Hong Hanh

Practical Data Science Specialization

Week 3

Practice Quiz • 30 min

1. What are some of the challenges present in data labelling? (Select all that apply)

1 / 1 point

- [x] Scale - machine learning algorithms need large, labeled datasets, ranging in count from hundreds to many thousands of examples, depending on the application.

Correct
Correct! A machine learning algorithm needs a great deal of data to be trained effectively.

- [x] It's time consuming to build a training dataset. It is a time-intensive process to label data - especially if it's a complex task, such as drawing bounding boxes around particular objects within images.

Correct
You're right! Often, you'll need to gather the help of others to label your training data in a reasonable period of time.

- [ ] Data labelling is a straightforward process and does not have any significant challenges.

- [x] Highly accurate data is required to train machine learning algorithms. It can be a challenge to obtain accurately and consistently labelled datasets, especially when you have a number of people and/or processes labelling your data.

Correct
Great work! That's right, if the data is not highly accurate, it's unreasonable to expect your algorithm to be highly accurate. It takes time and, in some cases, expertise in a particular domain to produce large amounts of highly accurate data.

2. Data scientists often spend 80% of their time on data preparation tasks such as finding, cleaning, and labeling data. The remaining 20% of their time is often spent actually developing the models and deriving insights. Which of the following can help a data scientist label data more efficiently to break this 80/20 rule? (Select all that apply)

1 / 1 point

- [x] Active learning is a feature of automated data labeling which learns how to automatically label new data based on previous data which was labeled manually by humans. The actively-learned labeling model can automatically apply labels to new data and greatly reduce the amount of manual human effort.

Correct
Right on! You can save your labeling team a lot of time by using active learning to automatically label a large portion of your training dataset.

- [ ] Try different models, oftentimes people use the wrong model for the job. Once you find the right model, you don't need to spend as much time perfecting your data.

- [x] Human-in-the-loop workflows, where machine learning does the heavy lifting of assigning a label (or labels), and the human is only involved in reviewing and approving the final results.

Correct
That's right! If a machine learning algorithm is able to accurately complete the task most of the time, it's more resource-efficient for the person to simply review the work and make a relatively-small number of corrections.

Hong Hanh

Practical Data Science Specialization

Week 3

Practice Quiz • 30 min

3. Which of the following statements are true about Amazon SageMaker Ground Truth? (Select all that apply)

1 / 1 point

- [x] When creating custom labelling tasks using Amazon SageMaker Ground Truth, you can create AWS Lambda functions that run before and after each data object is sent to the worker.

Correct

That's right! For example, you can pass an S3 URI to a Lambda function, which will then extract the text and transform it into the `taskInput` JSON format required by Amazon SageMaker Ground Truth.

- [x] When setting up your labelling task, Amazon SageMaker Ground Truth has several built-in task types which come with pre-built worker task templates (ex. Semantic Segmentation for Image Data).

Correct

You got it! There are a variety of built-in task types available for image, video, and text data.

- [x] Amazon SageMaker Ground Truth is a data labelling service where you provide the input dataset, define the labelling task, and recruit a human workforce to crowdsource labeling the dataset.

Correct

That's right! Ground Truth is a fully managed data labeling service that makes it easy to build highly accurate training datasets for machine learning.

- [ ] When selecting a human workforce to label your data, one option is to use a private workforce which, by default, restricts the IP addresses allowed to connect to the worker portal.

4. Your company is crowdsourcing the task of labelling a dataset. They are providing labellers with the following instructions.

1 / 1 point

We are building a machine learning algorithm to identify dogs in images. Please place a bounding box around any dogs you see in the images.

An example of an image from the dataset can be seen below:

Hong Hanh

Practical Data Science Specialization

What concerns might you have about the quality of labels that result?

- [x] The instructions don't explain what to do if there are overlapping dogs. Do you draw a bounding box only around the part of the dog that you can visually see, or do you extend the box to include parts of the dog not visible from the current view point?

Correct
You're right! The instruction is ambiguous in this way. Labellers will end up with differently positioned bounding boxes around each dog. Such discrepancies between labels will result in degraded performance of the machine learning algorithm.

- [x] The instructions are not specific enough - some labellers may draw one bounding box around all four dogs, while others will draw four bounding boxes, one around each dog. Inconsistencies between labelling methods may cause degraded performance.

Correct
That's correct! If your labelling instructions are ambiguous, each labeller may interpret them differently, resulting in inconsistently labelled data. Machine learning models learn from labelled examples, and will

- [ ] The instructions are adequately clear and you expect every labeller to have close to identically placed bounding boxes for each image. Once all of the labelling is complete, the model will have a collection of consistently labelled samples to learn from.

5. Good machine learning models are built on large volumes of high-quality training data but the process of creating this data can be costly, complex, and time consuming. What are some techniques that can be used to improve the efficiency and accuracy of data labeling? (Select all that apply)

1 / 1 point

- [x] Consolidate annotations across multiple labellers and require consensus across the provided labels

Correct
That's right! To prevent error or bias from individual annotators, labeler consensus is frequently used.

- [ ] Avoid automated data labeling
- [x] Clear communication and concise instructions to labelers

Correct
You're right! High accuracy can be achieved by streamlining tasks and communicating clear instructions..

- [x] Auditing labels

6. Amazon A2I makes it easy to build and manage human reviews for machine learning applications.

1 / 1 point

True/False: In Amazon A2I, machine learning predictions with high-confidence scores are sent for human review while those with low-confidence scores are returned directly to the client application.

- ( ) True
- (•) False

Correct
That's right! Amazon A2I allows humans to step in when the model is unable to make a high confidence prediction. If needed, the human reviewers correct the predictions. The consolidated predictions across all labelers is then returned to the client application. These updated predictions can also be used to re-train and improve the model.