Name ____________________ Student No. ____________________

Final Examination

CSE 474/574: Introduction to Machine Learning (Instructor: Sargur Srihari)

7:15 PM - 10:15 PM Hoch 114

Wednesday, December 14, 2016

No.	**Topic**	**Points**
1	Linear Regression	30
2	Linear Classification	20
3	Logistic Regression	20
4	Neural Networks	30
	Total	100

The total time to complete is 180 minutes. The questions are multiple choice. Some questions have multiple correct answers. Please use the mark-sense sheet to fill-in your answers.

1 Linear Regression

1. In linear regression we use a model of the form $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x})$ where $\mathbf{x} = (x_1, .., x_D)$ is an input vector, $\mathbf{w} = (w_0, .., w_{M-1})$ is a vector of weights, and $\phi = (\phi_0, .., \phi_{M-1})$ is a set of basis functions of the form $\phi_j(\mathbf{x})$. It is called linear regression because we use linear functions of $\mathbf{x}$.

   A) True

   [✓ B) False]

2. In linear regression if we use radial basis functions of the form $\phi_j(\mathbf{x}) = \exp\Big(-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{j}})^T\Sigma^{-1}(\mathbf{x} - \mu_{\mathbf{j}})\Big)$ then the vectors μ_j can be determined from training data using (choose one):

   [✓ A) k-means clustering]

   B) gradient descent

   C) Newton-Raphson

3. Consider formulating linear regression as a probabilistic model where target t is given by a deterministic function $y(\mathbf{x}, \mathbf{w})$ that is corrupted by zero mean noise with precision β. Then we can write $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$.

   [✓ A) True]

   B) False

4. The probabilistic model for linear regression leads to the objective function of maximizing the log-likelihood function over N samples: $\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1})$. This objective is different from the objective of minimizing the sum-of-squares error function $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$.

   A) True

   [✓ B) False]

5. Defining the design matrix $\mathbf{\Phi}$ as consisting of N rows of basis functions corresponding to N samples, the closed form solution to sum-of-squares minimization is $\tilde{\mathbf{w}} = (\mathbf{\Phi^T\Phi})^{-\mathbf{1}}\mathbf{\Phi^T t}$. The reason it is not used often is (choose one):

   A) Its derivatives become constant

   [✓ B) The product $\mathbf{\Phi^T\Phi}$ can become close to singular]

   C) The matrix multiplications are too time consuming

6. In gradient descent we update the weight vector using the gradient of the objective function ∇E. The update equation is(choose one):

   [✓ A) $\mathbf{w}^{(\tau+1)} = \mathbf{w}^{\tau} - \eta\nabla E$.]

   B) $\mathbf{w}^{(\tau+1)} = \mathbf{w}^{\tau} * \eta\nabla E$.

   C) $\mathbf{w}^{(\tau+1)} = \mathbf{w}^{\tau} + \eta\nabla E$.

7. In stochastic gradient descent we update the weight vector in the following way(choose one):

A) add noise to the parameters

[✓ B)] update the parameters using minibatches

C) update the parameters after all training samples are observed

8. Regularized least squares involves(choose one):

   [✓ A)] adding a term $\frac{\lambda}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}$ to the least squares objective $E(\mathbf{w})$.

   B) multiplying the least squares objective $E(\mathbf{w})$ by $\frac{\lambda}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w}$

   C) adding the vector of derivatives of the least squares objective $E(\mathbf{w})$ to its previous value

9. The solution to regularized least squares linear regression has the following property(choose one):

   A) does not have a closed-form solution and therefore requires gradient descent

   [✓ B)] it is a simple extension of the closed-form least squares solution obtained by adding $\lambda\mathbf{I}$ to $(\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi})$ in the least squares solution $\tilde{\mathbf{w}} = (\mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi})^{-\mathbf{1}}\mathbf{\Phi}^{\mathbf{T}}\mathbf{t}$

10. Lasso regularization(choose one):

    [✓ A)] uses a regularization penalty term based on the L^1 norm

    B) recursively uses weight decay regularization

11. Regularization serves the following purpose(choose one):

    [✓ A)] allows complex models to be trained without severe overfitting by limiting model complexity

    B) allows using large training sets

    C) improves performance over the training set

12. In Bayesian linear regression, if we assume a prior distribution over the parameters $\mathbf{w}$ to be Gaussian: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ then this is equivalent to (choose one):

    [✓ A)] adding a term $\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$ to the least squares objective $E(\mathbf{w})$

    B) multiplying the least squares objective $E(\mathbf{w})$ by $\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$

    C) adding the vector of derivatives of the least squares objective $E(\mathbf{w})$ to its previous value scaled by α

13. In Bayesian linear regression, we obtain a predictive distribution as the output. The mean of the predictive distribution can be expressed as a linear combination of the training set target variables t_n in the form $\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n)t_n$. The equivalent kernel $k(\mathbf{x}, \mathbf{x}')$ can be related to Gaussian processes by(choose one):

    A) expressing it as the covariance between $\mathbf{x}$ and $\mathbf{x}'$

    [✓ B)] expressing it as $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^T\psi(\mathbf{x}')$

14. In the *Learning to Rank* problem, although the target values lie in a discrete set $\{0, 1, 2\}$, we choose to perform regression because(choose one):

    A) it does not need a nonlinear output function and therefore it is easier than performing classification

[✓ B)] it provides a continuous value which is useful to rank inputs

15. A limitation of fixed basis functions is (choose one):

    A) The basis functions are nonlinear

    [✓ B)] The number of basis functions that are needed increases exponentially with the number of input dimensions

    C) As the number of samples increase the basis functions diverge

2 Linear Classification

16. A probabilistic classifier is designed to classify an input $\mathbf{x}$ into one of K classes $C_k, k = 1.., K$. It uses the model $p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_{k=1}^{K} p(\mathbf{x}|C_k)p(C_k)}$. Such a classifier is best referred to as a: (choose one)

    A) Mixture model

    [✓ B)] Generative model

    C) Discriminative model

17. In the model described above, if we choose $p(\mathbf{x}|C_k)$ to be multivariate Gaussian with a common covariance matrix, is it possible for the boundary between two classes to be parabolic?

    A) Yes

    [✓ B)] No

18. In a two-class probabilistic classifier is it necessary for $p(\mathbf{x}|C_1) + p(\mathbf{x}|C_2) = 1$?

    A) Yes

    [✓ B)] No

19. A two-class linear classifier has a decision boundary in D-dimensional input space of the form $\mathbf{w}^T\mathbf{x} + w_0 = 0$. The bias term w_0 can be eliminated by (choose one):

    A) Moving it to the right-hand side of the equation

    B) Dividing both terms of the summation by w_0

    [✓ C)] Adding an additional input whose value is set to 1

20. The main advantage of eliminating the bias term w_0 is(choose one):

    A) We only need to learn the weights associated with D inputs

    [✓ B)] The bias term can be determined using the same algorithm as the rest of the weights

    C) Derivative of w_0 is undetermined

21. Fisher's linear discriminant is a method for classification that operates by (choose one):

    A) finding the best linear decision surface between training data in input space

[✓ B)] projecting data to a one-dimensional space that provides the best separation of classes

C) multiplying learning data by the Fisher criterion

22. Where was the first machine learning system known as the Perceptron invented (choose one)?

A) San Jose, CA

B) Cambridge, MA

C) Cambridge, UK

[✓ D)] Buffalo, NY

23. The perceptron algorithm learns by (choose one):

A) using a method based on adding perceived weights

[✓ B)] adding or subtracting at each step a scaled version of the input to the weight

C) using a closed-form solution

24. When training data are not linearly separable the perceptron learning algorithm (choose one):

[✓ A)] does not stop

B) determines the best linear approximation

C) ends up in a local minimum

25. When the input data is a vector of binary values that are statistically independent then the optimal solution is (choose one):

A) a quadratic function of the input vector

[✓ B)] a linear function of the input vector

C) to determine whether the bits sum to a value greater than a threshold

3 Logistic Regression

26. A probabilistic classifier is designed to classify an input $\mathbf{x}$ into one of 2 classes $C_k, k = 1, 2$. It uses the model $P(C_1|\mathbf{x}) = y = \sigma(\mathbf{w}^T\mathbf{x})$ where $\mathbf{w}$ is a set of weights to be learnt. Such a classifier is best referred to as a: (choose one)

A) Mixture model

B) Generative model

[✓ C)] Discriminative model

27. Is it possible to use logistic regression and obtain a non-linear decision boundary in input space? (choose one):

[✓ A)] yes, by using basis functions

B) no

C) yes, it directly creates a non-linear decision boundary in input space

28. Both logistic regression and least squares classification obtain linear solutions. How are the results different? (choose one):

[✓ A)] Logistic regression is more robust (less sensitive to noise)

B) Least squares is more robust (less sensitive to noise)

C) The results are exactly the same

29. Given a training set of samples $\{\mathbf{x}_n, t_n\}, n = 1, .., N$, the two-class log-likelihood function based on the Bernoulli distribution leads us to minimize the following objective function, where $y_n = \sigma(\mathbf{w}^T\mathbf{x}_n)$ (choose one):

A) $E_S = -\sum_{n=1}^{N}(y_n - t_n)^2$

[✓ B)] $E_C = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$

30. The first and second derivatives of the objective function $E(\mathbf{w})$ are known as the gradient $\nabla E(\mathbf{w})$ and Hessian $H = \nabla\nabla E(\mathbf{w})$ respectively. They are useful as follows(choose one):

A) both are used in the weight vector update equation

[✓ B)] gradient is used to update the weight vector, Hessian is used to determine solution robustness

C) Hessian is used to update the weight vector when the gradient vanishes

For the remainder of the questions in Section 3, consider a classification problem where we are building a logistic regression classifier. The task is to determine the probability of breast cancer in a patient. The input data has two variables, *mass*: x_1 and *distortion*: x_2. The target is binary t (0-no cancer and 1-cancer). Consider the following training data set:

x_1	x_2	t
2	0.5	1
1	1.5	0

Consider the following weights for the logistic regression output $y = \sigma(\mathbf{w}^T\mathbf{x})$: $w_0 = 1$, $w_1 = 3$, $w_2 = -2$, in which w_0 is the bias term. Also, $\sigma'(a) = \sigma(a)(1 - \sigma(a))$:

31. What is the output of logistic regression y? (choose one):

[✓ A)] $\sigma(6)$

B) $\sigma(1)$

C) $\sigma(2)$

32. What is the Sum-of-Squared Error E_S? (choose one):

[✓ A)] $\frac{1}{2}\left[(\sigma(6)-1)^2+\sigma^2(1)\right]$

B) $\frac{1}{2}\left[(\sigma(1)-1)^2+\sigma^2(6)\right]$

C) $\frac{1}{2}\left[(\sigma(6)-2)^2+\sigma^2(1)\right]$

D) $\frac{1}{2}\left[(\sigma(1)-2)^2+\sigma^2(6)\right]$

33. What is the Cross-entropy Error E_C? (choose one):

[✓ A)] $-\log\left[\sigma(6)(1-\sigma(1))\right]$

B) $-\log\left[\sigma(1)(1-\sigma(6))\right]$

C) $\log\left[\sigma(6)(1-\sigma(1))\right]$

D) $\log\left[\sigma(1)(1-\sigma(6))\right]$

34. What is $\frac{\partial E_S}{\partial w_1}$? (choose one):

[✓ A)] $2\sigma^2(6)(1-\sigma(6))+\sigma^2(1)(1-\sigma(1))$

B) $\sigma^2(6)(1-\sigma(6))+2\sigma^2(1)(1-\sigma(1))$

C) $2\sigma^2(6)(1-\sigma(1))+\sigma^2(1)(1-\sigma(6))$

D) $\sigma^2(6)(1-\sigma(1))+2\sigma^2(1)(1-\sigma(6))$

35. What is $\frac{\partial E_C}{\partial w_1}$ (choose one):

[✓ A)] $2\sigma(6)+\sigma(1)-2$

B) $\sigma(6)+2\sigma(1)-2$

C) $2\sigma(6)+\sigma(1)-3$

D) $\sigma(6)+2\sigma(1)-3$

4 Neural Networks

36. A neural network uses a nonlinear activation function. To obtain a probability as output, the following function can be used. (choose one)

A) ReLu

[✓ B)] sigmoid

C) tanh

D) Gaussian

37. A neural network classifier has K outputs corresponding to K classes. We wish to interpret the outputs as probabilities that sum to one. We can achieve this by using (choose one)

    A) ReLu

    B) sigmoid

    [✓ C)] sotmax

    D) tanh

38. Early stopping minimizes the error rate on the training set (choose one)

    A) True

    [✓ B)] False

39. Tangent propagation is a method of(choose one)

    [✓ A)] regularization

    B) finding gradients of the objective function

40. A convolutional neural network is a special kind of neural network– one in which a kernel is slid over the input. Any CNN could be converted into an equivalent neural network. Consider a simple case in which we use a one dimensional kernel. Suppose the input layer $\mathbf{x}_0$ is a vector with 6 inputs $[x_0^0, .., x_0^5]^\top$. Suppose the convolutional layer uses a kernel (filter) which is the 3-element one dimensional vector $[a, b, c]^\top$. The second layer $\mathbf{x}_1$ is a $6 - 3 + 1 = 4$ node vector $[x_1^0, .., x_1^3]^\top$. If we want to convert this convolutional layer into a regular neural network layer that performs the computation $\mathbf{x}_1 = W_1 \mathbf{x}_0$, we can determine the weight matrix W_1 as given below. (choose one)

    A) $W_1 = \begin{bmatrix} a & a & a & & & \\ & b & b & b & & \\ & & c & c & c & \\ & & & 1 & 1 & 1 \end{bmatrix}$

    [✓ B)] $W_1 = \begin{bmatrix} a & b & c & & & \\ & a & b & c & & \\ & & a & b & c & \\ & & & a & b & c \end{bmatrix}$

    C) $W_1 = \begin{bmatrix} a & b & c & & & \\ & & & a & b & c \\ a & b & c & & & \\ & & & a & b & c \end{bmatrix}$

    D) $W_1 = \begin{bmatrix} a+b+c & a+b+c & a+b+c & & & \\ & a+b+c & a+b+c & a+b+c & & \\ & & a+b+c & a+b+c & a+b+c & \\ & & & a+b+c & a+b+c & a+b+c \end{bmatrix}$

41. Continue with the previous question. Suppose a subsampling layer $\mathbf{x}_2$ follows the convolution layer and it performs averaging in an area of two contiguous nodes resulting in a vector $\mathbf{x}_2 = [x_2^0, x_2^1]^\top$. Please convert this subsampling layer into a regular neural network layer that computes $\mathbf{x}_2 = W_2 \mathbf{x}_1$ where the weight matrix W_2 can be determined using: (choose one)

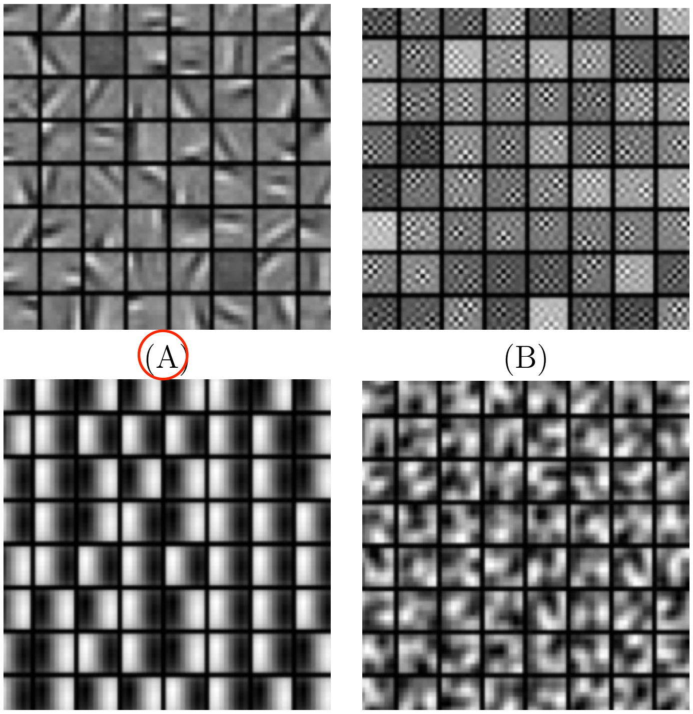
A) $W_2 = \begin{bmatrix} 1 & 1 & & \\ & & 1 & 1 \end{bmatrix}$

B) $W_2 = \begin{bmatrix} 1 & & 1 & \\ & 1 & & 1 \end{bmatrix}$

[✓ C)] $W_2 = \begin{bmatrix} 0.5 & 0.5 & & \\ & & 0.5 & 0.5 \end{bmatrix}$

D) $W_2 = \begin{bmatrix} 0.5 & & 0.5 & \\ & 0.5 & & 0.5 \end{bmatrix}$

42. Suppose we use convolutional neural network to classify digit images. Suppose we visualize the filters learned. Which of the following in Figure 1 do you think is the desired set of filters.



[✓ (A)] (B)

(C) (D)

Figure 1: CNN filters for question 42.

43. What is the computational complexity of the backpropagation method of determining derivatives when W is the number of of weights and biases? (choose one)

[✓ A)] $O(W)$

B) $O(W^{1.5})$

C) $O(W^2)$

D) $O(W^3)$

For questions numbered 44, 45 and 46 consider the simple neural network with only one output node in Figure 2. Ignore the bias node for this example. The values on the edges indicate the weights associated with the "receiving node".

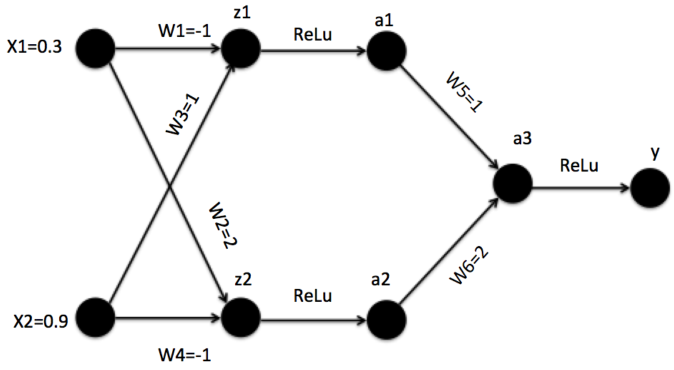


Figure 2: Neural Networks

Assume that the neurons in the hidden layer and output layer have a ReLU activation function defined as:

$$f(x) = \max(0, x)$$

Assume that we use the *sum-of-squared errors* as the objective function.

44. Performing a forward pass on the network in Figure above, we get for the output y a value of: (choose one)

    [✓ A)] 0.6

    B) 0.5

    C) 0.4

    D) 0.3

45. Performing a backward pass on the network with t = 0.5., the derivatives $\frac{\partial E}{\partial z_1}$, and $\frac{\partial E}{\partial z_2}$ are: (choose one)

    A) 0.6, 0

    B) 0, 0.6

    [✓ C)] 0.3, 0

    D) 0, 0.3

46. Update the weights at the output and hidden nodes using learning rate η =1. $\frac{\partial E}{\partial w_1}$, $\frac{\partial E}{\partial w_2}$, $\frac{\partial E}{\partial w_3}$, $\frac{\partial E}{\partial w_4}$ (choose one)

    [✓ A)] 0.18, 0, 0.54, 0

    B) 0.09, 0, 0.27, 0

    C) 0, 0.18, 0, 0.54

    D) 0, 0.09, 0, 0.27

47. What is meant by neural network regularization? (choose one)

    A) making it congruent to a standard neural network

    [✓ B)] improving its ability to generalize to previously unseen data

    C) multiplying the objective function by a hyper-parameter

48. A deep neural network has the advantage that (choose one)

    [✓ A)] the representation of the input is learnt automatically

    B) it needs fewer parameters than a shallow neural network

49. Which of the following statements about Tensorflow are correct:

    [✓ A)] A TensorFlow computation graph is a description of computations.

    [✓ B)] A Session places the graph ops onto Devices, such as CPUs or GPUs, and provides methods to execute them.

    C) If we use TensorFlow to implement a neural network, the forward propagation will be computed during the computation graph construction phase.

    D) We use a tf.Variable object to represent the activations of a neural network.

    E) A tf.placeholder object is used to store the parameters of the neural network.

50. When implementing neural networks, we prefer to “vectorize” the code, which means substiting loops with matrix manipulations to make the code faster. Which of the following loops could be vectorized into matrix manipulations?

    [✓ A)] Loops through data points

    [✓ B)] Loops through elements in a vector

    [✓ C)] Loops through filters in the same layer in a CNN

    [✓ D)] Loops caused by sliding the filter on an input layer to perform convolutions