Midterm Sample Questions

#1 Based on Linear algebra method, what will be the predicted values of the new square foot houses prices. Describe the output matrix.

$$X = \begin{bmatrix} 2104 \\ 1416 \\ 1532 \\ 832 \end{bmatrix}$$

(I) $h_\theta(x) = -60 + 0.25x$

(II) $h_\theta(x) = 100 + 0.1x$

(III) $h_\theta(x) = -120 + 0.4x$

#2 The following algorithm is for one feature univarite Linear regression?

repeat until Convergence {

$$\theta_0 \leftarrow \theta_0 - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)\right)$$

$$\theta_1 \leftarrow \theta_1 - \alpha\left(\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{i}\right)$$

}

(for, θ_0, θ_1)

Write the updates for multi-features multivariate Linear regression?

OR

Modify the above algorithm for the multivariate Linear Regression.

Q:3 The following algorithm is for stochastic Gradient descent: Explain each line of this algorithm (numbered 1, 2, ...). Explain how does it work and the purpose of each line in the algorithm?

① Initialize parameters $(\theta_0 = \theta_1 = 0)$

② Randomly shuffle the data set

③ repeat K times $(K \in [1, 10])$ {

④ for $i = 1, 2, 3 \cdots m$ {

⑤ $\theta_j \leftarrow \theta_j - \alpha \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$

for $j = 0, 1 \cdots n$

$x_0^{(i)} = 1$

}

}

* why do we shuffle data.
* why There is no summation here.

4) Explain the Logistic Regression Decision boundary: describe the boundary based on sigmoid function?
how do you read it, Explain.

$$h_\theta(x) = P(y=1 \mid x, \theta)$$

5) The below cost function is for logistic Regression classification for one Example.

$$J(\theta) = -y \log(h_\theta(x)) - (1-y) \log(1-h_\theta(x))$$

Describe how does it work for Right prediction and wrong prediction.

#6 prove $g'(z) = g(z)(1 - g(z))$

#7 The below $\frac{\partial J}{\partial \theta_j}$ term is identical for both Linear Regression and Logistic Regression —

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

But There is a difference, what is That? Explain briefly?

#8 write the Discriminant function of Logistic Regression? how does this Discriminant function work as a decision boundary.

#9 Describe what modification need to do for regularization to the below Linear regression (gradient descent) algorith

① Initialize $\theta_0 = \theta_1 = \cdots \theta_n$.

② repeate {

$$\theta_0 \leftarrow \theta_0 - \alpha \left(\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_5^{(i)} \right)$$

$$\theta_j \leftarrow \theta_j \leftarrow \left(\alpha/m \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_5^{(i)} \right)$$

}

Write the final θ_j.

#10 prove the Bayes rule based on Conditional probability. write the name of each term in the equation.

#11) write the decesion boundary based on Bayesian rule? Explain how does it work.

Midterm review
Sample questions

12. Describe supervised learning and unsupervised learning with examples. Identify each of the below examples?

    Some of the problems below are best addressed by a supervised learning algorithm and others with an unsupervised learning algorithm. For each problem below, identify whether it is a supervised or unsupervised learning problem.

    - Given historical data of children's ages and heights, predict the children's height as a function of their age.
    - Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of patients for which we might tailor separate treatments.
    - Have a computer examine an audio clip of a piece of music and classify whether or not there are vocals (i.e. a human voice singing) in that audio clip or if it is a clip of just musical instruments (i.e. no vocals).
    - Given data on how 1000 medical patients respond to an experimental drug (such as effective of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug and if so, what these categories are.

13. What is the purpose of a training dataset in supervised learning?

14. Suppose you are working on weather prediction and you would like to predict whether it will be raining at 5 p.m. the next day. What category of supervised learning would this belong to?

15. Supervised learning can be split up primarily into two categories. What are they and what is the purpose of each category? Explain with examples

16. Define linear regression and classification for one variable? Explain with examples.

17. What is the hypothesis/Model in machine learning?

18. What is the Goal of linear regression using one variable?

19. Draw a block diagram of supervised learning model/hypothesis?

20. What is the prediction model used in univariate linear regression? How would you use this parameters to perform predictions?

21. State the cost function $J(\theta_0, \theta_1)$ for univariate linear regression. Why do we square the differences between the predicted outputs and true outputs rather than, say, taking the absolute value?

22. How do we choose the parameters $\theta_1 \; and \; \theta_2$ to find the line of best fit? What is simultaneous update in linear regression? What will happen if we don't do follow simultaneous update rule?

23. State the cost function for univariate linear regression.

24. What do we need to do with the cost function in order to retrieve the correct parameters for linear regression?

25. Cost 0, what does it mean?

26. Why do we keep bias term always separate?

27. The following algorithm is the Gradient descent algorithm. Explain how it works?

28. Explain the gradient term of gradient descent ($\frac{\partial}{\partial \theta_j} J(\theta_1)$)? How does it minimize the cost function?

29. Does it matter how we initialize the parameters for gradient descent? What effect does the initialization have in gradient descent?

30. What is the purpose of the learning rate? What happens if this value is too small? What happens if this value is too large?

31. What means "Batch gradient" descent? What means "stochastic gradient" or "on-line" gradient descent?
32. Why Linear Regression has a single minimum
33. Why is matrix vector multiplication useful?
34. Do you believe that using just one feature would allow regression to be accurate for real-world problems? Why or why not
35. Specify the elements in the below table?
36. In the first class, using multiple features may help in increasing the accuracy of our prediction model. Suppose we have the following table that illustrates our training dataset.

Size (feet^2)	Number of bedrooms	Number of floors	Age of house (years)	Price ($1000s)
2014	5	1	45	460
1416	3	2	40	232
1534	2	2	30	315
832	1	1	36	178

The first four columns represent different features x_1, x_2, x_3, x_4 where the last column is the observed price of a house y. Therefore, each training example is occupied by one row where the first four columns are features and the last column is the expected output. From class, we used the notation $x_j^{(i)}$ which describes the j^{th} feature for the i^{th} training example in our dataset and $y^{(i)}$ is the expected output for the i^{th} training example in our dataset. To ensure that you have this notation correct in your head, identify the following quantities from the table:

- $x_2^{(3)}, x_1^{(4)}, x_4^{(2)}, x_3^{(1)}$
- $y^{(2)}, y^{(4)}$

37. Write the hypothesis of multivariate linear regression

38. Write Gradient Descent By linear algebra

39 There is a condition that $\theta_0$must have when computing the parameter update that no other parameter has. What would this condition be?

40. Usually in practice, with multiple features come different dynamic ranges for each feature. How does this affect the gradient descent algorithm?

41. Define Zero Mean Unit Variance

42. What would you do to your training data set before you found the model parameters with gradient descent? What benefits would this have if you decide to do this before finding your model parameters?

43. How do you choose learning rate?

44. How do we confirm that Gradient descent is working properly?

45. What is expected to happen with the cost function as the number of iterations increase when using gradient descent to find the model parameters?

46. What does it mean if the cost function increases at each iteration?

47. Is there a strategy you can use to select the right learning rate to ensure that the gradient descent algorithm converges?

48. Describe Polynomial Regression and where do we use polynomial Regression?

49. What potential problems would polynomial regression have when you include more features to accommodate this task? Do you need to do something to these new features to allow for convergence?

50. As seen in class, if the dataset is huge, using the current definition of gradient descent will be unsuitable as you have to sum over every training example before computing an update. Is there another way we can compute an update that is faster

51. Write the algorithm for batch gradient descent and stochastic gradient descent

52. How do we check for convergence (stochastic gradient descent?)

53. When approaching the parameters that minimize the overall cost, what are the differences between stochastic gradient descent and batch gradient descent? Describe this in terms of the contour plots seen in class.

54. How does Classification works
55. What is the goal of classification?

56. What is the difference between classification and logistic regression?

57.

What is the hypothesis $h_\theta(x)$ used in logistic regression?

58. We need to set a threshold on where we denote the transition between positive class and negative. Or describe the decision boundaries

59 How does logistic regression work where our predicted output is continuous?

60 How do you interpret the output of the hypothesis $h_\theta(x)$?

61

What is the definition of the likelihood $L(\theta)$ for the parameters θ of our classification prediction model?

62

Why do we take the log of the likelihood (i.e. $l(\theta)$) instead of considering it normally?

63

How do we convert the likelihood into a cost function $J(\theta)$?

64

Why is it generally not recommended to use linear regression for classification purposes?

65

What is a decision boundary? Provide the definition with regards to the sigmoid hypothesis.

66 Describe one –VS- all algorithm

The theory covered in classification assumes the binary classification case, or classifying an input into on of two categories. How do we perform multi-class classification?

67. What is the problem of over fitting?

68. What is the technique called to reduce over fitting? What do we do to the cost function to facilitate this?

69 We do regularization by introducing an additional constant into $J(\theta)$. For linear regression the cost function with regularization is below

$$J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=0}^{m} \theta_j^2]$$

Is this correct? If not, explain why not,

70 How does λ affect our model and Explain it with by taking $different\ \lambda$ value (Such as 0, 10, 100)

71. How do you convert the likelyhood function to a cost function?

72. For two probabilistic events and , state the conditional probability of event given that event has occurred in terms of Bayes Rule.

73 What does the prior probability $P(\omega_j)$symbolize?

74
What does the class-conditional or likelihood probability $\boldsymbol{P(x|\omega_j)}$ symbolize? What is the underlying assumption we are taking in this course in terms of how the features $\boldsymbol{x}$ are ***distributed***?

75 State what the class-conditional or likelihood probability would be assuming our features are Gaussian distributed.

76 State the Bayes decision Rule for binary classification. Do this in terms of the likelihood and class-conditional probabilities.