UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Midterm 2019

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 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 **40** questions and **61** points (19 one point questions, 21 two point questions).

1. [2 points] Compared to MLE, MAP solutions tend to have

   (a) lower bias and lower variance

   (b) higher bias and lower variance

   (c) lower bias and higher variance

   (d) higher bias and higher variance

2. [2 points] We trained a three-way logistic regression and obtained weights

$$w_a = (1, 1, 0), w_b = (-1, 1, 1), w_c = (2, 1, 2)$$

   What label would be given to the point $x = (0, 1, 1)$?

   (a) A

   (b) B

   (c) C

3. [2 points] As part of building a decision tree, we want to measure the information gain from asking a question about a binary split on feature X_1 of the following 4 samples:

X_1	y
T	0
F	1
F	1
F	0

   Which of the following measures the information gain of splitting the sample on the value of X_1?

   (a) $-(\frac{2}{3}) \log_2(\frac{2}{3}) + (\frac{1}{2}) \log_2(\frac{1}{2})$

   (b) $-\log_2(\frac{1}{2}) + \frac{1}{4}(\frac{2}{3} \log_2(\frac{2}{3}) + \frac{1}{3} \log_2(\frac{1}{3})) + \frac{3}{4}(\frac{2}{3} + \log_2(\frac{2}{3}) + \frac{1}{3} \log_2(\frac{1}{3}))$

   (c) $-\log_2(\frac{1}{2}) + \frac{3}{4}(\frac{2}{3} \log_2(\frac{2}{3}) + \frac{1}{3} \log_2(\frac{1}{3}))$

   (d) $-\log_2(\frac{1}{2}) + \frac{1}{4}(\frac{2}{3} \log_2(\frac{2}{3}) + \frac{1}{3} \log_2(\frac{1}{3}))$

   (e) none of the above

4. [2 points] The depth of a decision tree is most likely to scale as

   (a) the number of training examples

   (b) the square root of the number of training examples

   (c) the log of the number of the training examples

5. [2 points] You are using a dataset for housing with one feature (area) in squared meters. Your friend Margaret is also using the same data set but with the mentioned feature in squared feet. (A meter is roughly 3 feet.) You both are using nearest neighbors with L2 distance. What do you expect to see?

   (a) The two models will both give the exact same predictions.

   (b) Area is **more** important when measured in feet than in meters.

   (c) Area is **less** important when measured in feet than in meters.

   (d) The two models will give different predictions, but we can't know the relative effect of area.

6. [2 points] Similar to the question above, you are using a dataset for housing with one feature (area) in squared meters. Your friend Margaret is also using the same data set but with the mentioned feature in squared feet. However, this time, you are both are experimenting with decision trees. What do you expect to see?

   (a) The two models will both give the exact same predictions.

   (b) Area is **more** important when measured in feet than in meters.

   (c) Area is **less** important when measured in feet than in meters.

   (d) The two models will give different predictions, but we can't know the relative effect of area.

7. [2 points] Traditionally, when we have a real-valued input attribute during decision-tree learning, we consider 9 binary splits according to whether the attribute is in the lowest 10%, lowest 20%, etc. Your friend Pat suggests that instead we should just have a 10-way split with one branch for each of the 10 "bins" of the attribute (lowest 10%, next 10%, etc). The single biggest problem with Pat's suggestion is:

   (a) It is too computationally expensive.

   (b) It would probably result in a decision tree that scores badly on the training set and a test set.

   (c) It would probably result in a decision tree that scores well on the training set but badly on a test set.

   (d) It would probably result in a decision tree that scores well on a test set but badly on a training set.

8. [2 points] Suppose we compute the MAP estimate of the mean μ of a Gaussian with a fixed variance, n samples and the following prior on μ:

$$p(\mu|\mu_0, \sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}}$$

   Under what condition(s) does the MAP estimate of μ converge to the MLE estimate of μ?

   (a) $\mu_0 \to 0$

   (b) $n \to \infty$

   (c) $\sigma_0 \to \infty$

   (d) two of the above

   (e) all of the above

9. [2 points] As k in k-NN increases:

   (a) the bias increases and the variance increases

   (b) the bias increases and the variance decreases

   (c) the bias decreases and the variance increases

   (d) the bias decreases and the variance decreases

10. [1 points] *True or False?* The larger the regularization penalty λ in penalized regression, the more the model will tend to overfit the training data.

11. [1 points] Which norm most heavily shrinks large weights (weights that are much bigger than 1)?

    (a) L_0

    (b) L_1

    (c) L_2

    (d) not enough information to tell

12. [2 points] Suppose you have a friend who is an esteemed doctor. Your friend has a dataset containing gene expression data for patients. For each patient, there is a vector containing the expression levels of $10,000$ genes, along with a label indicating whether the patient developed a certain disease. Using this, your friend's goal is to learn a model where, given a new patient (also represented by a similar $10,000$-dimensional vector), the model can output the probability of this patient developing the disease. Your friend also believes a reasonable model will depend on at least $9,000$ of the genes. Which of the following methods would be most suitable for this problem?

    (a) Linear least squares regression with L_2 penalty

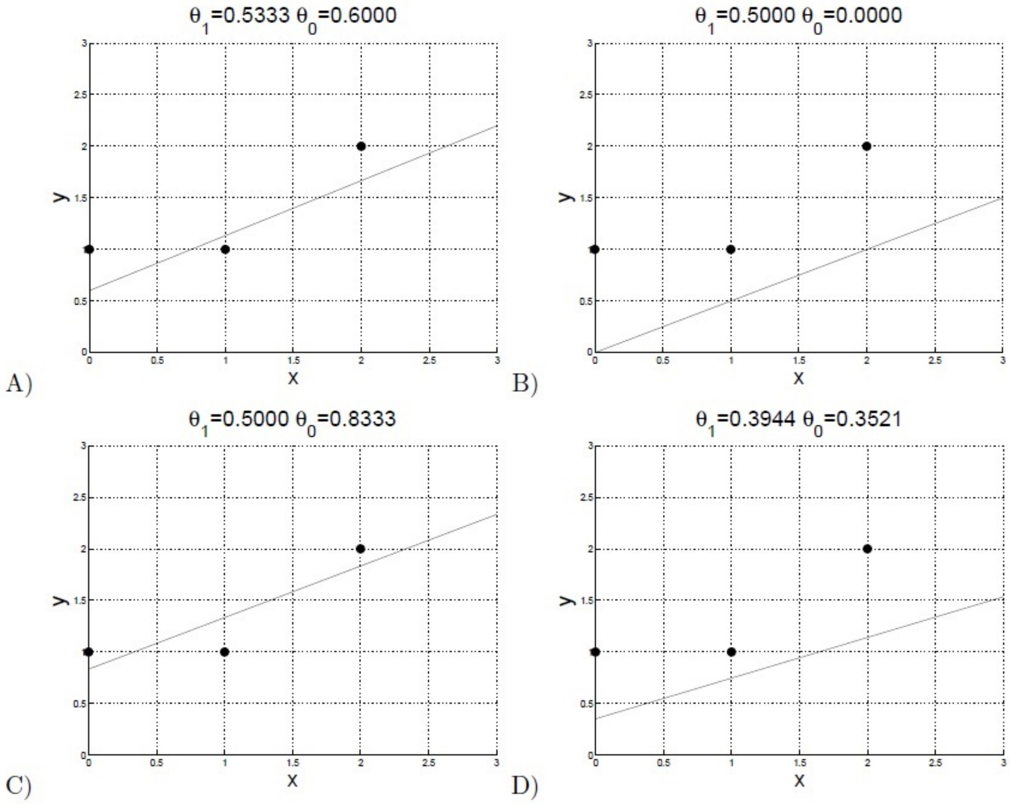
    (b) Linear least squares regression with L_1 penalty

    (c) Logistic regression with L_2 penalty

    (d) Logistic regression with L_1 penalty

    (e) Decision Tree

Models A-D below are fit by linear regression on a simple three-example data set $\{(1,1),(0,1),(2,2)\}$ with one independent variable x and one dependent variable y, using OLS or L_1 or L_2-penalized regression.



Use the plots above to answer **the next three questions**:

13. [2 points] Which model was most likely to have been generated using OLS?

    (a) Model A

    (b) Model B

    (c) Model C

    (d) Model D

    (e) Not enough information

14. [2 points] Which model was most likely to have been generated using L_1 regularization?
    (a) Model A
    (b) Model B
    (c) Model C
    (d) Model D
    (e) Not enough information

15. [2 points] Which model was most likely to have been generated using L_2 regularization?
    (a) Model A
    (b) Model B
    (c) Model C
    (d) Model D
    (e) Not enough information

---

16. [2 points] In order to check whether a function $k(x_1, x_2)$ is a kernel function, we used a matrix X to generate a potential kernel matrix K. If the matrix K has eigenvalues 1.0, 2.3, 3.7, and 4.2 then we can conclude that
    (a) $k(x_1, x_2)$ is definitely a kernel function
    (b) $k(x_1, x_2)$ cannot be a kernel function
    (c) $k(x_1, x_2)$ might be a kernel function

17. [2 points] In order to check whether a function $k(x_1, x_2)$ is a kernel function, we used a matrix X to generate a potential kernel matrix K. If the matrix K has eigenvalues -1.0, 2.3, 3.7, and 4.2 then we can conclude that
    (a) $k(x_1, x_2)$ is definitely a kernel function
    (b) $k(x_1, x_2)$ cannot be a kernel function
    (c) $k(x_1, x_2)$ might be a kernel function

18. [1 points] *True or False?* Stepwise linear regression will always find at least as accurate a model as streamwise regression, assuming the same regularization penalty is used in both cases.

19. [1 points] *True or False?* The result of streamwise regression depends on the order in which features are added.

20. [1 points] *True or False?* In stepwise regression, the model which includes all the features will always be considered.

21. [2 points] You want to fit a linear regression model with "the best" 10 features out of 200 candidate features. Which of the following will work best?

    (a) Perform L_2 regularization

    (b) Perform L_1 regularization

    (c) Use the hinge loss instead of square error loss.

22. [1 points] *True or False?* For logistic regression, gradient descent can converge to a local minimum and fail to find a global minimum. More advanced optimization methods are thus often used.

23. [1 points] *True or False?* A larger stepsize in gradient descent can lead to faster convergence, but lower accuracy.

24. [1 points] *True or False?* The key idea of AdaGrad is to learn slowly from frequent features but pay attention to rare but informative features.

25. [1 points] *True or False?* The decision boundary for logistic regression in p dimensions is a $p-1$-dimensional hyperplane.

26. [2 points] The radial basis functions used in RBFs transform a feature vector of dimension p to a new space of dimension k such that

    (a) $k < p$

    (b) $k = p$

    (c) $k > p$

    (d) there is no constraint on how k and p are related

27. [1 points] *True or False?* SVMs, unlike neural nets, have a global optimum.

28. [1 points] *True or False?* The error surface followed by the gradient descent backpropagation algorithm changes if we change the training data.

29. [2 points] Which of the following statement about KL-divergence is **not** true?

    (a) KL-divergence is a measure of how different two distributions are

    (b) KL-divergence is 0 if and only if two distributions P and Q are equal

    (c) KL-divergence can sensibly be measured between any two vectors of equal length

    (d) KL-divergence is always non-negative.

30. [1 points] *True or False?* A max pooling layer in a ConvNet has an equal number of weights to learn as a filter of the same size.

31. [1 points] *True or False?* (Standard) GANS are neural networks composed of two subcomponent neural nets. One neural net (the generator) takes as input an image and generates a different, fake image. The other neural net (the discriminator) takes as input either a real image or a fake image and tries to determine whether it was real or fake.

32. [2 points] Assume a convolutional neural net where the input is a 4×4 RGB image (i.e., 4×4×3), with 2 filters, each of size 2×2×3, a stride of 2, and we zero pad the image with zeros on all four sides (giving a 6×6×3 "input").

    How many nodes are there in the first hidden layer?

    (a) 6×6×3×2
    (b) 5×5×2
    (c) 4×4×2
    (d) 3×3×2
    (e) none of the above

33. [1 points] *True or False?* When fitting a model using Adaboost, training should stop before the training error reaches zero.

34. [2 points] Suppose we are fitting a model using gradient boosting, where we call η the learning rate of the model update:

    $$F_m(x) = F_{m-1}(x) + \eta \cdot \gamma_{m-1} h_{m-1}(x),\ 0 < \eta \leq 1$$

    Additionally, each base learner $h_{m-1}(x)$ is fit with a randomly subsampled fraction f, $0 < f \leq 1$ of the training data.

    Which of the following would most reduce overfitting?

    (a) increasing η, increasing f
    (b) decreasing η, increasing f
    (c) increasing η, decreasing f
    (d) decreasing η, decreasing f

35. [1 points] *True or False?* When fitting a model using gradient boosting and a squared loss, we fit the base learner to the (standard) residuals at each iteration.

36. [1 points] *True or False?* In AdaBoost, in each iteration the weights of the misclassified examples on any given iteration all go up by the same multiplicative factor.

37. [1 points] *True or False?* AdaBoost will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

38. [2 points] Which of the following **least** penalizes misclassifications as a function of how far they are from the decision boundary?
    (a) hinge loss
    (b) logistic loss
    (c) exponential loss

39. [1 points] *True or False?* Hinge loss functions only penalize points for which the predicted classification label is wrong.

40. [1 points] *True or False?* The number of support vectors found by an SVM is independent of the magnitude of the penalty on the slack variables.