b) $L1$ norm

    (c) $L2$ norm

    (d) either (A) or (B)

    (e) any of the above

17. [1 points] *True or False?* 1-nearest neighbor is a consistent estimator.

18. [1 points] *True or False?* In the limit of infinite training and test data, consistent estimators always give at least as low a test error as biased estimators.

19. [1 points] *True or False?* Stepwise regression with an RIC penalty is an unbiased estimator.

20. [1 points] *True or False?* Leave-one out cross validation (LOOCV) generally gives less accurate estimates of true test error than 10-fold cross validation.

21. [2 points] The optimal code for a long series of symbols A, B, C, D drawn with probabilities $P(A) = 1/2$, $P(B) = 1/4$, $P(C) = P(D) = 1/8$ should take on average how many bits per symbol?

    (a) 0.5

    (b) 1

    (c) 1.5

    (d) 1.75

(e) none of the above

22. [2 points] If we we want to encode a true distribution $P(A) = 1/2$, $P(B) = 1/2$, $P(C) = P(D) = 0$ with an estimated distribution $P(A) = 1/2$, $P(B) = 1/4$, $P(C) = P(D) = 1/8$ what is the KL divergence (using log based 2) between them?

    (a) 0.5
    (b) 1
    (c) 4/3
    (d) 1.5
    (e) none of the above

23. [2 points] Given a coding that would be optimal for a stream of A's, B's, and C's generated from the probability distribution ($p_A = 1/2, p_B = 1/4, p_C = 1/4$), How many bits would it take to code the data stream "AABB"?

    (a) 1
    (b) 2
    (c) 4
    (d) 6
    (e) none of the above

24. [2 points] We have a stream of A's, B's, and C's generated from the probability distribution ($p_A = 1/2, p_B = 1/2, p_C = 0$). However, we instead use a coding that would be optimal if the distribution was ($p_A = 1/2, p_B = 1/4, p_C = 1/4$). Our coding will (in expectation)

    (a) take more bits than the optimal coding
    (b) take fewer bits than the optimal coding
    (c) not enough information is provided to tell

25. [1 points] *True or False?* Adding a feature to a linear regression model during streamwise regression always decreases model bias.

26. [1 points] *True or False?* Adding a feature to a linear regression model during streamwise regression always increases model variance.

27. [1 points] *True or False?* 5-NN has lower bias than 1-NN.

28. [1 points] *True or False?* 5-NN is more robust to outliers than 1-NN.

29. [1 points] *True or False?* Choosing different norms does not affect the decision boundary of 1-NN.

    In the following four questions, assume both trees are trained on the same data.

30. [1 points] *True or False?* A tree with depth of 3 has higher variance than a tree with depth of 1.

31. [1 points] *True or False?* A tree with depth of 3 has higher bias than a tree with depth 1.

32. [1 points] *True or False?* A tree with depth of 3 never has higher training error than a tree with depth 1.

33. [1 points] *True or False?* A tree with depth of 3 never has higher test error than a tree with depth 1.

34. [1 points] *True or False?* If decision trees such as the ones we built in class are allowed to have decision nodes based on questions that can have many possible answers (e.g. "What country are you from) in addition to binary questions, they will in general tend to add the multiple answer questions to the tree before adding the binary questions

    In the following three questions, assume models are trained on the same data without transformations or interactions.

35. [1 points] *True or False?* K-nearest neighbors will always give a linear decision boundary.

36. [1 points] *True or False?* Decision trees with depth one will always give a linear decision boundary.

37. [1 points] *True or False?* Gaussian Naive Bayes will always give a linear decision boundary.

38. [1 points] *True or False?* Logistic Regression will always give a linear decision boundary.

39. [2 points] Suppose you have picked the parameter θ for a model using 10-fold cross validation. The best way to pick a final model to use and estimate its error is to

(a) pick any of the 10 models you built for your model; use its error estimate on the held-out data

(b) pick any of the 10 models you built for your model; use the average CV error for the 10 models as its error estimate

(c) average all of the 10 models you got; use the average CV error as its error estimate

(d) average all of the 10 models you got; use the error the combined model gives on the full training set

(e) train a new model on the full data set, using the θ you found; use the average CV error as its error estimate

40. [2 points] Suppose we want to compute 10-Fold Cross-Validation error on 100 training examples. We need to compute error N_1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N_2, and test the model on the data of size N_3.

    What are the appropriate numbers for N_1, N_2, N_3?

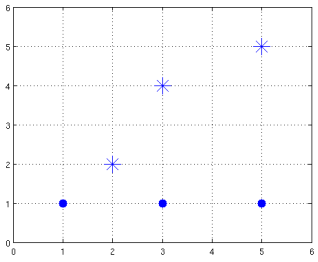
    (a) $N_1 = 10, N_2 = 90, N_3 = 10$

    (b) $N_1 = 1, N_2 = 90, N_3 = 10$

    (c) $N_1 = 10, N_2 = 100, N_3 = 10$

    (d) $N_1 = 10, N_2 = 100, N_3 = 100$

    The following question refers to this picture, which shows a set of 6 points in a two-dimension space, where each point is either labeled "." or "".*



41. [2 points] What is the training error (fraction labeled wrong) in the above picture using 1-nearest neighbor and the L_1 distance norm between points? *If there is more than one equally close nearest neighbor, use the majority label of all the equally close nearest neighbors. Please make sure you really are computing training error, not testing error.*

(a) 0
(b) 1/6
(c) 2/6
(d) 3/6
(e) 4/6

42. [2 points] MLE estimates are often undesirable because

(a) they are biased
(b) they have high variance
(c) they are not consistent estimators
(d) None of the above

43. [2 points] For a data set with p features, of which q will eventually enter the model, streamwise feature selection will test approximately how many models?

(a) p
(b) q
(c) pq
(d) p^2

44. [2 points] For a data set with p features, of which q will eventually enter the model, stepwise feature selection will test approximately how many models?

(a) p
(b) q
(c) pq
(d) p^2

45. [2 points] Which penalty should you use if you are doing stepwise regression and expect 10 out of 100,000 features, with n = 100 observations?

(a) AIC
(b) BIC
(c) RIC

46. [2 points] Which penalty should you use if you are doing stepwise regression and expect 200 out of 1,000 features, with n = 10,000,000 observations?

    (a) AIC
    (b) BIC
    (c) RIC

47. [2 points] You are doing feature selection with 1,000 features and 1,000 observations. Assume roughly half of the features will be selected. What would be the best (or least bad) MDL model?

    (a) AIC
    (b) BIC
    (c) RIC
    (d) Both AIC and BIC are equally good
    (e) Both AIC and RIC are equally good

48. [2 points] Which of the following best describes what discriminative approaches try to model? (w are the parameters in the model)

    (a) $p(y|x, w)$
    (b) $p(y, x)$
    (c) $p(x|y, w)$
    (d) $p(w|x, y)$
    (e) $p(y|w)$

49. [2 points] Which of the following tends to work best on small data sets (few observations)?

    (a) Logistic regression
    (b) Naive Bayes

50. [2 points] Suppose you have a three class problem where class label $y \in 0, 1, 2$ and each training example X has 3 binary attributes $X_1, X_2, X_3 \in 0, 1$. How many parameters do you need to know to classify an example using the Naive Bayes classifier?

    (a) 5

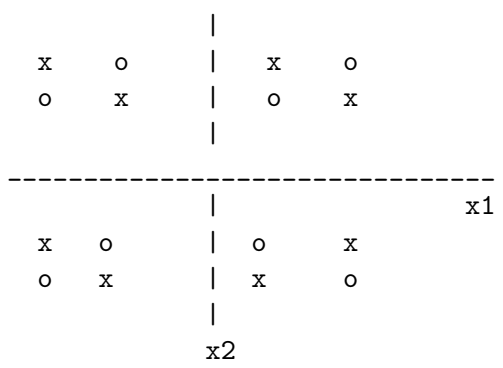
(b) 9

(c) 11

(d) 13

(e) 23

51. [1 points] *True or False* Radial basis functions could be used to make the following dataset be linearly separable.



52. [2 points] Given the following table of observations, calculate the information gain $IG(Y|X)$ that would result from learning the value of X

X	Y
Red	True
Green	False
Brown	False
Brown	True

(a) 1/2

(b) 1

(c) 3/2

(d) 2

(e) none of the above

53. [2 points] We are comparing two models A and B for the same data set.
We find that model A requires 127 bits to code the residual and 260

bits to code the model.
We find that model B requires 160 bits to code the residual and 250 bits to code the model.

(a) model B is underfit compared to A
(b) model B is overfit compared to A
(c) not enough information is provided to tell if model B is over- or under-fit relative to A

54. [2 points] In fitting some data using radial basis functions with kernel width σ, we compute a training error of 345 and a testing error of 390.

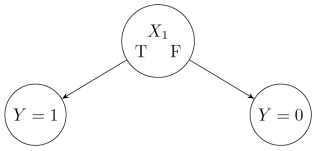
(a) increasing σ will most likely reduce test set error
(b) decreasing σ will most likely reduce test set error
(c) not enough information is provided to determine how σ should be changed

55. [2 points] When choosing one feature from $X_1, ..., X_n$ while building a Decision Tree, which of the following criteria is the most appropriate to maximize? (Here, H() means an entropy, and P() means a Probability)

(a) $P(Y|X_j)$
(b) $P(Y) - P(Y|X_j)$
(c) $H(Y) - H(Y|X_j)$
(d) $H(Y|X_j)$
(e) $H(Y) - P(Y)$

56. [2 points] How many bits are required to code the decision tree (below) which is derived from the data set below it?

(a) 3
(b) 5
(c) 7
(d) None of above



X_1	X_2	X_3	Y
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	0

57. [2 points] How many bits are required to code the residual from using the above decision tree to code the above data set?

    (a) 1
    (b) 2
    (c) 3
    (d) None of above

58. [2 points] In each round of AdaBoost, the misclassification penalty for a particular training observation is increased going from round t to round $t + 1$ if the observation was...

    (a) classified incorrectly by the weak learner trained in round t
    (b) classified incorrectly by the full ensemble trained up to round t
    (c) classified incorrectly by a majority of the weak learners trained up to round t
    (d) B and C
    (e) A, B, and C

59. [1 points] *True or False?* Boosting minimizes an exponential loss function (subject to the model constraints).

60. [2 points] Which of the following regularization methods are scale-invariant

    (a) L0 and L2 but not L1
    (b) L1 and L2 but not L0
    (c) L0 but not L1 or L2
    (d) L2 but not L0 or L1
    (e) None of the above