Sample questions for “Fundamentals of Machine Learning 2018”

Teacher: Mohammad Emtiyaz Khan

A few important informations:

- In the final exam, no electronic devices are allowed except a calculator. Make sure that your calculator is only a calculator and cannot be used for any other purpose.
- No documents allowed apart from one A4 sheet of your own notes.
- You are not allowed to talk to others
- For derivations, clearly explain your derivation step by step. In the final exam you will be marked for steps as well as for the end result.
- For multiple-choice questions, you also need to provide explanations. You will be marked for your answer as well as for your explanations.
- We will denote the output data vector by $\mathbf{y}$ which is a vector that contains all y_n, and the feature matrix by $\mathbf{X}$ which is a matrix containing features $\mathbf{x}_n^T$ as rows. Also, $\widetilde{\mathbf{x}}_n = [1, \mathbf{x}_n^T]^T$.
- N denotes the number of data points and D denotes the dimensionality.

1 Multiple-Choice/Numerical Questions

1. Choose the options that are correct regarding machine learning (ML) and artificial intelligence (AI),

   (A) ML is an alternate way of programming intelligent machines.

   (B) ML and AI have very different goals.

   (C) ML is a set of techniques that turns a dataset into a software.

   (D) AI is a software that can emulate the human mind.

   Answer: (A), (C), (D)

2. Which of the following sentence is FALSE regarding regression?

   (A) It relates inputs to outputs.

   (B) It is used for prediction.

   (C) It may be used for interpretation.

   (D) It discovers causal relationships.

Answer: (D)

3. What is the rank of the following matrix?

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{1}$$

Answer: 1

4. What is the dimensionality of the null space of the following matrix?

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{2}$$

Answer: 2

5. What is the dimensionality of the null space of the following matrix?

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & -9 \\ -6 & -4 & 18 \\ 12 & 8 & -36 \end{bmatrix} \tag{3}$$

Answer: 2

6. For the one-parameter model, mean-Square error (MSE) is defined as follows: $\frac{1}{2N}\sum_{n=1}^{N}(y_n - \beta_0)^2$. We have a half term in the front because,

(A) scaling MSE by half makes gradient descent converge faster.

(B) presence of half makes it easy to do grid search.

(C) it does not matter whether half is there or not.

(D) none of the above

Answer: C

7. Grid search is,

(A) Linear in D.

(B) Polynomial in D.

(C) Exponential in D.

(D) Linear in N.

Answer: C,D

8. The advantage of Grid search is (are),

(A) It can be applied to non-differentiable functions.

(B) It can be applied to non-continuous functions.

(C) It is easy to implement.

(D) It runs reasonably fast for multiple linear regression.

Answer: A,B,C.

9. Gradient of a continuous and differentiable function

   (A) is zero at a minimum

   (B) is non-zero at a maximum

   (C) is zero at a saddle point

   (D) decreases as you get closer to the minimum

   Answer: A,C,D

10. Consider a linear-regression model with $N = 3$ and $D = 1$ with input-ouput pairs as follows: $y_1 = 22$, $x_1 = 1$, $y_2 = 3$, $x_2 = 1$, $y_3 = 3$, $x_3 = 2$. What is the gradient of mean-square error (MSE) with respect to β_1 when $\beta_0 = 0$ and $\beta_1 = 1$? Give your answer correct to two decimal digits.

    Answer: -1.66 (deviation 0.01)

11. Let us say that we have computed the gradient of our cost function and stored it in a vector $\mathbf{g}$. What is the cost of one gradient descent update given the gradient?

    (A) $O(D)$

    (B) $O(N)$

    (C) $O(ND)$

    (D) $O(ND^2)$

    Answer: (A)

12. Let us say that we are fitting one-parameter model to the data, i.e. $y_n \approx \beta_0$. The average of $y_1, y_2, \ldots, y_N$ is 1. We start gradient descent at $\beta_0^{(0)} = 0$ and set the step-size to 0.5. What is the value of β_0 after 3 iterations, i.e., the value of $\beta_0^{(3)}$?

    Answer: 0.875 (deviation 0.01)

13. Let us say that we are fitting one-parameter model to the data, i.e. $y_n \approx \beta_0$. The average of $y_1, y_2, \ldots, y_N$ is 1. We start gradient descent at $\beta_0^{(0)} = 10$ and set the step-size to 0.5. What is the value of β_0 after 3 iterations, i.e., the value of $\beta_0^{(3)}$?

    Answer: CA: 2.125 (deviation 0.01)

14. Computational complexity of Gradient descent is,

    (A) linear in D

    (B) linear in N

    (C) polynomial in D

    (D) dependent on the number of iterations

    Answer: C

15. Generalization error measures how well an algorithm perform on unseen data. The test error obtained using cross-validation is an estimate of the generalization error. Is this estimate unbiased?

    Answer: (No)

16. K-fold cross-validation is

    (A) linear in K

    (B) quadratic in K

    (C) cubic in K

    (D) exponential in K

    Answer: A

17. You observe the following while fitting a linear regression to the data: As you increase the amount of training data, the test error decreases and the training error increases. The train error is quite low (almost what you expect it to), while the test error is much higher than the train error.

    What do you think is the main reason behind this behavior. Choose the most probable option.

    (A) High variance

    (B) High model bias

    (C) High estimation bias

    (D) None of the above

    Answer: A

18. Adding more basis functions in a linear model... (pick the most probably option)

    (A) Decreases model bias

    (B) Decreases estimation bias

    (C) Decreases variance

(D) Doesn't affect bias and variance

Answer: A

2 Multiple-output regression

Suppose we have N regression training-pairs, but instead of one output for each input vector $\mathbf{x}_n \in \mathbb{R}^D$, we now have 2 outputs $\mathbf{y}_n = [y_{n1}, y_{n2}]$ where each y_{n1} and y_{n2} are real numbers. For each output y_{n1}, we wish to fit a separate linear model:

$$y_{n1} \approx f_1(\mathbf{x}_n) = \beta_{10} + \beta_{11}x_{n1} + \beta_{12}x_{n2} + \ldots + \beta_{1D}x_{nD} = \boldsymbol{\beta}_1^T\widetilde{\mathbf{x}}_n \tag{4}$$

$$y_{n2} \approx f_2(\mathbf{x}_n) = \beta_{20} + \beta_{21}x_{n1} + \beta_{22}x_{n2} + \ldots + \beta_{2D}x_{nD} = \boldsymbol{\beta}_2^T\widetilde{\mathbf{x}}_n \tag{5}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of β_{1d} and β_{2d} respectively, for $d = 0, 1, 2, \ldots, D$, and $\widetilde{\mathbf{x}}_n^T = [1\, \mathbf{x}_n^T]$.

Our goal is to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for which we choose to minimize the following cost function:

$$\mathcal{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) := \sum_{n=1}^{N} \left[\frac{1}{2}\left(y_{n1} - \boldsymbol{\beta}_1^T\widetilde{\mathbf{x}}_n\right)^2 + \frac{1}{2}\left(y_{n2} - \boldsymbol{\beta}_2^T\widetilde{\mathbf{x}}_n\right)^2\right] + \lambda_1\sum_{d=0}^{D}\beta_{1d}^2 + \lambda_2\sum_{d=0}^{D}\beta_{2d}^2. \tag{6}$$

(A) Derive the gradient of $\mathcal{L}$ with respect to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

(B) Suppose $N = 20$ and $D = 15$. Do we need to regularize? Explain your answer.

(C) Suppose we increase the number of data points from $N = 20$ to $N = 200$. Should we decrease the value of λ_1 and λ_2? Explain your answer.

(D) What is the computation complexity with respect to N and D?

Answer:

(A) $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}_1} := -\sum_{n=1}^{N}\left[\left(y_{n1} - \boldsymbol{\beta}_1^T\widetilde{\mathbf{x}}_n\right)^2\widetilde{\mathbf{x}}_n\right] + \lambda_1\boldsymbol{\beta}_1$, same for $\boldsymbol{\beta}_2$.

(B) The number of parameters is equal to 30 and the number of data points is equal to 40. It is good to regularize, but just a mild regularization will do since the number of parameters is still less than number of data points.

(C) Yes, we expect this to be the case because, if the data points are i.i.d., then we might need less regularization.

(D) Same as gradient descent (please put an exact number here for the final exam).

3 Eigenvalues

Given a real-valued matrix $\mathbf{X}$, show that all the non-zero eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ are the same.

Answer: To prove this, you can use the SVD of $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Then $\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{S}^2\mathbf{U}^T$ and $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^2\mathbf{V}$. The non-zero eigenvalues are the same, although the number of eigenvalues are different.

4 Artificial Neural Networks

Consider the following artificial neural network with the nonlinear transformation $z_{nm} = \sigma(a_{nm})$ (see figure below). Here, n is the data index and m is the index of hidden units. There are two binary outputs y_{n1} and y_{n2} taking values in $\{0, 1\}$.

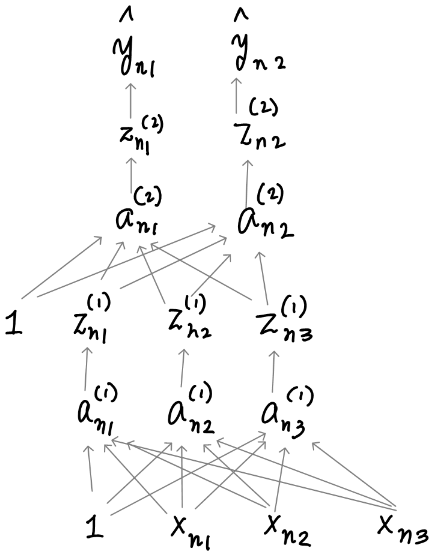

Figure 1: Artificial neural network

Suppose you have $N = 200$ data points but $M = 200$ hidden units for each layer. What problem(s) are you likely to encounter when training such a network? How would you solve the problem(s)?

Answer: Overfitting. There are multiple ways to tackle this problem as discussed in the lecture.