OXMEDICA
Driving Global Education

Dataset Division

Overview

1. Recap: What is a dataset?
2. Training datasets
3. Why is this story incomplete?
4. Test datasets
5. Validation datasets and hyperparameter tuning

Datasets: Recap

Discussion

- What is a dataset?
- What does it look like?
- How do we define it mathematically?
- Why do we need data to train models?

Training datasets

Task in pairs

- Think of a supervised machine learning problem you might want to solve. **[It can’t be anything we’ve mentioned so far]**

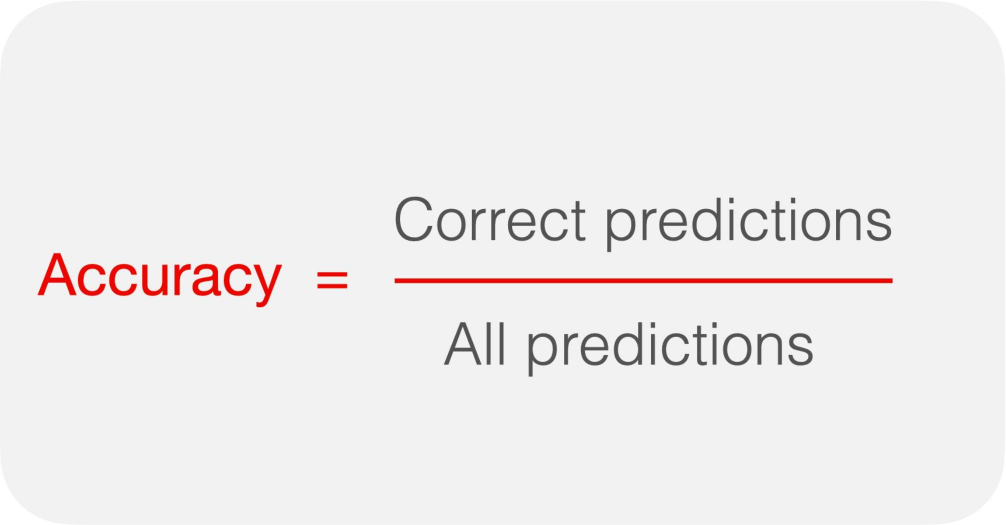
- What data would you need to have to train this model?

- How might you go about collecting this data?

Evaluating Model Performance

Evaluating model performance

- We want to use some metric to evaluate our model’s performance.
- The most common metric with classification problems is **accuracy**. I.e. the percentage of predictions which are correct.



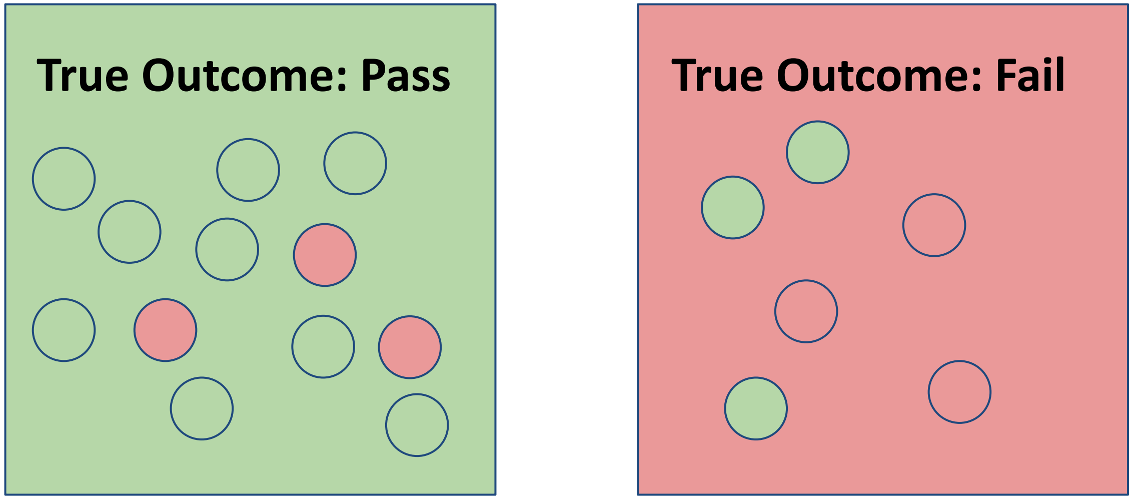
$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

Source

What is the accuracy here?

Machine Learning Problem: Predicting whether students will pass or fail a course.

○ = model predictions

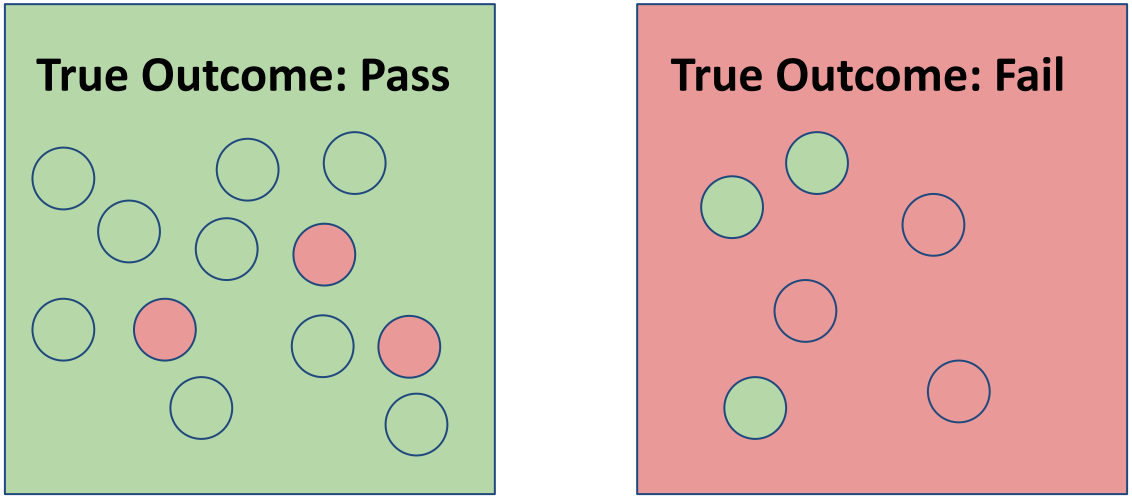


True Outcome: Pass

True Outcome: Fail

Source

What is the accuracy here?



Number of correct predictions = 12
Number of incorrect predictions = 6
Total predictions = 18

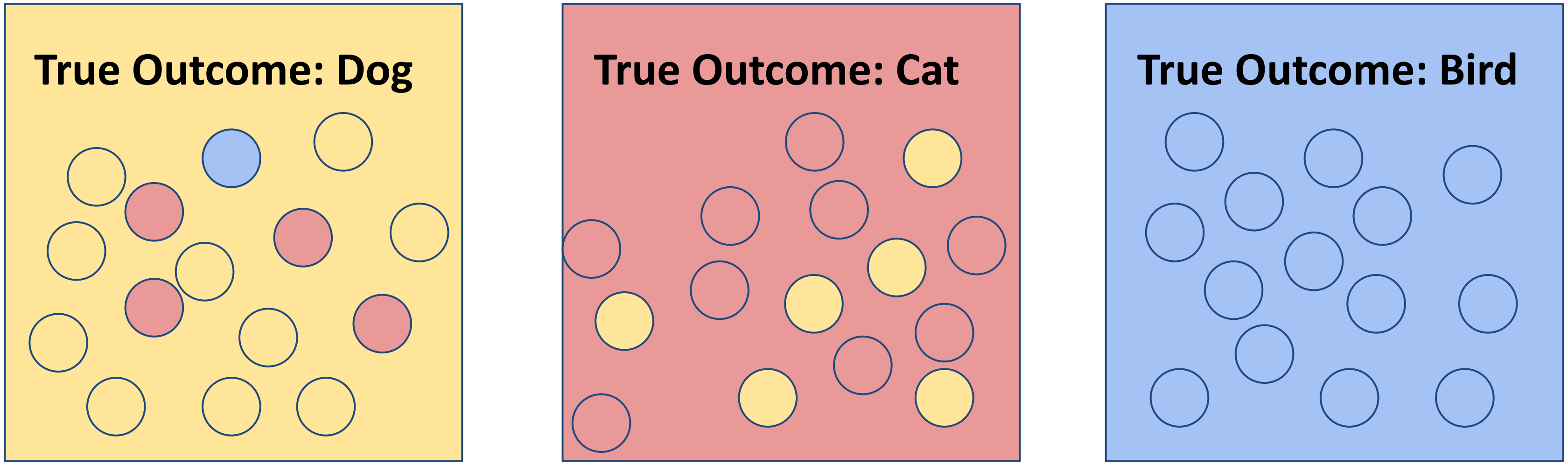
Accuracy = correct predictions/total predictions = 12/18 = 66.67%

Source

What is the accuracy here?

Machine Learning Problem: Predicting whether an image of an animal is a dog, cat or bird.

○ = model predictions



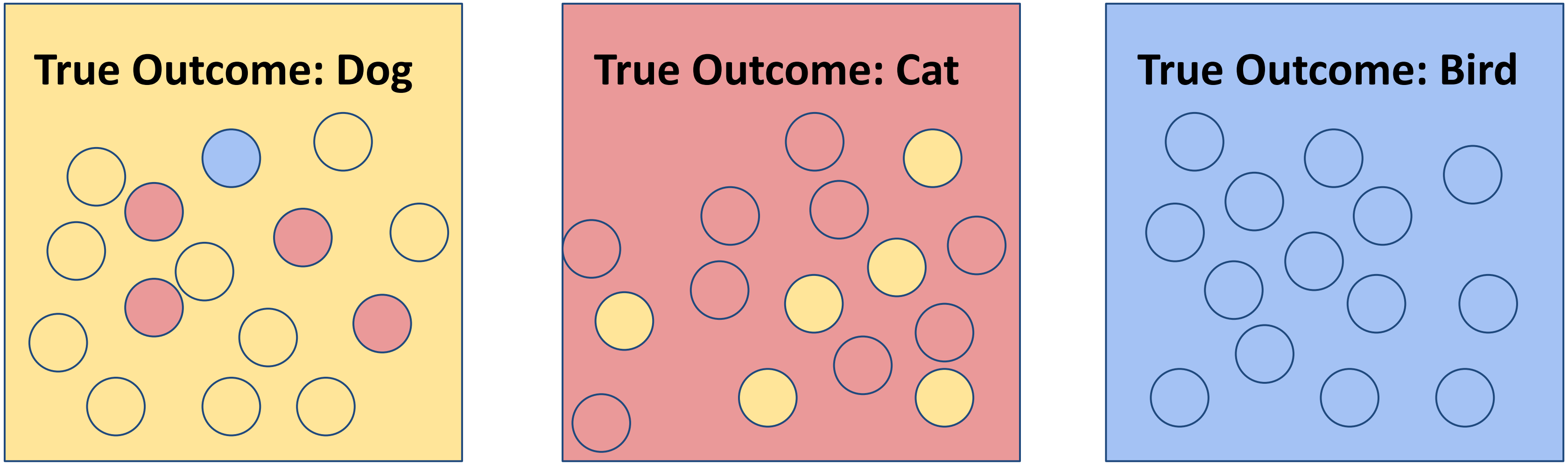
True Outcome: Dog

True Outcome: Cat

True Outcome: Bird

Source

What is the accuracy here?

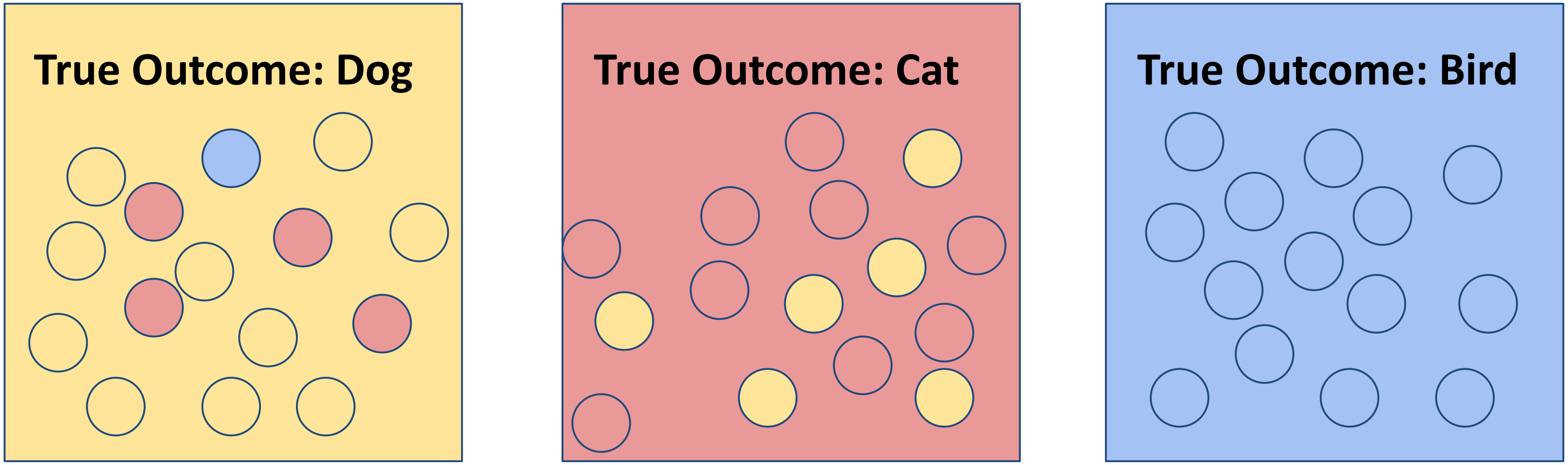


Number of correct predictions = 10 + 9 + 14 = 33
Number of incorrect predictions = 5 + 6 + 0 = 11
Total predictions = 44

Accuracy = correct predictions/total predictions = 33/44 = 75%

Source

What is the accuracy here?



- What do you notice about this classification model?
- Why might this be the case?
- Given this, is accuracy a good metric of the model's performance?

Source

Training and Test Datasets

Training datasets

Discussion

Should we evaluate model performance on this training dataset?

(**NOTE**: there are valid arguments both ways)

Training datasets

✅ It is an indication of whether our model has learnt from the training data. It tells us whether training is working!

❌ It might misrepresent our model's performance on unseen data, which is ultimately all we care about.

Training datasets

Discussion

Why might it misrepresent the model’s performance on unseen data?
(hard)

Test datasets

- Solution!
- Instead have a held-out dataset we call the **‘test’ dataset**
- The model is not trained on this dataset! It is only used for model evaluation

- Is this a good idea?
- What are the downsides of this approach?

[Extra:] Hyperparameters and Validation Dataset

Hyperparameters and Validation Datasets

See whiteboard

Recap questions

1. Why do we need to evaluate models?
2. What is accuracy and how do we calculate it?
3. What is wrong with evaluating on the training data?
4. How do we get around this?
5. What are the downsides of this solution?
6. What are hyperparameters?
7. Why might we want to use a validation (val) dataset?