Name _____ Student No. _____

Midterm Examination

CSE 474/574: Introduction to Machine Learning

6:30 PM - 7:50 PM Hoch 114

Monday, Oct 17, 2016

No. of Questions	**Topic**	**Points**
5	Linear Algebra	10
10	Probability Theory	30
15	Regression	60
	Total	100

1 Linear Algebra

1. Is the following statement true or false? An array arranged on a regular grid with variable number of axes is referred to as a tensor.

   A) True
   B) False

   A. Refer to the slides

2. Which of the following properties does matrix multiplication have? (Could have more than one possible choice)

   A) Distributivity over addition, $A(B+C) = AB + AC$
   B) Associativity, $A(BC) = (AB)C$
   C) Commutativity, $AB = BA$
   D) $(AB)^\top = A^\top B^T$
   E) $(AB)^{-1} = A^{-1}B^{-1}$

   A, B

3. What are the advantages of using on-line gradient descent to solve simultaneous linear equations compared to using the closed-form solution? (Could have more than one possible choice)

   A) On-line gradient descent requires less memory.
   B) On-line gradient descent is more numerically stable.
   C) On-line gradient descent can deal with the case when the matrix to be inversed in the closed-form solution is singular.
   D) On-line gradient descent gives more accurate solution.
   E) On-line gradient descent is faster.

   A, B, C. On-line gradient descent takes one data point at a time while the closed-form solution requires the design matrix which consists all data to be stored in memory. Gaussian elimination becomes numerically unstable when the matrix is close to singularity. In such cases, using online gradient descent is more advantageous.

4. Is the following statement true or false? If simultaneous equations $A\mathrm{x} = \mathrm{b}$ has n equations and n unknown variables, it must have one unique solution.

   A) True
   B) False

   B. If there are contradictory equations there won't be valid solutions. If some of the equations are not linearly independent, we could have infinite solutions.

5. Given matrix A whose inverse A^{-1} exists, it is necessary to have

   A) A is non-singular
   B) A is square
   C) $A = A^\top$
   D) $\mathrm{tr}A = 0$

   A, B

2 Probability Theory

6. Is the following statement true or false? $P(A) = \sum_B P(A|B)$

   A) True
   B) False

   B. Sum rule: $P(A) = \sum_B P(A, B)$

7. Is the following statement true or false? $P(A, B, C) = P(A|B, C)P(B|C)P(C)$

   A) True
   B) False

   A. Apply the product rule to the l.h.s. twice:$P(A, B, C) = P\left(A, B|C\right)P\left(C\right) = P(A|B, C)P(B|C)P(C)$

8. Is the following statement true or false? For bivariate Gaussian distribution, the conditional distribution of one variable conditioned on the other is Gaussian distribution.

   A) True
   B) False

   A. This is the property of multivariate Gaussian, refer to textbook P85.

2.1 Discrete Probability Distribution 1

Questions in this subsection concern properties of the random variables A and B which have the following joint probability distribution $P(A, B)$.

		B	
		0	1
A	0	1/4	1/4
	1	3/8	1/8

9. A and B are independent random variables _____

A) True

B) False

C) Not enough information to tell

B. Events A and B are independent events if and only if $P(A, B) = P(A) \times P(B)$.

10. Calculate the mean of A:_____.

A) 1

B) $\frac{1}{2}$

C) $\frac{1}{4}$

D) $\frac{2}{3}$

B. $E(A) = 0ŒP(A = 0) + 1ŒP(A = 1) = 1/2$

11. Calculate the variance of A:_____.

A) 1

B) $\frac{1}{2}$

C) $\frac{1}{4}$

D) $\frac{2}{3}$

C. $var(A) = \sum_A [AE(A)]^2ŒP(A) = 1/4$

12. Calculate the entropy of A:_____.

A) 1

B) $\frac{1}{2}$

C) $\frac{1}{4}$

D) $\frac{2}{3}$

A. $Entropy(A) = -\sum_{i=0}^{N} p_i log p_i = log(1/2) = log 2 = 1$

13. Calculate the covariance between A and B:_____.

A) $-\frac{1}{2}$

B) $-\frac{1}{4}$

C) $-\frac{1}{8}$

D) $-\frac{1}{16}$

D. $E[AB] = 1/8; E[A] = 1/2; E[B] = 3/8; Cov(A, B) = E[AB]E[A]E[B] = 1/16$

2.2 Discrete Probability Distribution 2

14. Suppose we are told that data is generated randomly from one of two distributions:

	$P(0)$	$P(1)$	$P(2)$	$P(3)$
Distribution A	0.1	0.2	0.3	0.4
Distribution B	0.4	0.3	0.2	0.1

Suppose that we see a 0's, b 1's, c 2's, and d 3's. Compute the likelihood (or log-likelihood if you prefer) of distributions A and B, in terms of a, b, c, and d. Under what circumstances is A more likely than B? Identify in the list below the one data distribution that makes A the more likely underlying distribution._____

A) $a = 2; \quad b = 8; \quad c = 1; \quad d = 4.$

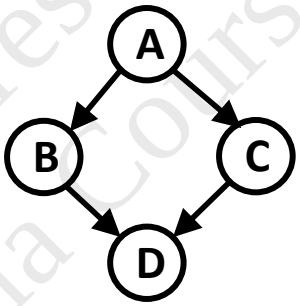
B) $a = 2; \quad b = 8; \quad c = 2; \quad d = 4.$

C) $a = 6; \quad b = 1; \quad c = 4; \quad d = 5.$

D) $a = 3; \quad b = 3; \quad c = 9; \quad d = 1.$

B. The likelihood of seeing a 0s, b 1s, c 2s and d 3s given distribution A is $(.1)^a(.2)^b(.3)^c(.4)^d$, while for distribution B, it is $(.4)^a(.3)^b(.2)^c(.1)^d$. If $(.1)^a(.2)^b(.3)^c(.4)^d > (.4)^a(.3)^b(.2)^c(.1)^d$ then $(1/4)^a(2/3)^b(3/2)^c(4)^d > 1$. If we take logarithms to base 2, and use the approximation $log_2 3 = 1.59$, we getthe simpler condition: $2(d-a) + 0.59(c-b) > 0$.

15. Given the following Bayesian network structure, which is the correct decomposition for the joint distribution?



A) $P(A)P(B)P(C)P(D)$

B) $P(A)P(B|A)P(C|A)P(D)$

C) $P(A|B,C)P(B|D)P(C|D)P(D)$

D) $P(A)P(B|A)P(C|A)P(D|B,C)$

D

3 Regression

16. Is the following statement true or false? In polynomial curve fitting, getting very large value of weights indicate the model over-fits the data.

    A) True
    B) False

    A. Regularization term is used to reduce the value of weights, refer to textbook P11.

17. Is the following statement true or false? Overfitting leads to low training error and low testing error.

    A) True
    B) False

    B. Overfitting leads to low training error and high testing error.

18. Is the following statement true or false? Algorithm A is better than algorithm B if the training error of algorithm A is better than that of B.

    A) True
    B) False

    B. Testing error is more important than training error. When overfitting happens, training error is also low.

19. Is the following statement true or false? In linear regression with squared error function, regularization coefficient λ is non-negative.

    A) True
    B) False

    A. Regularization term is used to reduce the value of weights, so is non-negative.

20. Is the following statement true or false? $\phi_j(x) = x^j$ are called polynomial basis functions.

    A) True
    B) False

    A. This is the definition of polynomial basis functions.

21. When performing linear regression using basis functions for non-linear quadratic in the value of a D-dimensional vector x of input variables, it is better to use sigmoidal basis functions than Gaussian basis functions

    A) True
    B) False

Gaussian Basis Function is better because it is non-linear in the input, x and can be more accurate. In Gaussian Basis Function, we use (x-mu)^2 which makes it non-linear (quadratic) in the input, x. However, Sigmoidal Basis Function is linear in the input, x, to most part and plays a little below Gaussian Function. Because Sigmoidal Basis Function uses just (x-mu).

22. Choose the right combination of closed-form maximum likelihood solution for linear regression with a regularization term. (Could have more than one possible choice)

A) $\mathrm{w}_{\mathrm{ML}} = \left(\lambda I + \Phi^{\top}\Phi\right)^{-1}\Phi^{\top}\mathrm{t}$

B) $\mathrm{w}_{\mathrm{ML}} = \left(\Phi^{\top}\Phi\right)^{-1}\Phi^{\top}\mathrm{t}$

A. This is the definition of the closed-form solution with a regularizor.

23. Which form of the design matrix is suitable for the computation of w_{ML}? (Could have more than one possible choice)

A) $\Phi = \begin{pmatrix} \phi_0(\mathrm{x}_1) & \phi_1(\mathrm{x}_1) & \cdots & \phi_{M-1}(\mathrm{x}_1) \\ \phi_0(\mathrm{x}_2) & & & \vdots \\ \vdots & & & \vdots \\ \phi_0(\mathrm{x}_N) & \cdots & \cdots & \phi_{M-1}(\mathrm{x}_N) \end{pmatrix}$

B) $\Phi = \begin{pmatrix} \phi_0(\mathrm{x}_1) & \phi_0(\mathrm{x}_2) & \cdots & \phi_0(\mathrm{x}_N) \\ \phi_1(\mathrm{x}_1) & & & \vdots \\ \vdots & & & \vdots \\ \phi_{M-1}(\mathrm{x}_1) & \cdots & \cdots & \phi_{M-1}(\mathrm{x}_N) \end{pmatrix}$

A. This is the definition of a design matrix.

24. Choose the correct formula of objective function with quadratic regularization. (Could have more than one possible choice)

A) $E_D(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^{\top}\phi(\mathrm{x}_n)\right\}^2 + \lambda\mathrm{w}^{\top}\mathrm{w}$

B) $E_D(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^{\top}\phi(\mathrm{x}_n)\right\}^2 + \lambda|\mathrm{w}|$

C) $E_D(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^{\top}\phi(\mathrm{x}_n)\right\}^2 + \lambda\mathrm{w}$

A. This the definition of the objective funtion with a quadratic regularizor.

25. Choose the correct formula of gradient descent with quadratic regularization, given $\mathrm{w}^{(\tau+1)} = \mathrm{w}^{(\tau)} - \eta\nabla E_n$. (Could have more than one possible choice)

A) $\nabla E_n = \left[-\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^{\top}\phi(\mathrm{x}_n)\,\phi(\mathrm{x}_n)^{\top}\right\}\right] + \lambda|\mathrm{w}|$

B) $\nabla E_n = \left[-\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^{\top}\phi(\mathrm{x}_n)\,\phi(\mathrm{x}_n)^{\top}\right\}\right] + \lambda\mathrm{w}$

C) $\nabla E_n = \left[-\sum_{n=1}^{N}\left\{t_n - \mathrm{w}^{\top}\phi(\mathrm{x}_n)\,\phi(\mathrm{x}_n)^{\top}\right\}\right] + \lambda\mathrm{w}^{\top}\mathrm{w}$

B.

26. Which of the following statement about $k-$fold cross validation is correct?

A) Divide the data into k partitions, and then pick $k-1$ of the partitions as held-out testing set. Train on one of the k partitions and then test on all $k-1$ partitions that is not used in the training. Repeat this procedure for all possible k choices of the training partition. Results from all the k runs are averaged.

B) Divide the data into k partitions, and then pick one of the partitions as held-out testing set. Train on all of the $k-1$ partitions and then test on the one that is not used in the training. Repeat this procedure for all possible k choices of the held out partition. Results from all the k runs are averaged.

C) Divide the data into k partitions, and then pick one of the partitions as held-out testing set. Train on all of the $k-1$ partitions and then test on the one that is not used in the training. Repeat this procedure for all possible k choices of the held out partition. The best result from all the k runs is reported.

D) Pick some of the samples as testing data, and then divide the remaining data into k partitions. Train on each one of the k partitions separately and then test on the testing data. Results from all the k runs are averaged.

E) Divide the data into k partitions, and then pick one of the partitions as held-out testing set. Randomly pick a held-out testing set, and then train on all of the $k-1$ partitions that do not overlap with the testing set. Result from this run is reported.

B. Definition of cross-validation refers to textbook P33

27. Consider a regression problem where the two dimensional input points $\mathbf{x} = [x_1, x_2]^T$ are constrained to lie within $[-1, 1]$. The training and test input points $\mathbf{x}$ are sampled uniformly at random within the range. The target outputs y are governed by the following model

$$y \sim \mathcal{N}(\mu, 1)$$

where $\mu = x_1^2 x_2 - 8x_1x_2 + 5x_1^2 + 2x_2 - 3$. In other words, the outputs are normally distributed with mean given by $x_1^2 x_2 - 8x_1x_2 + 5x_1^2 + 2x_2 - 3$ and variance 1.
We learn to predict y given x using polynomial regression models with order from 1 to 9. The performance criterion is the mean squared error. We first train a 1st, 2nd and 9th order model using $n_{train} = 20$ training points, and then test the predictions on a large independently sampled test set ($n_{test} > 1000$).
Select one appropriate model that you would expect to have the lowest error for each column.

	Lowest training error	Highest training error	Lowest test error
1st order	[ ]	[ ]	[ ]
2nd order	[ ]	[ ]	[ ]
9th order	[ ]	[ ]	[ ]

A) 1st order: Lowest test error, 2nd order: Highest training error, 9th order: Lowest training error

B) 1st order: Highest training error, 2nd order: Lowest training error, 9th order: Lowest test error

C) 1st order: Lowest test error, 2nd order: Highest training error, 9th order: Lowest training error

D) 1st order: Highest training error, 2nd order: Lowest test error, 9th order: Lowest training error

D. This question tests the understanding of overfitting and underfitting. Refer textbook P11

28. To find a local minimum of a function using gradient descent, we should take steps proportional to the ( ) of the gradient of the function at the current point. If instead one takes steps proportional to the ( ) of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent.

    A) Negative, Negative
    B) Negative, Positive
    C) Positive, Positive
    D) Positive, Negative

    B. Definition of gradient descent / ascent.

29. Is the following statement true or false? Bayesian linear regression using Gaussian prior is equivalent to maximum likelihood solution with quadratic regularizer.

    A) True
    B) False

    A. Bayesian linear regression using Gaussian prior ends up equivalent formula as maximum likelihood solution with quadratic regularizer.

30. In Bayesian linear regression, what happens to the variance of predictive distribution as the number of data points N the model sees increases?

    A) $\sigma_{N+1}^2(x) \leq \sigma_N^2(x)$
    B) $\sigma_{N+1}^2(x) \geq \sigma_N^2(x)$
    C) The relationship between $\sigma_{N+1}^2(x)$ and $\sigma_N^2(x)$ cannot be decided.

    A. As N increases, the variance σ^2 decreases.