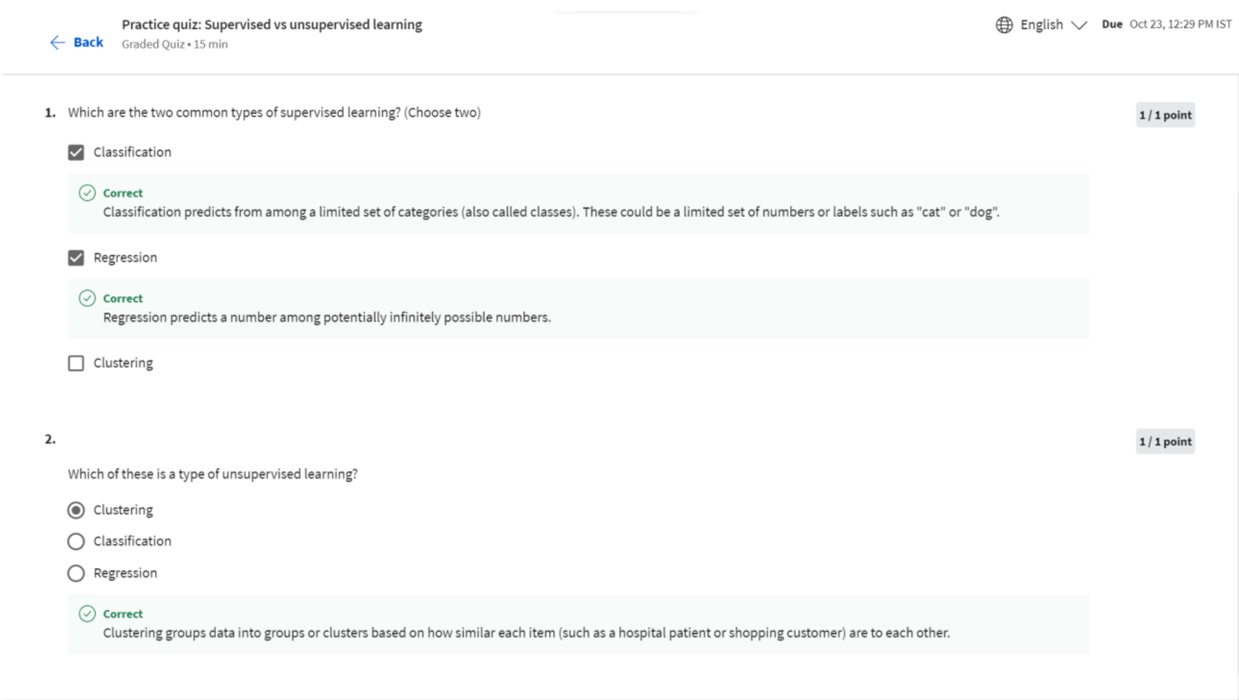
Machine Learning Specialization

Data Science -Mooc Course

Supervised Machine Learning: Regression and Classification

Practice quiz: Supervised vs unsupervised learning



Practice quiz: Supervised vs unsupervised learning
Back
Graded Quiz • 15 min

English Due Oct 23, 12:29 PM IST

1. Which are the two common types of supervised learning? (Choose two) 1 / 1 point

- [x] Classification

Correct
Classification predicts from among a limited set of categories (also called classes). These could be a limited set of numbers or labels such as "cat" or "dog".

- [x] Regression

Correct
Regression predicts a number among potentially infinitely possible numbers.

- [ ] Clustering

2. 1 / 1 point

Which of these is a type of unsupervised learning?

- (•) Clustering
- ( ) Classification
- ( ) Regression

Correct
Clustering groups data into groups or clusters based on how similar each item (such as a hospital patient or shopping customer) are to each other.

Practice quiz: Regression

← Back Practice quiz: Regression
Graded Quiz • 10 min

English Due Oct 23, 12:29 PM IST

1. 1 / 1 point

For linear regression, the model is $f_{w,b}(x) = wx + b$.

Which of the following are the inputs, or features, that are fed into the model and with which the model is expected to make a prediction?

- (●) x
- ( ) m
- ( ) w and b.
- ( ) (x, y)

Correct
The x, the input features, are fed into the model to generate a prediction $f_{w,b}(x)$

2. For linear regression, if you find parameters w and b so that $J(w, b)$ is very close to zero, what can you conclude? 1 / 1 point

- ( ) This is never possible -- there must be a bug in the code.
- (●) The selected values of the parameters w and b cause the algorithm to fit the training set really well.
- ( ) The selected values of the parameters w and b cause the algorithm to fit the training set really poorly.

Correct
When the cost is small, this means that the model fits the training set well.

Practice quiz: Train the model with gradient descent

1.

Gradient descent is an algorithm for finding values of parameters w and b that minimize the cost function J.

repeat until convergence {

$$w = w - \alpha \frac{\partial}{\partial w} J(w, b)$$

$$b = b - \alpha \frac{\partial}{\partial b} J(w, b)$$

When $\frac{\partial J(w,b)}{\partial w}$ is a negative number (less than zero), what happens to w after one update step?

- ( ) w stays the same
- (●) w increases.
- ( ) w decreases
- ( ) It is not possible to tell if w will increase or decrease.

Correct
The learning rate is always a positive number, so if you take W minus a negative number, you end up with a new value for W that is larger (more positive).

2.

For linear regression, what is the update step for parameter b?

- (selected) $b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)})$
- $b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$

Correct

The update step is $b = b - \alpha \frac{\partial J(w,b)}{\partial w}$ where $\frac{\partial J(w,b)}{\partial b}$ can be computed with this expression: $\sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)})$

Practice quiz: Multiple linear regression

1. In the training set below, what is $x_4^{(3)}$? Please type in the number below (this is an integer such as 123, no decimal points).

1 / 1 point

Size in feet²	Number of bedrooms	Number of floors	Age of home in years	Price ($) in $1000's
x_1	x_2	x_3	x_4	
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

30

Correct

Yes! $x_4^{(3)}$ is the 4th feature (4th column in the table) of the 3rd training example (3rd row in the table).

2.

Which of the following are potential benefits of vectorization? Please choose the best option.

- ○ It makes your code run faster
- ○ It can make your code shorter
- ○ It allows your code to run more easily on parallel compute hardware
- ◉ All of the above

Correct
Correct! All of these are benefits of vectorization!

3. True/False? To make gradient descent converge about twice as fast, a technique that almost always works is to double the learning rate $alpha$.

- ○ True
- ◉ False

Correct
Doubling the learning rate may result in a learning rate that is too large, and cause gradient descent to fail to find the optimal values for the parameters w and b.

Practice quiz: Gradient descent in practice

Graded Quiz. • 30 min

← Back

Practice quiz: Gradient descent in practice
Graded Quiz • 30 min

English Due Oct 30, 12:29 PM IST

1. 1 / 1 point

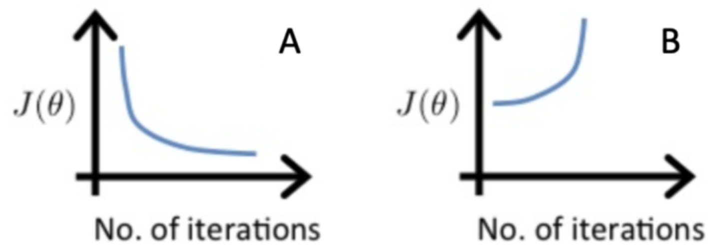
Which of the following is a valid step used during feature scaling?

- ○ Add the mean (average) from each value and and then divide by the (max - min).
- ◉ Subtract the mean (average) from each value and then divide by the (max - min).

Correct
This is called mean normalization.

Practice quiz: Gradient descent in practice
Back Graded Quiz • 30 min
English Due Oct 30, 12:29 PM IST

2. Suppose a friend ran gradient descent three separate times with three choices of the learning rate α and plotted the learning curves for each (cost J for each iteration). 1 / 1 point



For which case, A or B, was the learning rate α likely too large?

- ○ Both Cases A and B
- ○ case A only
- ◉ case B only
- ○ Neither Case A nor B

Correct
The cost is increasing as training continues, which likely indicates that the learning rate alpha is too large.

Practice quiz: Gradient descent in practice
Back Graded Quiz • 30 min
English Due Oct 30, 12:29 PM IST

3. Of the circumstances below, for which one is feature scaling particularly helpful? 1 / 1 point

- ○ Feature scaling is helpful when all the features in the original data (before scaling is applied) range from 0 to 1.
- ◉ Feature scaling is helpful when one feature is much larger (or smaller) than another feature.

Correct
For example, the "house size" in square feet may be as high as 2,000, which is much larger than the feature "number of bedrooms" having a value between 1 and 5 for most houses in the modern era.

4. 1 / 1 point

You are helping a grocery store predict its revenue, and have data on its items sold per week, and price per item. What could be a useful engineered feature?

- ◉ For each product, calculate the number of items sold times price per item.
- ○ For each product, calculate the number of items sold divided by the price per item.

Correct
This feature can be interpreted as the revenue generated for each product.

5. True/False? With polynomial regression, the predicted values f_w,b(x) does not necessarily have to be a straight line (or linear) function of the input feature x. 1 / 1 point

- ◉ True
- ○ False

Correct

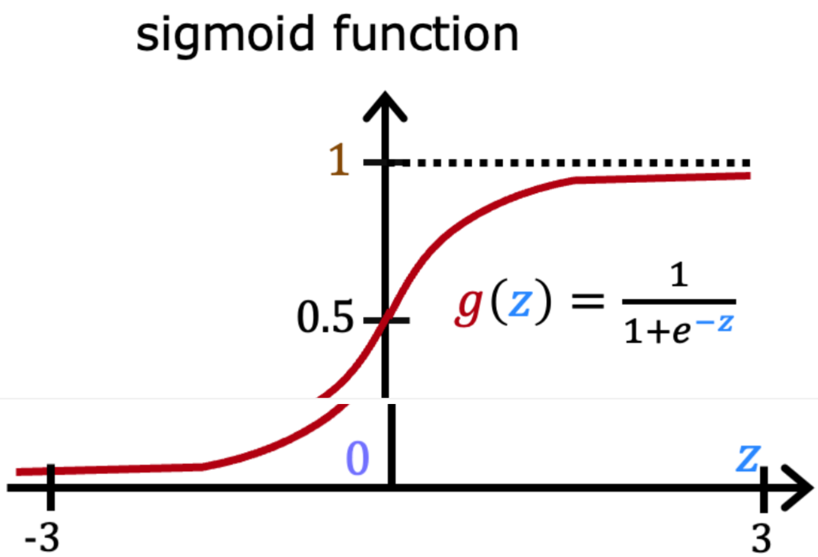
Practice quiz: Classification with logistic regression

1. Which is an example of a classification task?

- ○ Based on a patient's age and blood pressure, determine how much blood pressure medication (measured in milligrams) the patient should be prescribed.
- ○ Based on a patient's blood pressure, determine how much blood pressure medication (a dosage measured in milligrams) the patient should be prescribed.
- ◉ Based on the size of each tumor, determine if each tumor is malignant (cancerous) or not.

✓ **Correct**
This task predicts one of two classes, malignant or not malignant.

2. Recall the sigmoid function is $g(z) = \frac{1}{1+e^{-z}}$

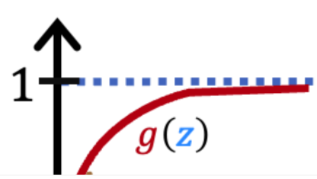


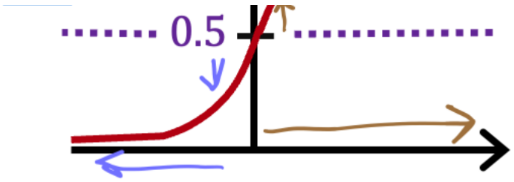
If z is a large positive number, then:

- ○ $g(z)$ is near negative one (-1)
- ○ $g(z)$ will be near 0.5
- ○ $g(z)$ will be near zero (0)
- ◉ $g(z)$ is near one (1)

✓ **Correct**
Say z = $+100$. So e^{-z} is then e^{-100}, a really small positive number. So, $g(z) = \frac{1}{1+\text{a small positive number}}$ which is close to 1

3.





A cat photo classification model predicts 1 if it's a cat, and 0 if it's not a cat. For a particular photograph, the logistic regression model outputs $g(z)$ (a number between 0 and 1). Which of these would be a reasonable criteria to decide whether to predict if it's a cat?

- (selected) Predict it is a cat if g(z) >= 0.5
- Predict it is a cat if g(z) < 0.5
- Predict it is a cat if g(z) < 0.7
- Predict it is a cat if g(z) = 0.5

Correct
Think of g(z) as the probability that the photo is of a cat. When this number is at or above the threshold of 0.5, predict that it is a cat.

4.

True/False? No matter what features you use (including if you use polynomial features), the decision boundary learned by logistic regression will be a linear decision boundary.

- (selected) False
- True

Practice quiz: Cost function for logistic regression

1.

$$\overbrace{J(\vec{w},b)}^{?} = \frac{1}{m}\sum_{i=1}^{m}\underbrace{L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right),y^{(i)}\right)}_{?}$$

In this lecture series, "cost" and "loss" have distinct meanings. Which one applies to a single training example?

- [x] Loss

Correct
In these lectures, loss is calculated on a single training example. It is worth noting that this definition is not universal. Other lecture series may have a different definition.

- [ ] Cost
- [ ] Both Loss and Cost
- [ ] Neither Loss nor Cost

2.

Simplified loss function

$$L(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)}) = \begin{cases} -\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right) & \text{if } y^{(i)} = 0 \end{cases}$$

$$\boxed{L(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)}) = -y^{(i)}\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) - (1 - y^{(i)})\log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right)}$$

For the simplified loss function, if the label $y^{(i)} = 0$, then what does this expression simplify to?

- ○ $-\log(1 - f_{\vec{w},b}(\mathbf{x}^{(i)})) - log(1 - f_{\vec{w},b}(\mathbf{x}^{(i)}))$
- ◉ $-\log(1 - f_{\vec{w},b}(\mathbf{x}^{(i)}))$
- ○ $\log(1 - f_{\vec{w},b}(\mathbf{x}^{(i)})) + log(1 - f_{\vec{w},b}(\mathbf{x}^{(i)}))$
- ○ $\log(f_{\vec{w},b}(\mathbf{x}^{(i)})$

Correct
When $y^{(i)} = 0$, the first term reduces to zero.

Practice quiz: Gradient descent for logistic regression

$$w_j = w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}) x_j^{(i)} \right]$$

$$b = b - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} (f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}) \right]$$

} simultaneous updates

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$

Which of the following two statements is a more accurate statement about gradient descent for logistic regression?

- (selected) The update steps look like the update steps for linear regression, but the definition of $f_{\vec{w},b}(\mathbf{x}^{(i)})$ is different.
- The update steps are identical to the update steps for linear regression.

Correct
For logistic regression, $f_{\vec{w},b}(\mathbf{x}^{(i)})$ is the sigmoid function instead of a straight line.

Practice quiz: The problem of overfitting

1. Which of the following can address overfitting?

- [x] Apply regularization

Correct
Regularization is used to reduce overfitting.

- [x] Collect more training data

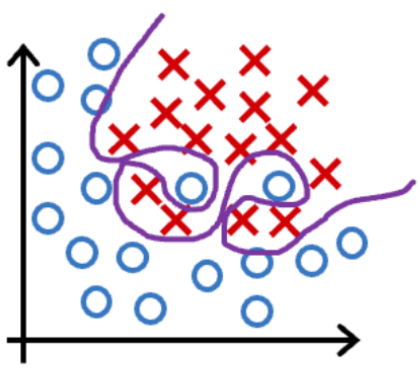
Correct
If the model trains on more data, it may generalize better to new examples.

- [x] Select a subset of the more relevant features.

Correct
If the model trains on the more relevant features, and not on the less useful features, it may generalize better to new examples.

- [ ] Remove a random set of training examples

2. You fit logistic regression with polynomial features to a dataset, and your model looks like this.



What would you conclude? (Pick one)

- ○ The model has high variance (overfit). Thus, adding data is, by itself, unlikely to help much.
- ○ The model has high bias (underfit). Thus, adding data is, by itself, unlikely to help much.
- ○ The model has high bias (underfit). Thus, adding data is likely to help
- ◉ The model has high variance (overfit). Thus, adding data is likely to help

Correct
The model has high variance (it overfits the training data). Adding data (more training examples) can help.

3\. # Regularization

$$\min_{\vec{w},b} J(\vec{\mathrm{w}}, b) = \min_{\vec{w},b} \left[\overbrace{\frac{1}{2m} \sum_{i=1}^{m} \left(f_{\vec{\mathrm{w}},b}(\vec{\mathrm{x}}^{(i)}) - y^{(i)}\right)^2}^{\text{mean squared error}} + \overbrace{\frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2}^{\text{regularization term}} \right]$$

Suppose you have a regularized linear regression model. If you increase the regularization parameter λ, what do you expect to happen to the parameters $w_1, w_2, \ldots, w_n$?

- (•) This will reduce the size of the parameters $w_1, w_2, \ldots, w_n$
- ( ) This will increase the size of the parameters $w_1, w_2, \ldots, w_n$

Correct
Regularization reduces overfitting by reducing the size of the parameters $w_1, w_2, \ldots w_n$.