Exercise: Check your understanding

What parts of the linear regression equation are updated during training?

- [ ] The prediction
- [x] The bias and weights

  During training, the model updates the bias and weights.

  Correct answer.
- [ ] The feature values

Which of the two linear models shown in the preceding plots has the higher Mean Squared Error (MSE) when evaluated on the plotted data points?

- [x] The model on the right.

  The eight examples on the line incur a total loss of 0. However, although only two points lay off the line, both of those points are *twice* as far off the line as the outlier points in the left figure. Squared loss amplifies those differences, so an offset of two incurs a loss four times as great as an offset of one:
  $MSE = \frac{0^2+0^2+0^2+2^2+0^2+0^2+0^2+2^2+0^2+0^2}{10} = 0.8$

  Correct answer.
- [ ] The model on the left.

Linear regression: Parameters exercise

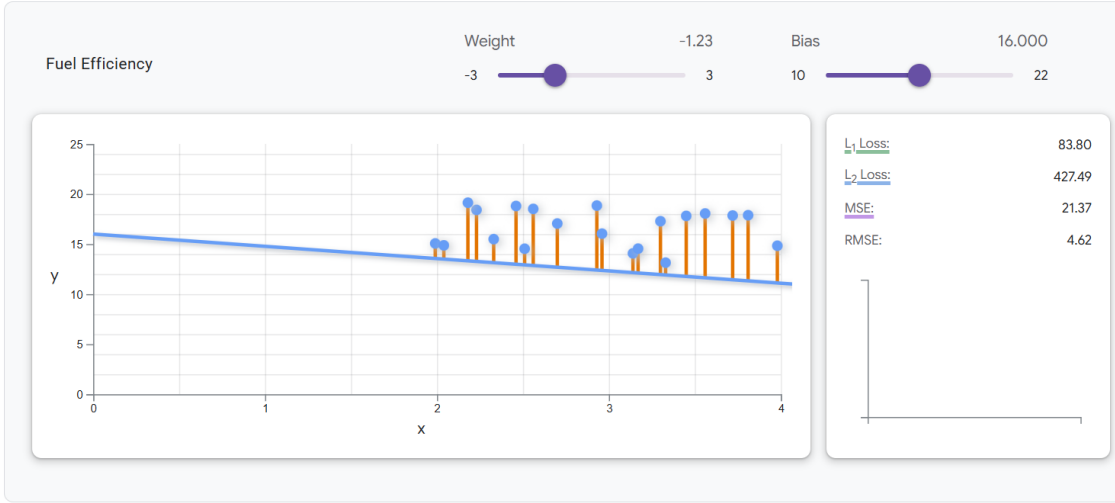
Send feedback

The graph below plots 20 examples from a fuel-efficiency dataset, with the feature (car heaviness in thousands of pounds) plotted on the x-axis and the label (miles per gallon) plotted on the y-axis.

Your task: Adjust the **Weight** and **Bias** sliders above the graph to find the linear model that minimizes MSE loss on the data.

Questions to consider:

- What is the lowest MSE you can achieve?
- What weight and bias values produced this loss?



Exercise: Check your understanding

1. What's the best batch size when using mini-batch SGD?

- [x] It depends

  The ideal batch size depends on the dataset and the available compute resources

  Correct answer.

- [ ] 100 examples per batch
- [ ] 10 examples per batch

2. Which of the following statements is true?

- [ ] Larger batches are unsuitable for data with many outliers.
- [x] Doubling the learning rate can slow down training.

  This statement is true. Doubling the learning rate can result in a learning rate that is too large, and therefore cause the weights to "bounce around," increasing the amount of time needed to converge. As always, the best hyperparameters depend on your dataset and available compute resources.

  Correct answer.

Answer the following two questions.

1. What is the value of *z* for these input values?

- [ ] −1
- [ ] 0
- [ ] 0.731
- [x] 1

  Correct! The linear equation defined by the weights and bias is $z = 1 + 2x_1 - x_2 + 5x_3$. Plugging the input values into the equation produces z = 1 + (2)(0) - (10) + (5)(2) = 1

  Correct answer.

2. What is the logistic regression prediction for these input values?

- [ ] 0.268
- [ ] 0.5
- [x] 0.731

  As calculated in #1 above, the log-odds for the input values is 1. Plugging that value for *z* into the sigmoid function:

  $$y = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-1}} = \frac{1}{1+0.367} = \frac{1}{1.367} = 0.731$$

  Correct answer.

- [ ] 1

1. Imagine a phishing or malware classification model where phishing and malware websites are in the class labeled 1 (true) and harmless websites are in the class labeled 0 (false). This model mistakenly classifies a legitimate website as malware. What is this called?

A false positive ✓

A negative example (legitimate site) has been wrongly classified as a positive example (malware site).

Correct answer.

A true positive ☐

A true negative ☐

A false negative ☐

2. In general, what happens to the number of false positives when the classification threshold increases? What about true positives? Experiment with the slider above.

Both true and false positives increase. ☐

True positives increase. False positives decrease. ☐

Both true and false positives decrease. ✓

As the threshold increases, the model will likely predict fewer positives overall, both true and false. A spam classifier with a threshold of .9999 will only label an email as spam if it considers the classification to be at least 99.99% likely, which means it is highly unlikely to mislabel a legitimate email, but also likely to miss actual spam email.

Correct answer.

Exercise: Check your understanding

A model outputs 5 TP, 6 TN, 3 FP, and 2 FN. Calculate the recall.

- [x] 0.714

  Recall is calculated as $\frac{TP}{TP+FN} = \frac{5}{7}$.

  Correct answer.

- [ ] 0.625
- [ ] 0.455

A model outputs 3 TP, 4 TN, 2 FP, and 1 FN. Calculate the precision.

- [ ] 0.429
- [ ] 0.75
- [x] 0.6

  Precision is calculated as $\frac{TP}{TP+FP} = \frac{3}{5}$.

  Correct answer.

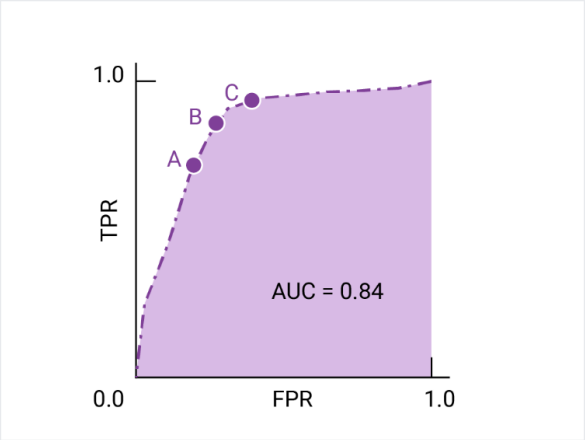
You're building a binary classifier that checks photos of insect traps for whether a dangerous invasive species is present. If the model detects the species, the entomologist (insect scientist) on duty is notified. Early detection of this insect is critical to preventing an infestation. A false alarm (false positive) is easy to handle: the entomologist sees that the photo was misclassified and marks it as such. Assuming an acceptable accuracy level, which metric should this model be optimized for?

- [ ] False positive rate (FPR)
- [ ] Precision
- [x] Recall

  In this scenario, false alarms (FP) are low-cost, and false negatives are highly costly, so it makes sense to maximize recall, or the probability of detection.

  Correct answer.

Imagine a situation where it's better to allow some spam to reach the inbox than to send a business-critical email to the spam folder. You've trained a spam classifier for this situation where the positive class is spam and the negative class is not-spam. Which of the following points on the ROC curve for your classifier is preferable?



Point A ✓

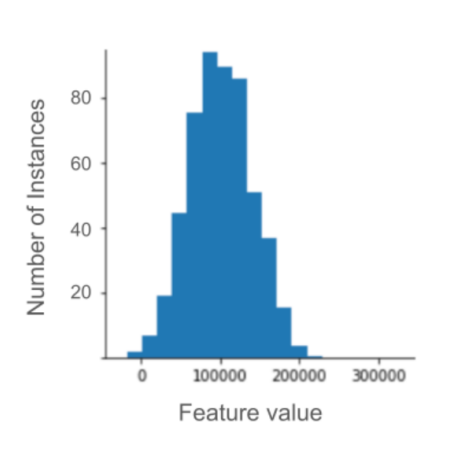
In this use case, it's better to minimize false positives, even if true positives also decrease.

Correct answer.

Point B ☐

Point C ☐

Which technique would be most suitable for normalizing a feature with the following distribution?



- [ ] Clipping
- [x] Z-score scaling

  The data points generally conform to a normal distribution, so Z-score scaling will force them into the range –3 to +3.

  Correct answer.
- [ ] Log scaling
- [ ] Linear scaling

Suppose you are developing a model that predicts a data center's productivity based on the temperature measured inside the data center. Almost all of the `temperature` values in your dataset fall between 15 and 30 (Celsius), with the following exceptions:

- Once or twice per year, on extremely hot days, a few values between 31 and 45 are recorded in `temperature`.
- Every 1,000th point in `temperature` is set to 1,000 rather than the actual temperature.

Which would be a reasonable normalization technique for `temperature`?

- [ ] Delete the outlier values between 31 and 45, but clip the outliers with a value of 1,000.
- [x] Clip the outlier values between 31 and 45, but delete the outliers with a value of 1,000

  The values of 1,000 are mistakes, and should be deleted rather than clipped.

  The values between 31 and 45 are legitimate data points. Clipping would probably be a good idea for these values, assuming the dataset doesn't contain enough examples in this temperature range to train the model to make good predictions. However, during inference, note that the clipped model would therefore make the same prediction for a temperature of 45 as for a temperature of 35.

  Correct answer.
- [ ] Clip all the outliers
- [ ] Delete all the outliers

Then explore the model, and use it to answer the following questions.

How many parameters (weights and biases) does this neural network model have?

- 4 ☐
- 12 ☐
- 16 ☐
- 21 ✓

There are 4 parameters used to calculate each of the 4 node values in the hidden layer—3 weights (one for each input value) and a bias—which sums to 16 parameters. Then there are 5 parameters used to calculate the output value: 4 weights (one for each node in the hidden layer) and a bias. In total, this neural network has 21 parameters.

Correct answer.

Try modifying the model parameters, and observe the effect on the hidden-layer node values and the output value (you can review the Calculations panel below to see how these values were calculated).

Can this model learn nonlinearities?

- Yes ☐
- No ✓

If you click on each of the nodes in the hidden layer and review the calculations below, you'll see that all of them are linear (comprising multiplication and addition operations).

If you then click on the output node and review the calculation below, you'll see that this calculation is also linear. Linear calculations performed on the output of linear calculations are also linear, which means this model cannot learn nonlinearities.

Correct answer.

Exercise: Check your understanding

Which language model makes better predictions for English text?

- **A language model based on 6-grams**
- **A language model based on 5-grams**

- The language model based on 6-grams. ☐
- The language model based on 5-grams. ☐
- The answer depends on the size and diversity of the training set. ✓

If the training set spans millions of diverse documents, then the model based on 6-grams will probably outperform the model based on 5-grams.

Correct answer.

Exercise: Check your understanding

Which of the following statements is true about LLMs?

- [ ] A fine-tuned LLM contains fewer parameters than the foundation language model it was trained on.
- [x] A distilled LLM contains fewer parameters than the foundation language model it sprung from.

  Yes, distillation reduces the number of parameters.

  Correct answer.
- [ ] As users perform more prompt engineering, the number of parameters in an LLM grows.

Exercise: Check your understanding

Which language model makes better predictions for English text?

- **A language model based on 6-grams**
- **A language model based on 5-grams**

- [ ] The language model based on 5-grams.
- [x] The answer depends on the size and diversity of the training set.

  If the training set spans millions of diverse documents, then the model based on 6-grams will probably outperform the model based on 5-grams.

  Correct answer.
- [ ] The language model based on 6-grams.