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?

- P(y=1|x) =$\sigma(\mathbf{w}^T x)$ where $\sigma(t) = 1/(1+\exp(-t))$
- **y = $\mathbf{x}^T\mathbf{w}$**
- y = $(\mathbf{w}^T x)^2$
- 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 $1/2||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$?

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

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 $||\mathbf{w}||$ with respect to $\mathbf{w}$ at $\mathbf{w}=\mathbf{0}$?

- $2\,\mathbf{w}$
- $1/2\,(\mathbf{w}^T\mathbf{w})^{-1/2}$
- **it is not differentiable at that point**
- $(\mathbf{w}^T\mathbf{w})^{-1/2}\,\mathbf{w}$

5 Multiple Choice 1 point

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

- **The fact that the algorithm tries to decrease the value of the objective function**
- The fact that you reduce the step size at each iteration
- The fact that it decreases the running time by using a small batch of training examples
- The fact that it tries to reduce the noise in the estimate of the gradient

6 Multiple Choice 1 point

Which of the following best describes gradient descent?

- It exactly minimizes the objective function
- **At every step it exactly minimizes a local approximation of the objective function**
- It is the best optimization algorithm we know of
- It is a randomized algorithm, that is, it uses randomness to speed-up certain computations

7 Multiple Choice 1 point

Suppose that we draw two independent observations Y_1, Y_2 from a Gaussian distribution with mean m and variance 1. What is the maximum likelihood estimate of the parameter m in terms of the two observations?

- Any one of the observations
- Both of the observations
- The sum of the observations
- **The mean of the observations**

8 Multiple Choice 1 point

How does the running time of one iteration of minibatch stochastic gradient descent increase as a function of minibatch size?

- It is constant
- **Increases linearly**
- Increases logarithmically
- Increases exponentially

9 Multiple Choice 1 point

Suppose $\mathbf{A}$ is a d x d matrix with all 1 entries. Suppose weight vector $\mathbf{w}$ is of size d x 1. What is the gradient of $F(\mathbf{w}) = \mathbf{w}^T \mathbf{A} \mathbf{w}$ with respect to $\mathbf{w}$?

- $\mathbf{w}$
- $\mathbf{A}\mathbf{w}$
- **$2\mathbf{A}\mathbf{w}$**
- $2\mathbf{w}$

10 Multiple Choice 1 point

Which of the following is NOT a reasonable loss function for regression? y is the ground truth, y' is model prediction

- $(y-y')^2$
- absolute value of (y-y')
- **$(y-y')^3$**
- $(y-y')^4$

11 Multiple Choice 1 point

Which of the following is a synonym for "deep learning"?

- Machine learning using nonlinear functions
- Machine learning using tensorflow
- **Machine learning using multilayer neural networks**
- Machine learning using Python and numpy

12 Multiple Choice 1 point

Suppose a distribution has the density p(x) proportional to exp(x) * exp(- $x^2/8$). What kind of a distribution is it?

- Can't be determined. Need to know the constant of proportionality.
- **Gaussian distribution**
- Standard Gaussian distribution
- Laplace distribution

13 Multiple Choice 1 point

Suppose **A** is a matrix of size n x n with all 1s in it. What is the rank of **A**?

- 0
- **1**
- n
- None of the other choices are correct

14 Multiple Choice 1 point

Suppose q represents a parameter in a statistical model and P(data | q) is the likelihood of the observed data under parameter q. Which of the following is equivalent to the principle of maximum likelihood?

- (●) Maximize square root of P(data | q)
- ( ) Minimize P(data | q)
- ( ) Maximize -log(P(data | q))
- ( ) Maximize exp(-P(data | q))

15 Multiple Choice 1 point

Suppose a fully connected layer inside of a neural network takes in m scalar inputs and produces n scalar outputs. How does the number of trainable parameters in such a layer scale with m and n?

- (●) O(m * n)
- ( ) O(m)
- ( ) O(n)
- ( ) O(1)

16 Multiple Choice 1 point

Which of the following is a correct statement about biological neural networks?

- (●) They consist of biological neurons connected with each other
- ( ) They are trained using backpropagation
- ( ) They are trained using stochastic gradient descent
- ( ) They are incapable of computing non-linear functions

17 Multiple Choice 1 point

Which of the following is a regression problem?

- Predicting whether or not a student will pass STATS 315
- Predicting whether or not the end-of-semester letter grade of a student in STATS 315 will be an A+
- Predicting whether or not a student initially enrolled in STATS 315 will drop the class within the first 4 weeks of the semester
- **Predicting the time a student will need to finish the final exam in STATS 315**

18 Multiple Choice 1 point

What is the maximum possible value of the entropy (in bits) of a distribution of a random variable that can take N distinct values.

- N
- $+\infty$
- **$\log_2(N)$**
- 1

19 Multiple Choice 1 point

Suppose we have a coin whose probability of landing HEADS is p and the probability of landing TAILS is 1-p. What is the entropy of the outcome of this coin's toss?

- **- p log p - (1-p) log (1-p)**
- p log p + (1-p) log (1-p)
- 1
- 0

20 Multiple Choice 1 point

Suppose Y takes values either 0 or 1 and that

$P(Y = 1) = 1/(1 + \exp(-t))$.

What happens to the entropy of the distribution of Y as $t \to +\infty$?

- [x] It tends to 0
- [ ] It tends to 1
- [ ] It tends to plus infinity
- [ ] It tends to 0.5

21 Multiple Choice 1 point

Consider linear regression with squared loss. Suppose I am running minibatch stochastic gradient descent with a batch size of one and positive learning rate. The current weight vector is **w**. I choose a random labeled example with feature vector **x** and associated response y from the dataset. After the update, the weight vector is now **w** + **x**. What can be correctly concluded from this?

- [ ] The training process has finished and we can stop training
- [ ] **w** perfectly predicts the response y associated with **x**
- [x] **w** under-predicts the response y associated with **x**
- [ ] **w** over-predicts the response y associated with **x**

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 maximum possible value of the cross-entropy loss for such a true label?

- [ ] 0
- [ ] H
- [ ] 1
- [x] ∞

23 Multiple Choice 1 point

Suppose p is a probability distribution with entropy H. What is the cross-entropy between p and p itself?

- (●) H
- ( ) 0
- ( ) 1
- ( ) -∞

24 Multiple Choice 1 point

Which package had its first release earlier, Keras or Tensorflow?

- ( ) Tensorflow
- (●) Keras
- ( ) They were released together
- ( ) They have not been released yet to the general public

25 Multiple Choice 1 point

Why does the following code not run properly?

```
import tensorflow as t

x = t.constant(3.0)
with t.GradientTape() as g:
  y = x * x
dy_dx = g.gradient(y, x)
print(dy_dx.numpy())
```

- ( ) tensorflow should be imported as tf, not t
- ( ) GradientTape() should be named tape, not g
- (●) x is a constant and so its gradients are not tracked by default
- ( ) all of the other choices are correct

26 Multiple Choice 1 point

The main goal of the Gradient Tape API in tensorflow is to:

- [x] make it easy to perform automatic differentiation
- [ ] make it easy to design new deep learning architectures
- [ ] make it easy to select good quality datasets
- [ ] make it easy to manipulate tensors

27 Multiple Choice 1 point

Why is backpropagation called by that name?

- [x] Because it has a backward pass in addition to a forward pass
- [ ] Because it is backwards compatible with propagation
- [ ] Because it serves as a backup in case gradient descent doesn't work
- [ ] Because it keeps coming back to the same expression differentiating it over and over again

28 Multiple Choice 1 point

Suppose I have a binary classification problem and I want to model the probability of the label being 1 linearly. I write down my model as:

$P(y = 1| \mathbf{x}) = \mathbf{w}^T\mathbf{x}$

What's the problem with such a model?

- [ ] We should not model probabilities to solve a classification problem
- [x] This model does not ensure that we always get valid probabilities
- [ ] This model will suffer from having too many parameters
- [ ] This model will suffer from having too few parameters

29 Multiple Choice 1 point

What trick can we use to get rid of the bias b in a linear model $\mathbf{w}^T\mathbf{x} + b$?

- Add a dummy feature to all examples that is always 0
- **Add a dummy feature to all examples that is always 1**
- Add a dummy feature to all examples that is chosen randomly
- Add a dummy feature that is 1 for all examples with a positive response and is 0 for all examples with a negative response

30 Multiple Choice 1 point

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

- linear algebra
- information theory
- multivariable calculus
- **topology**