UNIVERSITY OF PENNSYLVANIA

CIS 520: Machine Learning
Midterm, 2018

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 **46** questions, worth a total of **70** points.

1. [1 points] MLE estimation of a model $y = f(x; \theta) + \epsilon$ where ϵ is mean zero Gaussian noise is the same as minimizing the ___ error.
   (a) L_0
   (b) L_1
   (c) L_2
   (d) L_∞

2. [2 points] Increasing k in k-nearest neighbor models will:
   (a) Increase bias, increase variance
   (b) Increase bias, decrease variance
   (c) Decrease bias, increase variance
   (d) Decrease bias, decrease variance

3. [2 points] In a least-squares linear regression problem, adding an L_2 regularization penalty always decreases the expected sum of squares error of the solution w on unseen test data.
   (a) True
   (b) False

4. [1 points] In a least-squares linear regression problem, adding an L_2 regularization penalty cannot decrease the sum of squares error of the solution w on the training data.
   (a) True
   (b) False

5. [2 points] We have some data $D = \{x_i, y_i\}$, and we assume a simple linear model of this data with Gaussian noise as follows:

   $Y = w^\top X + b + Z$, with $Z \sim N(0, \sigma^2)$

   We will further assume a prior on w, that means $w_j \sim N(0, \lambda^2)$. Then, in which case does MAP **not** reduce to MLE?
   (a) $\lambda \to \infty$
   (b) $\sigma \to \infty$
   (c) $N \to \infty$ (Here N means the number of samples)
   (d) $\frac{\lambda}{\sigma} \to \infty$

6. [2 points] Suppose we want to fit the following regression prediction model: $h(x) = c$, which is constant for all x. Suppose the actual underlying model that generated the data is $y = ax$, where a is a constant slope. In other words, we are modeling the underlying linear relation with a constant model. Let us now try to compute the bias and variance of our method. Assume that $x \sim N(\mu, \sigma^2)$. Compute the **average** hypothesis $h(x)$ over datasets $D = \{x_1, ..., x_n\}$ (Here we use the ordinary least squares estimate to estimate $h(x; D)$):

   (a) $a\mu$
   (b) $\frac{a}{\sigma^2}$
   (c) $\frac{a}{\sigma}$
   (d) $\frac{a}{\mu}$

7. [1 points] For any probability model, the log-likelihood function of data generated from this model is always guaranteed to be concave.

   (a) True
   (b) False

8. [1 points] Any Naïve Bayes classifier that assumes the features are drawn from Gaussian distributions can always be written as a linear classifier.

   (a) True
   (b) False

9. [1 points] Which type of regularization leads to sparser solutions (fewer non-zero weights)?

   (a) L_2
   (b) L_1
   (c) neither

10. [1 points] If the complexity of a model increases, then which of the following is expected to increase?

   (a) Bias
   (b) Variance

11. [2 points] The L_0 pseudo-norm of a vector $\mathbf{w}$ of length n is defined as

(a) $\sum_{i=1}^{n} |w_i|$

(b) $\sum_{i=1}^{n} \mathbb{I}(w_i \neq 0)$, where $\mathbb{I}$ takes value 1 when $w_i \neq 0$, and 0 otherwise.

(c) $\sqrt{\sum_{i=1}^{n} w_i^2}$

(d) None of the above

12. [2 points] The solution to the following L_1 regularized least-squares regression

$$\operatorname*{argmin}_{\mathbf{w}} \|Y - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 ,$$

for $\lambda > 0$ is:

(a) $(X^\top X)^{-1}X^\top Y$

(b) $(X^\top X + \lambda I)^{-1}X^\top Y$

(c) The objective is unbounded, i.e. the solution is $-\infty$.

(d) None of the above

13. [1 points] Which of the following sequences has a higher description length (in bits) – a) 111111111111.... or b) 010101010101.....?

(a) a

(b) b

14. [2 points] In MDL, reducing the description length of a model also reduces the description length of the residual error of that model.

(a) True

(b) False

15. [1 points] Kernels such as radial basis functions can often be used with support vector machines to make data linearly separable, even when the data was not linearly separable before the kernel was applied.

(a) True

(b) False

16. [1 points] When doing 10-fold cross-validation, standard practice for making predictions on new points is to take an ensemble (e.g. average or majority vote) of the ten models which are built.

(a) True

(b) False

17. [1 points] Given features $X = [X_1, X_2, \ldots, X_p]$ and output Y, the reduction in the entropy of Y from observing feature X_p is given by $H(X_p) - H(X_p|Y)$

    (a) True

    (b) False

18. [1 points] The entropy for a Gaussian, Y, given by $H(Y) = -\int p(y)\log(p(y))dy$ is independent of its mean, $E[Y]$.

    (a) True

    (b) False

19. [2 points] $\mathbf{E}(f(X))$ for a continuous random variable X is given by

    (a) $f(\mathbf{E}(X))$ for all functions f

    (b) $f(\mathbf{E}(X))$ if $f(X)$ is a linear function of X

    (c) $\int f(x)p(x)dx$ where $p(x)$ is the probability density function of X

    (d) both (b) and (c)

    (e) all (a), (b) and (c)

20. [1 points] In logistic regression, adding Gaussian priors to the parameters w, i.e. $w_j \sim \mathcal{N}(0, \lambda^2)$ is equivalent to adding a quadratic penalty on the parameters in the objective function defined by the loglikelihood.

    (a) True

    (b) False

21. [1 points] Adding an L_1 penalty on the parameters in a regression problem is equivalent to a prior that the weights are small and will, in general, shrink all of the parameters.

    (a) True

    (b) False

22. [1 points] The curve defined by $\|x\|_{0.5} = 2$ (where $\|x\|_{0.5}$ is the $L_{0.5}$ norm) is a convex set.

    (a) True

(b) False

23. [2 points] The following function can be interpreted as a probability density:

$$f(x) = \begin{cases} 2, & |x| \leq 1/2 \\ 0, & |x| > 1/2 \end{cases}$$

(a) True

(b) False

24. [1 points] The expected value of the testing error in approximating a true y by a model $\widehat{y} = f(x;\theta)$ is equal to the sum of the expected value of the bias on the training set plus the expected value of the variance on the training set, plus the irreducible uncertainty (the variance of the noise).

(a) True

(b) False

25. [2 points] Assume a variable can take on three values, A, B, and C with probabilities either given by $p = [p_A, p_B, p_C] = [1/2, 1/4, 1/4]$ (I.e., $p_A = 1/2, p_B = 1/4, p_C = 1/4$)) or by $q = [q_A, q_B, q_C] = [1/2, 1/2, 0]$. The KL divergence $KL(p, q)$ is equal to

(a) 0

(b) $-(1/2)\log(1/2) - (1/2)\log(1/4)$

(c) infinity

(d) none of the above

26. [2 points] The MLE estimate of weights $\widehat{w}$ for ordinary least squares gives estimates of the (unknown) true weight w that are distributed as $\widehat{w} \sim N(w, \sigma^2/n)$. If we use Ridge regression instead, the expected value of $\widehat{w}$ will

(a) remain $\widehat{w}$

(b) become larger in magnitude

(c) become smaller in magnitude

(d) we can't say

27. [2 points] The MLE estimate of weights for ordinary least squares gives $\widehat{w} \sim N(w, \sigma^2/n)$. If we use Ridge regression instead, the variance of $\widehat{w}$ will

    (a) remain σ^2/n
    (b) become larger in magnitude
    (c) become smaller in magnitude
    (d) we can't say

28. [1 points] The training error of 1-NN is always zero.

    (a) True
    (b) False

29. [1 points] A classifier trained on less training data is less likely to overfit.

    (a) True
    (b) False

30. [1 points] A gradient descent algorithm with a properly chosen fixed step size for training a logistic regression model almost always converges to the exact value of the optimal regression weights.

    (a) True
    (b) False

31. [2 points] Let $X_1, X_2 ... X_n$ be iid samples from Uniform$(-w, w)$, i.e.

$$f_X(x) = \begin{cases} 0 & \text{if } |x| > w \\ \frac{1}{2w} & \text{if } |x| \leq w \end{cases}$$

    where $w > 0$ is an unknown parameter. The MLE estimate of w is

    (a) $\frac{\sum_{i=1}^{n} |X_i|}{n}$
    (b) $\frac{n}{\sum_{i=1}^{n} 2|X_i|}$
    (c) $\max_i |X_i|$
    (d) $\max_i \frac{1}{2|X_i|}$

32. [2 points] Consider the two class problem where class label $y \in \{T, F\}$ and each training example X has 2 binary attributes X_1, $X_2 \in \{T, F\}$. How many parameters will you need to know/estimate if you are to classify an example using the Naïve Bayes classifier?

    (a) 5
    (b) 8
    (c) 3
    (d) 7

33. [2 points] Again, consider the two class problem where class label $y \in \{T, F\}$ and each training example X has 2 binary attributes X_1, $X_2 \in \{T, F\}$. How many parameters will you need to estimate if we do **not** make the Naïve Bayes conditional independence assumption?

    (a) 3
    (b) 5
    (c) 7
    (d) 8

34. [2 points] Consider the following two sets in the two-dimensional plane:

$$C = \left\{\mathbf{x} \in \mathbb{R}^2 \;\middle|\; 0 \leq x_1 \leq 2\right\} \cap \left\{\mathbf{x} \in \mathbb{R}^2 \;\middle|\; 2 \leq x_1 \leq 4\right\}$$
$$D = \left\{\mathbf{x} \in \mathbb{R}^2 \;\middle|\; 0 \leq x_1 \leq 2\right\} \cup \left\{\mathbf{x} \in \mathbb{R}^2 \;\middle|\; 2 \leq x_1 \leq 4\right\}$$

    Select the most accurate statement:

    (a) Both C and D are convex.
    (b) C is convex, D is not convex.
    (c) C is not convex, D is convex.
    (d) Neither C nor D is convex.

35. [2 points] Consider the following functions of two variables:

$$f(\mathbf{x}) = \max(x_1, -x_2)$$
$$g(\mathbf{x}) = -x_1^2 - x_2^2$$

    Select the most accurate statement:

    (a) Both f and g are convex.

(b) f is convex, g is concave.
(c) f is convex, g is neither convex nor concave.
(d) Neither f nor g is convex.

36. [2 points] Consider a primal optimization problem with two inequality constraints, and suppose you form the Lagrange dual problem over two dual variables λ_1, λ_2. Which of the following functions of two variables *cannot* be the objective function of the dual problem?

    (a) $\phi(\boldsymbol{\lambda}) = \lambda_1 - \lambda_2$
    (b) $\phi(\boldsymbol{\lambda}) = \lambda_1^2 + \lambda_2^2$
    (c) $\phi(\boldsymbol{\lambda}) = \lambda_1^2 - \lambda_2^2$
    (d) All three can be dual objective functions.

37. [2 points] Let $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be two symmetric, positive definite kernel functions. Which of the following *cannot* be a valid kernel function?

    (a) $K(x, x') = 5 \cdot K_1(x, x')$
    (b) $K(x, x') = K_1(x, x') + K_2(x, x')$
    (c) $K(x, x') = K_1(x, x') + \dfrac{1}{K_2(x, x')}$
    (d) All three are valid kernels.

38. [2 points] You are trying to impress your friend with your photographs. Over the last few months, you have observed which photographs she gives a 'like' to and which she does not. Based on these examples, you want to estimate the probability that she will like your new photograph. Which of the following machine learning methods would be **least** useful for this problem?

    (a) Logistic regression
    (b) k-nearest neighbor
    (c) Support vector machines

39. [2 points] The linear (soft margin) support vector machine algorithm learns a weight vector $\mathbf{w}$ (and possibly a bias term b). What sort of regularization does it effectively perform on $\mathbf{w}$?

    (a) L_1 regularization

(b) L_2 regularization

(c) regularization other than L_1 and L_2

(d) no regularization

40. [2 points] Suppose you are training a support vector machine classifier using a polynomial kernel $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^q$ for degrees $q = 1, 2, 3, 4$. Assuming that for each value of q, you select the SVM parameter C by cross-validation over a large enough range of values, which of the following scenarios are realistic (could be expected to arise in practice)?

(a)

q	1	2	3	4
Train error	0.35	0.30	0.24	0.17
Test error	0.25	0.21	0.18	0.23

(b)

q	1	2	3	4
Train error	0.17	0.24	0.29	0.33
Test error	0.29	0.25	0.32	0.36

(c)

q	1	2	3	4
Train error	0.24	0.21	0.19	0.13
Test error	0.31	0.26	0.23	0.27

(d) All three scenarios are realistic.

41. [2 points] Suppose you are solving a regression problem. You have 1000 labeled examples $\left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{1000}$ and decide to use ridge regression:

$$\min_{\mathbf{w}} \sum_{i=1}^{1000} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2 \,.$$

To choose λ, you perform cross-validation over some range of values; let's say a value λ_1 is selected. After some time, you get another 1000 labeled examples $\left\{(\mathbf{x}_i, y_i)\right\}_{i=1001}^{2000}$. You decide to re-learn your model using all the data; you again select λ using cross-validation. Let's call this second value λ_2. What is the relationship you would expect between λ_1 and λ_2 ?

(a) λ_1 is expected to be smaller than λ_2

(b) λ_2 is expected to be smaller than λ_1

(c) λ_1 and λ_2 will be approximately the same size

(d) there is not enough information to tell

42. [2 points] In a convolutional neural net with an image of size 5x5x3 (where 3 is red/green/blue), we pad with a single zero all around the image and then use 4 local receptive fields ('filters') of size 3x3x3 and a stride of size 2. The outputs of these local receptive fields are sent to a single output. Assuming that there are no bias terms in this model, the total number of weights in the network is.

    (a) 3*3*4 + 4
    (b) 3*3*3*4 + 4
    (c) 7*7*4 + 4
    (d) 7*7*7*4 + 4
    (e) none of the above

43. [1 points] When learning neural networks, one should always use an L_2 loss function rather than a log-likelihood.

    (a) True
    (b) False

44. [1 points] When learning a decision tree, a feature x_j which is not correlated with the label y (i.e., $corr(x_j, y) = 0$) will never be split on and hence will never to used in the tree.

    (a) True
    (b) False

45. [1 points] When learning a decision tree, features that have many possible values (e.g. colors or countries) tend to be more likely to be selected than binary features.

    (a) True
    (b) False