1 Multiple Choice 1 point

Which of the following is an example of linear regression where y is the scalar output, **x** is the input vector, and **w** is a weight vector?

- ( ) $y = (\mathbf{w}^T x+1)^2 + (\mathbf{w}^T x)^2$
- ( ) $y = (\mathbf{w}^T x)^2$
- (•) $y = (\mathbf{w}^T x+1)^2 - (\mathbf{w}^T x)^2$
- ( ) $y = (\mathbf{w}^T x)^2 + \mathbf{w}^T x$

2 Multiple Choice 1 point

Suppose in a linear regression setting with squared loss, the design matrix is such that $\mathbf{X}^T\mathbf{X}$ is invertible. Suppose $\mathbf{w}^*$ is the minimizer of $|| \mathbf{y} - \mathbf{X}\,\mathbf{w} ||^2$. What will be the minimizer of $|| 2\,\mathbf{y} - \mathbf{X}\,\mathbf{w} ||^2$?

- ( ) $\mathbf{w}^*$
- (•) $2\ \mathbf{w}^*$
- ( ) $\mathbf{w}^* / 2$
- ( ) Cannot be written just in terms of $\mathbf{w}^*$

3 Multiple Choice 1 point

Suppose in a linear regression setting with squared loss, the design matrix is such that $\mathbf{X}^T\mathbf{X}$ is invertible. Suppose $\mathbf{w}^*$ is the minimizer of $|| \mathbf{y} - \mathbf{X}\,\mathbf{w} ||^2$. What will be the minimizer of $|| \mathbf{y} - 2\,\mathbf{X}\,\mathbf{w} ||^2$?

- ( ) $\mathbf{w}^*$
- ( ) $2\ \mathbf{w}^*$
- (•) $\mathbf{w}^* / 2$
- ( ) Cannot be written just in terms of $\mathbf{w}^*$

4 Multiple Choice 1 point

Recall that the norm $||\mathbf{w}||$ of a vector $\mathbf{w}$ is defined as $(\mathbf{w}^T\mathbf{w})^{1/2}$. What is the gradient of $||\mathbf{w}||^4$ with respect to $\mathbf{w}$?

- ( ) $2\,\mathbf{w}$
- ( ) $4\,||\,\mathbf{w}\,||^3$
- (•) $4\,\mathbf{w}\,\mathbf{w}^T\mathbf{w}$
- ( ) $2\,\mathbf{w}^T\mathbf{w}$

5 Multiple Choice 1 point

What does "minibatch" in "minibatch stochastic gradient descent" refer to?

- ( ) The fact that only a small number of entries of the weight vector are updated at each step
- (•) The fact that a randomly drawn small subsample of the entire training set is used at each step
- ( ) The fact that it uses a very small learning rate
- ( ) The fact that it attempts to make the loss function very small at each step

6 Multiple Choice 1 point

What does "stochastic" in "minibatch stochastic gradient descent" refer to?

- (•) The fact that randomness is involved in selecting the minibatch
- ( ) The fact that the learning rate is randomly selected
- ( ) The fact that the training set is a random draw from some underlying probability distribution
- ( ) The fact that the loss function is randomly selected

7 Multiple Choice 1 point

Suppose that we draw a single observation Y from a Gaussian distribution with mean m and variance 1. What is the maximum likelihood estimate of the parameter m in terms of the observation Y?

- (selected) Y
- m
- 1
- 0

8 Multiple Choice 1 point

When is it a good idea to use minibatch stochastic gradient descent?

- (selected) When the training set size is too large
- When the loss of the model weights is too large
- When the number of weights in your model is too large
- When the learning rate is too large

9 Multiple Choice 1 point

Suppose weight vector **w** is of size d x 1. What is the gradient of F(**w**) = **w** with respect to **w**?

- **w**
- 1
- (selected) Identity matrix of size d x d
- d x 1 vector consisting of all 1s

10 Multiple Choice 1 point

Which of the following is NOT a reasonable way to select model parameters given data?

- ( ) Choose a loss function and minimize the total loss over the dataset
- ( ) Setup a statistical model using the parameters and maximize the likelihood of the data given parameters
- ( ) Setup a statistical model using the parameters and minimize the negative log likelihood of the data given parameters
- (•) Choose a loss function and make sure that the losses for every training example are roughly equal

11 Multiple Choice 1 point

Why is "deep learning" called by that name?

- ( ) Because it uses very deep mathematics
- ( ) Because it has deep connections with neuroscience
- (•) Because it uses neural networks that have many layers
- ( ) Because the loss function landscape in the space of model parameters has deep valleys

12 Multiple Choice 1 point

Suppose a Gaussian distribution has the density p(x) proportional to exp(- x^2/8). What are the mean and standard deviation of this distribution?

- ( ) Can't be determined. Need to know the constant of proportionality.
- (•) Mean 0, Standard Deviation 2
- ( ) Mean 0, Standard Deviation 4
- ( ) Mean 0, Standard Deviation 8

13 Multiple Choice 1 point

Suppose f(**x**) = A **x** and g(**u**) = B **u** where A is m x n and B is k x m. Which of the following correctly describes the function g(f(**x**))?

- ( ) It's not well defined since the output dimension of f does not match the input dimension of g
- (•) It's a linear function given by B A **x**
- ( ) It's a linear function given by A B **x**
- ( ) It's a non linear function of **x**

14 Multiple Choice 1 point

When applying the principle of maximum likelihood, we can instead look at the negative log likelihood. In the context, which of the following statement is NOT correct?

- ( ) Under independence, taking logs gives a sum of terms to work with instead of a product of terms
- ( ) The negative sign allows us to express a maximization problem as a minimization problem and thus establish a link to loss minimization
- (•) Taking the negative log likelihood allows us to solve the optimization problem in closed form
- ( ) The solution of the problem of maximization of the likelihood is the same as that of minimization of the negative log likelihood

15 Multiple Choice 1 point

Suppose we view linear regression as a simple neural network. What happens to the number of layers of this network if we double the dimension of the input features?

- (•) Stays the same
- ( ) Number of layers gets doubled
- ( ) Number of layers gets halved
- ( ) Number of layers increases by 1

16 Multiple Choice 1 point

Which of the following is a point of similarity between artificial and biological neurons?

- (selected) They both have inputs and generate outputs
- In both cases, their weights are trained using gradient descent
- Both can only compute linear functions
- They are both incapable of aggregating their inputs

17 Multiple Choice 1 point

Which of the following is a classification problem?

- (selected) Predicting whether or not a student will pass STATS 315
- Predicting a student's GPA at graduation
- Predicting the time a student will take to solve the midterm exam
- Predicting the salary of a student in their first job after college

18 Multiple Choice 1 point

Which is of the following is a correct statement about the cross entropy loss function?

- It is incapable of handling the situation when true labels are soft
- It is not differentiable with respect to model predictions
- (selected) It is not bounded from above by a finite constant
- It is not bounded from below by a finite constant

19 Multiple Choice 1 point

Let R be the set of all real numbers and let S = { softmax(**x**) : **x** = (x_1, x_2) , x_1, x_2 ε R }. That is, S is the set obtained by applying the softmax function to all points in the two dimensional plane. Which of the following statements correctly describes S?

- (selected) $\{ (x_1, x_2) : 0 < x_1 < 1, 0 < x_2 < 1, x_1 + x_2 = 1 \}$
- $\{ (x_1, x_2) : 0 <= x_1 <= 1, 0 <= x_2 <= 1, x_1 + x_2 = 1 \}$
- $\{ (x_1, x_2) : -\infty < x_1 < +\infty, -\infty < x_2 < +\infty \}$
- $\{ (x_1, x_2) : -\infty <= x_1 <= +\infty, -\infty <= x_2 <= +\infty \}$

20 Multiple Choice 1 point

Suppose X has Gaussian distribution with mean 0 and standard deviation 1. Which of the following correctly describes X+1 ?

- (selected) It has a Gaussian distribution with mean 1 and standard deviation 1
- It has a Gaussian distribution with mean 0 and standard deviation 2
- Its distribution is no longer Gaussian
- It also has a Gaussian distribution with mean 0 and standard deviation 1

21 Multiple Choice 1 point

Consider softmax regression with cross-entropy loss. Suppose I am running minibatch stochastic gradient descent with a batch size of one and learning rate 0.1. I choose a random labeled example with a non-zero feature vector from the dataset. But the update does NOT change the weight vector. What can be correctly concluded from this?

- The training process has finished and we can stop training
- (selected) The label predicted by the current weight vector matches the true label of the chosen example
- Batch size is too small. We should increase it.
- The learning rate is not right. We should modify it.

22 Multiple Choice 1 point

We noted that cross-entropy loss can also be used when true labels in the labeled dataset are soft. Suppose you have a true label which is soft and whose entropy is H. What is the minimum value of the cross-entropy loss for such a true label?

- ( ) 0
- (•) H
- ( ) Insufficient information: we need to know the number of classes to compute the minimum value
- ( ) $-\infty$

23 Multiple Choice 1 point

Which of the following is a correct observation about cross-entropy and relative-entropy between two probability distributions?

- (•) Cross-entropy is at least as large as relative-entropy
- ( ) Relative-entropy is at least as large as cross-entropy
- ( ) Relative-entropy is always the same as cross-entropy
- ( ) Relative-entropy can be smaller than, larger than, or equal to, cross-entropy

24 Multiple Choice 1 point

Which technology company is primarily responsible for the development of Tensorflow?

- ( ) Facebook
- (•) Google
- ( ) Microsoft
- ( ) Amazon

25 Multiple Choice 1 point

The following code creates a neural network with how many layers?

```
from tensorflow import keras
from tensorflow.keras import layers
model = keras.Sequential([
    layers.Dense(512, activation="relu"),
    layers.Dense(10, activation="softmax")
])
```

- (selected) 2
- 512
- 10
- 522

26 Multiple Choice 1 point

What is the closest analogue of a numpy ndarray in tensorflow?

- (selected) tensor
- variable
- gradient tape
- layer

27 Multiple Choice 1 point

Having access to tensorflow/keras saves a programmer from which of the following activities?

- (selected) Implementing minibatch stochastic gradient descent
- Creating a neural network architecture
- Specifying a loss function
- Choosing a training dataset

28 Multiple Choice 1 point

What does the filetype extension ".ipynb" indicate?

- (●) jupyter notebook
- ( ) python notebook
- ( ) tensorflow notebook
- ( ) keras notebook

29 Multiple Choice 1 point

How do we deal with the restriction that tensorflow tensors are not assignable?

- ( ) By coding everything in numpy
- (●) By using tensorflow variables
- ( ) By using keras
- ( ) By using the gradient tape API

30 Multiple Choice 1 point

Which of the following branches of mathematics have we NOT used in this course?

- ( ) linear algebra
- ( ) probability theory
- ( ) multivariable calculus
- (●) differential equations