AP Machine Learning C Exam*

December 23, 2022

Name (Individual Copy): ______________________________

*C because this version of APML uses calculus.

Multiple Choice Questions (MCQ), 70% of grade. Please bubble in answers correctly.

Note: These questions are difficult. That was intentional.

1 Fundamentals

1. Which of the following algorithms is used to train all neural networks in updating their parameters?
   - ◯ Gradient Descent
   - ◯ Bayesian Optimization
   - ◯ Overfitting
   - ◯ Regularization
2. Gradient Descent employs which concept from calculus in its optimization for *multivariate* models.
   - ◯ Derivative
   - ◯ Partial Derivative
   - ◯ Taylor Series
   - ◯ Maclaurin Series
3. Which rule best exemplifies the fact that neural networks and regression models are really just strings of functions?
   - ◯ Product rule, because neural network's multiply weights and inputs together
   - ◯ Chain rule, because a neural network is a composite function
   - ◯ Sum rule, because matrix multiplication internal to a neural network involves summations
   - ◯ Quotient rule, because you divide the learning rate at each iteration
4. Exponential and polynomial regressions outperform linear regression when the data is nonlinear because...
   - ◯ The model learned by linear regression insufficiently captures nonlinear relationships
   - ◯ The linear model ends up learning a nonlinear relationship in such a case
   - ◯ Exponential and polynomial regressions are more computationally efficient
   - ◯ The derivative of linear models at any point will be greater than in nonlinear models
5. Why is it empirically inefficient to use a linear ANN over linear regression?
   - ◯ The linear ANN will be significantly more computationally expensive
   - ◯ The linear ANN learns the same relationship as that of the linear regression
   - ◯ The many functional compositions are unable to learn linear relationships
   - ◯ Both A and B
6. Which of the following correctly denotes the usage of exploratory data analysis (EDA)?
   - ◯ EDA is used as a means to gather more data
   - ◯ EDA is a method of statistical stratification
   - ◯ EDA is used to transform and reduce the complexity of the model
   - ◯ EDA is used to clean elements of the data to ensure the model's success
7. Which of the following are valid reasons for why the log loss is used in logistic models (*May be multiple correct answers*)?

- ◯ The log makes the function more concave
- ◯ Makes the objective function optimization convex
- ◯ The log makes the function more continous
- ◯ The function constraints the evaluation to only labels in a binary spectrum

8. Which of the following is the correct function for the logarithm loss function?
   - ◯ $\Sigma x P(x)$
   - ◯ $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
   - ◯ $\sum_{i=1}^{N}(y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i))$
   - ◯ $\sum_{i=1}^{N} y_i log(\hat{y}_i)$

9. Which of the following proofs are required to conclude that the optimization will be done on a convex function ($J(\theta)$ is the objective function)?
   - ◯ $\frac{\partial J}{\partial \theta} > 0$
   - ◯ $\frac{\partial J}{\partial \theta} \geq 0$
   - ◯ $\frac{\partial^2 J}{\partial \theta^2} \geq 0$
   - ◯ $\frac{\partial^2 J}{\partial \theta^2} > 0$

10. Linear regression...
    - ◯ is strictly convex in optimization
    - ◯ can be graphically visualized as approximating curves
    - ◯ is strictly non-convex in optimization
    - ◯ has a second-order partial derivative of the loss which is strictly negative

11. Artificial Neural Networks can be represented as which of the following (choose 2)?
    - ◯ Composite Functions
    - ◯ Integral of the loss function over optimal parameter values
    - ◯ Differentiable Functions
    - ◯ Sub-differentiable Function

12. Mean squared error (MSE) and log loss differ because...
    - ◯ MSE is the integral of the model's cost, whereas log loss is the log of the integral
    - ◯ Log loss utilizes a log operation whereas MSE utilizes Euler's number e
    - ◯ MSE is used in regression, whereas log loss is used in classification
    - ◯ MSE and log loss do not differ

2 Model Questions

13. All of the following model architectures EXCEPT which one will require logarithm loss?
    - ◯ Neural Network for Binary Prediction
    - ◯ Logistic Regression
    - ◯ Linear Neural Network
    - ◯ Autoencoder

14. What is Marc's favorite data set?
    - ◯ Anish Lakkapragada Feet Pics 2019-20
    - ◯ ImageNet
    - ◯ Wisconsin Breast Cancer Dataset
    - ◯ Vehicle Accidents in U.S. 2016-2021
15. Which of the following architectures is NOT for sequential data?
    - ◯ Recurrent Neural Networks
    - ◯ Convolutional Neural Networks
    - ◯ LSTMs
    - ◯ GRUs
16. Which of the following correctly indicates the probability distributions approximated by discriminative and generative models?
    - ◯ Whereas discriminative models approximate P(X), generative models approximate P(y | X)
    - ◯ Whereas discriminative models approximate P(y | X), generative models approximate P(X)
    - ◯ Whereas discriminative models approximate P(X | y), generative models approximate P(y)
    - ◯ Both models use the input space X, meaning neither approximate probability distributions
17. Which of the following would describe a problem statement requiring the use of recurrent models?
    - ◯ Classifying images as either cat or dog
    - ◯ A tabular dataset with $N > 10^6$
    - ◯ A dataset whose applied model takes the output from previous layers to help construct an output for its current layer
    - ◯ A model translating English input sentences into French
18. Which of the following correctly explains why neural loss landscapes (with 1 or more hidden layers) have local minima?
    - ◯ Neural networks are forced to use cross-entropy loss, introducing many local optima
    - ◯ The usage of many non-linear activation functions in a neural network yields non-convex optimization
    - ◯ The significant number of matrix operations accompanying a larger neural network yield non-convex optimization
    - ◯ The usage of any non-linear activation function, such as sigmoid or ReLU, entail a non-convex optimization problem
19. In what way are the discriminator and generator different?
    - ◯ The discriminator is a discriminative model, whereas the generator is a Bayesian model
    - ◯ The generator takes in the input data, the discriminator takes in Gaussian sampled random noise
    - ◯ The generator generates examples, and the discriminator detects whether its examples are real or fake
    - ◯ The discriminator is a GAN used for tabular cases, the generator- image cases.
20. The convolution operation...
    - ◯ is a matrix multiplication on the input image
    - ◯ is the result of a single scalar dot product between the filter and input

- ◯ is the integral of the multiplicative product between the filter and input
- ◯ Is the result of many strided dot products

Free Response Section (FRQ), 30% of grade. Use the space provided below. If it is not enough, take an L.

The questions below are also difficult. Some of them we don't know how to solve either (kidding of course). They increase in difficulty as you go down the list.

1. For an artificial neural network that takes in a dataset with k input variables, has two hidden layers with l_1 and l_2 nodes respectively and an output layer with l_3 nodes, write the shapes[1] for ALL the parameters.

2. Using first-order gradient descent, write the update rules for parameters α and β for a model with the prediction function $f(x_i) = \alpha sin(\beta x_i)$ on a training sample vector x_i. Assume the loss function to be MSE.

3. For every step in gradient descent, a delta Δ_t is added to the parameters θ of the model to yield the new parameters. Prove that this Δ_t at a given timestep has to be $-\frac{\partial J(\theta_t)}{\partial \theta_t}$, where $J(\theta_t)$ is the evaluation of the objective function for the parameters θ at timestep t.

4. In linear regression models, you can solve for the optimal weights θ in the equation $\hat{Y} = X\theta$ by solving[2] for θ in the equation $\frac{\partial MSE(\theta)}{\partial \theta} = 0$.

[1] Including whether they are a matrix or vector.

[2] $MSE(\theta)$ is the mean-squared error function. Ignore the bias term. Hf with this problem.