Datasets, generalization, and overfitting: Test Your Knowledge

Let's do a quick test! You must answer at least 4 questions correctly to pass this quiz.

1. Which of the following is an example of a stationary dataset?

- ○ Rainfall rates in Seattle, Washington
- ◉ Catalog of musical notes in Beethoven's symphonies
  - ✓ The set of musical notes used in Beethoven's symphonies is static and will not change over time.
- ○ Most frequently used words in spam emails
- ○ Daily ticket sales at a movie theater

2. You are preparing to train a model for predicting the sale price of used cars using a dataset that contains the following features: `year`, `model`, and `mileage`. When you examine the dataset, you discover that mileage values are missing for 150 out of 2,500 examples. Which of the following options would be reasonable actions to take? (Choose all that apply)

Choose as many answers as you see fit.

- ☑ Removing those 150 examples from the dataset
  - ✓ It is reasonable to remove the examples that are missing mileage values.
- ☐ Removing the `mileage` column from the dataset, and training the model on just year and model
- ☐ Inserting a value of `0` into all the empty mileage fields
- ☑ Inferring a `mileage` estimate for each empty field by multiplying the age of the car by the average yearly mileage for all cars in the dataset
  - ✓ It is reasonable to impute these missing values with an estimate that can be logically derived from other data.

3. You are training a streaming service's movie-recommendation model to predict whether or not a user will enjoy a given movie. Which of the following would be reasonable proxy labels for "User enjoyed the movie"? (Choose all that apply)

Choose as many answers as you see fit.

- ☐ User saved the movie to their "Want to Watch" list.
- ☐ User clicked "Start watching movie."
- ☑ User recommended the movie to another user.
  - ✓ If a user recommends a movie, that's also likely a good signal that they enjoyed it.
- ☑ User gave the movie a 5-star rating.
  - ✓ This is a reasonable proxy label, because a high rating typically correlates with user enjoyment.

4. True or False: Training your model until it achieves a low loss value on your test data is a good way to prevent overfitting.

- ○ True
- ◉ False
  - ✓ Training your model multiple times until it performs well on the test data may actually cause overfitting, as good results on the test set will no longer demonstrate the model's ability to perform well on never-before-seen data.

5. Fill in the blank in the following sentence:
Regularization improves your model's ability to generalize to new data by penalizing ___ during training.

- ○ Incorrect predictions
- ○ Learning rate
- ◉ Complexity
  - ✓ Regularization can improve a model's ability to generalize to new data by penalizing complexity, which encourages the model to fit the data as simply as possible.
- ○ Gradient descent

Results

You scored 5 out of 5. Congratulations! You have passed this quiz.

You've unlocked a new milestone: **Machine Learning Crash Course: Datasets, generalization, and overfitting!**

Sign in to claim your badge.

Share Sign in

Was this helpful?