1. Mark each statement True or False.

   (a) In linear regression, a reasonable alternative to the cost function *mean squared error* $= \frac{1}{N}\sum_{i=1}^{N}[f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]^2$ is *mean error* $= \frac{1}{N}\sum_{i=1}^{N}[f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]$.

   ◯ True
   ◯ False ●

   (b) A linear SVM with decision boundary $(2, 1, 2) \cdot \mathbf{x} + 2 = 0$ has a smaller margin between $+1$ and -1 support vectors than one with boundary $(6, 0, 8) \cdot \mathbf{x} - 3 = 0$.

   ◯ True
   ◯ False ANSWER: ●
   The margin for the first SVM is $\frac{2}{||\mathbf{w}||} = \frac{2}{\sqrt{2^2+1^2+2^2}} = \frac{2}{3}$, while the margin for the second is $\frac{2}{||\mathbf{w}||} = \frac{2}{\sqrt{6^2+0^2+8^2}} = \frac{2}{10}$.

   (c) The values for $\mathbf{w}$ and b that minimize negative log-likelihood in the logistic regression model also minimize the mean squared error $\frac{1}{N}\sum_{i=1}^{N}[P_{\mathbf{w},b}(y_i = 1)(\mathbf{x}_i) - y_i]^2$.

   ◯ True
   ◯ False ANSWER: ●
   A predicted probability $\hat{P}_{\mathbf{w},b}(y_i = 1)(\mathbf{x}_i) \in \{0, 1\}$, while a label $y_i \in [0, 1]$. The logistic curve does not fit the data points.

   (d) The entropy of a decision tree node containing the set of examples

x_1	x_2	y
2	6	0
5	7	0
3	8	1

   is ≈ 0.92.

   ◯ True ANSWER: ●
   The node's y values are 0, 0, 1, so $f_{ID3}(S) = P(y) = \frac{1}{|S|}\sum_{(\mathbf{x},y)\in S} y = \frac{1}{3}(0+0+1) = \frac{1}{3}$.

$$
\begin{aligned}
H(S) &= \sum_{y\in\{0,1\}} P(y)\,[-\log_2 P(y)] \\
&= -\left(1 - \frac{1}{3}\right)\log_2\left(1 - \frac{1}{3}\right) - \frac{1}{3}\log_2\frac{1}{3} \\
&\approx 0.92
\end{aligned}
$$

   ◯ False

   (e) While gradient descent's computation speed depends on the number of features D, stochastic gradient descent's computation speed does not depend on D.

- ◯ True
- ◯ False ANSWER: ●

  Even if SGD uses only one example for each iteration, it can still take longer to process a high-D example than a low-D example.

(f) For a logistic regression model on data with one feature, the midpoint of the logistic curve is always between the feature minimum and the feature maximum.

- ◯ True
- ◯ False ANSWER: ●

  Consider, e.g., a data set whose "sample proportions" are all small.

2. Consider the logistic regression model, $P(y_i = 1) = \dfrac{1}{1 + e^{-(\mathbf{wx}+b)}}$.

(a) For each function below that plays a role in the model, indicate its image from among these choices:

A. $\mathbb{Z}$ = integers

B. $\mathbb{Z}_+$ = positive integers

C. $\mathbb{R}$ = real numbers

D. $\mathbb{R}_+$ = positive real numbers

E. $(0, 1)$ = interval from 0 to 1

Hint: The *image* of a function is the set of all output values it may produce.

i. $f_1(\mathbf{x}) = \mathbf{wx} + b$ for $\mathbf{x} \in \mathbb{R}^D$:

A ○, B ○, C ○ANSWER: ● , D ○, E ○

ii. $f_2(t) = \frac{1}{1+e^{-t}}$ for $t \in \mathbb{R}$:

A ○, B ○, C ○, D ○, E ○ANSWER: ●

iii. $f_3(t) = e^{-t}$ for $t \in \mathbb{R}$:

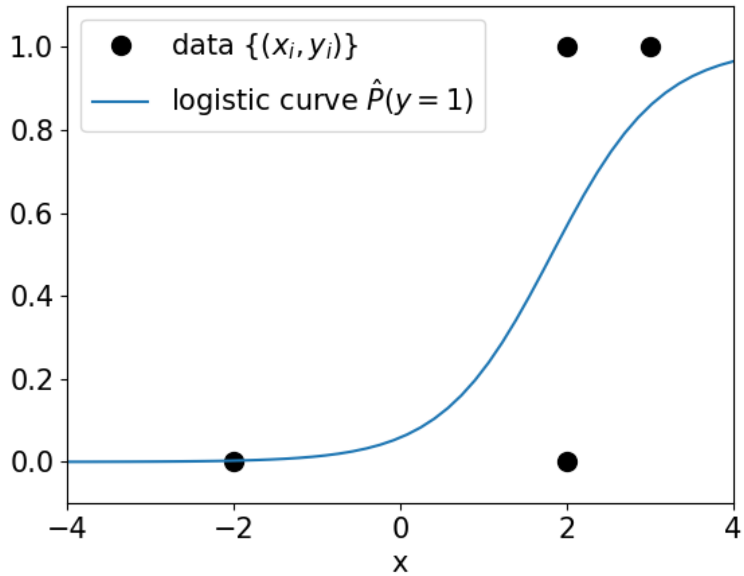
A ○, B ○, C ○, D ○ANSWER: ● , E ○

(b) For the model with $\mathbf{w} = (-3, 3)$ and $b = 3$, find $\hat{P}\left(y = 1 | \mathbf{x} = (2, 1)\right)$. $\boxed{}$

ANSWER:

$$\hat{P}(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{wx}+b)}} = \frac{1}{1 + e^{-[(-3,3)\cdot(2,1)+3]}} = \frac{1}{1 + e^0} = \frac{1}{2}.$$

(c) I ran some Python/scikit-learn code to make the model pictured here:



i. For each array below, indicate the line of code that could have produced it from among these choices:

A. `model.fit(X, y)`

B. `model.intercept_`

C. `model.coef_[0]`

D. `model.predict(X)`

E. `model.predict_proba(X)[:, 1]`

F. `model.score(X, y)`

1. array([0, 1, 1, 1]):

   A ○, B ○, C ○, D ○● , E ○, F ○

2. array([0.003, 0.569, 0.569, 0.859])

   A ○, B ○, C ○, D ○, E ○● , F ○

3. array([1.528])

   A ○, B ○, C ○● , D ○, E ○, F ○

4. array([-2.778])

   ○, B ○● , C ○, D ○, E ○, F ○

ii. How do we classify a new point at $x = 0.5$ if using a decision threshold of 0.5?

○ $\hat{y} = 0$ ●

○ $\hat{y} \approx 0.05$

○ $\hat{y} \approx 0.95$

○ $\hat{y} = 1$

ANSWER:
$\hat{y} = 0$. The graph shows $\hat{P}_{\mathbf{w},b}(y = 1|x = 0.5)$ is between 0 and 0.2 (Python says ≈ 0.12), less than the 0.5 threshold. So we assign $\hat{y} = 0$.

3. Here are some questions about decision trees.

(a) Consider a classification decision tree node containing the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = (x_1, x_2, x_3)$:

S

x_1	x_2	x_3	y
2	11	12	1
3	6	14	1
0	8	17	0
4	10	15	1
1	7	13	0
5	9	16	1

i. The entropy of this node in bits is $\boxed{}$.

ANSWER:
The node's y values are 1, 1, 0, 1, 0, 1, so $f_{ID3}(S) = \frac{1}{|S|}\sum_{(\mathbf{x},y)\in S} y = \frac{1}{6}(1+1+0+1+0+1) = \frac{2}{3}$.
$H(S) = \frac{2}{3}(-\log_2(\frac{2}{3})) + \frac{1}{3}(-\log_2(\frac{1}{3})) \approx -\frac{2}{3}(-0.585) + \frac{1}{3}(-1.585) \approx 0.918$

ii. The (feature, threshold) pair (j, t) that yields the best split for this node is feature $j = \boxed{}$ and threshold $t = \boxed{}$.

ANSWER:
Using feature $j = 1$ and threshold $t = 1.5$ (or any $t \in [1, 2)$) splits S into $S_- = \{(\mathbf{x}, y) \in S | x^{(j)} \leq t\} = \{0, 0\}$ and its complement $S_+ = \{(\mathbf{x}, y) \in S | x^{(j)} > t\} = \{1, 1, 1, 1\}$, each of which has entropy 0.

(b) Consider a regression decision tree with `max_depth=1` (that is, the root node is split once into two leaves) made from the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = x_1$:

S

x_1	y
0	10
1	11
2	21
3	22
4	23
5	24

What value does this tree predict for $x_1 = 4.5$? $\hat{y} = \boxed{}$

ANSWER:

The best split uses feature $j = 1$ and threshold $t = 1.5$, yielding a left subtree containing the first two examples and a right subtree containing the last four. Making a predition with $x_1 = 4.5$ would use the right subtree. Its average y is 22.5, so the tree would predict $\hat{y} = 22.5$.

4. Consider this training data set:

x_1	x_2	x_3	y
0	3	0	Red
2	0	0	Red
0	1	3	Red
0	1	2	Green
-1	0	1	Green
1	1	1	Red

We use this data set to make a prediction for y when $x_1 = x_2 = x_3 = 0$ using k-NN.

(a) Compute the Euclidean distance between each example and the test example, $(x_1, x_2, x_3) = (0, 0, 0)$.

(b) What is our prediction with $k = 1$?

(c) What is our prediction with $k = 3$?

ANSWER:

(a) The distances are (from top to bottom): $3, 2, \sqrt{10}, \sqrt{5}, \sqrt{2}, \sqrt{3}$.

(b) The shortest distance is $\sqrt{2}$ and the corresponding y is Green, so we predict $y =$ Green.

(c) The shortest three distances are $\sqrt{2}, \sqrt{3}, 2$ and the corresponding y are Green, Red, and Red, so we predict $y =$ Red.

5. When the number D of features is large, performance of k-NN (and other local approaches that predict using only examples near the test example) tends to deteriorate. This is known as the *curse of dimensionality.*

   (a) Suppose $D = 1$. Suppose x is uniformly distributed on $[0, 1]$, the unit interval. We predict a test example's y response using only the examples in the 10% of the x range nearest to that test example. For instance, to predict for $x = 0.6$, we use examples in the range $[0.55, 0.65]$. On average, what proportion of the examples will we use to make the prediction?
   (Hint: There is a simple answer. This is not a trick question.)
   ANSWER:
   proportion $= 10\% = 0.1$.

   (b) Now suppose $D = 2$ and our feature are x_1 and x_2, with (x_1, x_2) uniformly distributed on $[0, 1] \times [0, 1]$ (the unit square). For a given test example, we predict using examples in the closest 10% of the x_1 range and in the closest 10% of the x_2 range. For instance, in order to predict the response for a test example with $x_1 = 0.6$ and $x_2 = 0.35$, we use examples in the range $[0.55, 0.65]$ for x_1 and in the range $[0.3, 0.4]$ for x_2. On average, what proportion of the available examples will we use to make the prediction?
   ANSWER:
   proportion $= 0.1 \times 0.1 = 0.1^2 = 0.01$.

   (c) Now suppose $D = 100$. Again, each feature is uniformly distributed on $[0, 1]$, so each example is from $[0, 1]^{100}$ (the 100-dimensional "unit hypercube"). We predict a test example's response y using examples within the closest 10% of each feature's range. What proportion of the available examples will we use to make the prediction?
   ANSWER:
   proportion$=0.1^{100} = 10^{-100}$.

   (d) Now suppose that we wish to make a prediction for a test example by creating a D-dimensional hypercube centered around the test example that contains, on average, 10% of the training examples. What is the length l of each side of the hypercube for each value of D?
   Hint: Solve the equation $l^D = 0.1$.

      i. For $D = 1$, $l =$ _______
      ii. For $D = 2$, $l =$ _______
      iii. For $D = 100$, $l =$ _______

   Comment on what happens to the length l of each side as $D \to \infty$?
   ANSWER:
   For

      i. $D = 1$ we have $l = 0.1$;
      ii. $D = 2$ we have $l = 0.316$;
      iii. $D = 3$ we have $l = 0.977$.

   As $D \to \infty$, the solution for above equation tends toward 1, which means for every feature, we should use almost all its range only to make 10% of examples can be used for prediction.

6. Suppose we have a soft-margin SVM for which $\mathbf{w} = (6, -3, 2)$ and $b = 1$. Consider the example $(\mathbf{x} = (-1, 2, 1), y = -1)$.

   (a) How does the SVM classify $\mathbf{x}$?

   ANSWER:

   $\mathbf{wx} + b = -9 < 0 \implies \hat{y} = -1$

   (b) Does $(\mathbf{x}, y)$ satisfy the SVM constraint? (Answer Yes or No.)

   ANSWER:

   $\mathbf{wx} + b = -9 \leq -1 \implies$ yes.

   (c) What is the hinge loss associated with $(\mathbf{x}, y)$?

   ANSWER:

   $\max(0, 1 - y_i(\mathbf{wx}_i + b)) = \max(0, 1 - (-1)(-9)) = \max(0, -8) = 0$

7. In each situation, indicate the frequency (never, sometimes, or always) with which we will obtain 100% accuracy on training data that contain no pairs of examples with identical feature vectors. Suppose any hyperparameters are set to optimum values for training performance.

(a)	Decision tree with $N = 1$ example	always ○● ,	sometimes ○,	never ○
(b)	Decision tree with $N = 5$ examples	always ○,	sometimes ○● ,	never ○
(c)	Hard-margin linear SVM on linearly-separable data	always ○● ,	sometimes ○,	never ○
(d)	Hard-margin linear SVM on non-linearly-separable data	always ○,	sometimes ○,	never ○●
(e)	Soft-margin linear SVM on non-linearly-separable data	always ○,	sometimes ○,	never ○●
(f)	SVM with RBF kernel on non-linearly-separable data	always ○● ,	sometimes ○,	never ○
(g)	k-NN with $k = 1$	always ○● ,	sometimes ○,	never ○
(h)	k-NN with $k = 3$	always ○,	sometimes ○● ,	never ○
(i)	Logistic regresion	always ○,	sometimes ○● ,	never ○

8. Suppose we have the regression model $y = 3x + 4$.

   (a) If y is converted to $1000y$ (e.g., y units are changed from kilograms (kg) to grams (g)),

      i. the slope will be ________________ and
      ii. the intercept will be ________________.

   (b) If x is converted to $2x$,

      i. the slope will be ________________ and
      ii. the intercept will be ________________.

   ANSWER:

   (a) i. $3 \times 1000 = 3000$
   ii. $4 \times 1000 = 4000$

   (b) i. $3 \times \frac{1}{2} = \frac{3}{2} = 1.5$
   ii. 4

9. Consider gradient descent to minimize the loss function $L = x^3 + 2y^2 - 4xy + 3$.

(a) What is the gradient of L starting at $(x_0, y_0) = (1, 2)$?

ANSWER:
The gradient is $(3x^2 - 4y, 4y - 4x)$; at (1, 2), this is $\big(3(1^2) - 4(2), 4(2) - 4(1)\big) = (-5, 4)$.

(b) If we set the step size $\alpha = 1$, the next point visited by gradient descent is $(x_1, y_1) =$ ____________

ANSWER: $(x_1, y_1) = (1, 2) - 1 \times (-5, 4) = (6, -2)$