← Back Bird Recognition in the City of Peacetopia (Case Study)

Graded Quiz • 45 min Due Feb 14, 2:59 AM EST

Fellow Learners on Discourse! Setting Up your Goal Comparing to Human-level Performance Lecture Notes (Optional) Machine Learning Flight Simulator Reading: Machine Learning Flight
Simulator Quiz: Bird Recognition in the City of Peacetopia (Case Study)
 15 questions Heroes of Deep Learning (Optional)

Grade Recognition in the City of Peacetopia (Case

Quiz • 45 min This example adapted from a real production application, but with details disguised to protect confidentiality. Submit your assignment 91.66% We keep your highest score

You are a famous researcher in the City of Peacetopia. The people of Peacetopia have a common characteristic:

overPeacetopia and alert the population. The City Council gives you a dataset of 10,000,000 images of the sky above Peacetopia, taken from the city's security cameras. They are labelled: • y = 0: There is no bird on the image y = 1: There is a bird on the image

they are afraid of birds. To save them, you have **to build an algorithm that will detect any bird flying**

Your goal is to build an algorithm able to classify new images taken by security cameras from Peace topia.There are a lot of decisions to make: What is the evaluation metric?

Metric of success The City Council tells you that they want an algorithm that Has high accuracy 2. Runs quickly and takes only a short time to classify a new image. 3. Can fit in a small amount of memory, so that it can run in a small processor that the city will attach to many

Note: Having three evaluation metrics makes it harder for you to quickly choose between two different algorithms, and will slow down the speed with which your team can iterate. True/False? True ○ False

How do you structure your data into train/dev/test sets?

different security cameras.

2. After further discussions, the city narrows down its criteria to: • "We need an algorithm that can let us know a bird is flying over Peacetopia as accurately as possible." • "We want the trained model to take no more than 10sec to classify a new image." "We want the model to fit in 10MB of memory."

If you had the three following models, which one would you choose? ○ B: Test Accuracy Memory size Test Accuracy Runtime Memory size O Test Accuracy Runtime Memory size 97% O Test Accuracy 97%

 Correct
 Correct! As soon as the runtime is less than 10 seconds you're good. So, you may simply maximize the test
 accuracy after you made sure the runtime is <10sec.

3. Based on the city's requests, which of the following would you say is true?

Accuracy is a satisficing metric; running time and memory size are an optimizing metric. Accuracy, running time and memory size are all optimizing metrics because you want to do well on all three. Accuracy is an optimizing metric; running time and memory size are a satisficing metrics. Accuracy, running time and memory size are all satisficing metrics because you have to do sufficiently well on all three for your system to be acceptable.

1 / 1 point

4. Structuring your data Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of these do you

think is the best choice? 9,500,000 3,333,334 3,333,333 3,333,333 O Train 6,000,000 O Train 1,000,000 6,000,000 Correct Yes.

5. After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the 1/1 point "citizens' data". Apparently the citizens of Peacetopia are so scared of birds that they volunteered to take pictures of the sky and label them, thus contributing these additional 1,000,000 images. These images are different from $\,$ $the\ distribution\ of\ images\ the\ City\ Council\ had\ originally\ given\ you,\ but\ you\ think\ it\ could\ help\ you\ algorithm.$ Notice that adding this additional data to the training set will make the distribution of the training set different from the distributions of the dev and test sets.

Is the following statement true or false? "You should not add the citizens' data to the training set, because if the training distribution is different from the dev and test sets, then this will not allow the model to perform well on the test set." ○ True False

⊘ Correct False is correct: Sometimes we'll need to train the model on the data that is available, and its distribution may not be the same as the data that will occur in production. Also, adding training data that differs from the dev set may still help the model improve performance on the dev set. What matters is that the dev and test set have the same distribution.

6. One member of the City Council knows a little about machine learning, and thinks you should add the 1,000,000 1/1 point citizens' data images to the test set. You object because: A bigger test set will slow down the speed of iterating because of the computational expense of evaluating

models on the test set. ☐ The 1,000,000 citizens' data images do not have a consistent x-->y mapping as the rest of the data (similar to the New York City/Detroit housing prices example from lecture). This would cause the dev and test set distributions to become different. This is a bad idea because you're not aiming where you want to hit. **⊘** Correct

The test set no longer reflects the distribution of data (security cameras) you most care about. **⊘** Correct

Correct

✓ Correct

⊘ Correct

You should try to get a bigger dev set.

You should get a bigger test set. You have underfit to the dev set.

✓ Correct

⊘ Correct

 $ig\otimes$ Incorrect

Correct

7. You train a system, and its errors are as follows (error = 100%-Accuracy): 1 / 1 point Training set error

This suggests that one good avenue for improving performance is to train a bigger network so as to drive down the 4.0% training error. Do you agree? O No, because this shows your variance is higher than your bias. Yes, because having 4.0% training error shows you have a high bias. No, because there is insufficient information to tell. Yes, because this shows your bias is higher than your variance.

8. You ask a few people to label the dataset so as to find out what is human-level performance. You find the following levels of accuracy: Bird watching expert #1 Bird watching expert #2 0.5% error Normal person #1 (not a bird watching expert) 1.0% error Normal person #2 (not a bird watching expert) 1.2% error If your goal is to have "human-level performance" be a proxy (or estimate) for Bayes error, how would you define O.4% (average of 0.3 and 0.5)

O.75% (average of all four numbers above) 0.3% (accuracy of expert #1) 0.0% (because it is impossible to do better than this) ✓ Correct

9. Which of the following statements do you agree with? A learning algorithm's performance can be better than human-level performance and better than Bayes A learning algorithm's performance can never be better than human-level performance nor better than A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.

O A learning algorithm's performance can never be better than human-level performance but it can be better

10. You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so 1/1 point you define that as "human-level performance." After working further on your algorithm, you end up with the following: Human-level performance 2.0% Training set error Dev set error 2.1% Based on the evidence you have, which two of the following four options seem the most promising to try? (Check two options.) Try decreasing regularization.

☐ Try increasing regularization. Train a bigger model to try to do better on the training set. **⊘** Correct ☐ Get a bigger training set to reduce variance.

11. You also evaluate your model on the test set, and find the following: 0.1% Human-level performance 2.0% 2.1% 7.0%

Test set error What does this mean? (Check the two best options.) You have overfit to the dev set. ✓ Correct

12. After working on this project for a year, you finally achieve: 0.10% Human-level performance Training set error What can you conclude? (Check all that apply.)

1 / 1 point

 $\begin{tabular}{ll} \hline & With only 0.09\% further progress to make, you should quickly be able to close the remaining gap to 0\% \\ \hline \\ \hline \end{tabular}$ ☐ This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance. lacksquare If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤ 0.05 ✓ Correct It is now harder to measure avoidable bias, thus progress will be slower going forward.

13. It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor's systems, they conclude they actually like your competitor's system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do? Ask your team to take into account both accuracy and false negative rate during development. Rethink the appropriate metric for this task, and ask your team to tune to the new metric. Look at all the models you've developed during the development process and find the one with the lowest false negative error rate. O Pick false negative rate as the new metric, and use this new metric to drive all further development.

14. You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the 1/1 point citizens from birds! But over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your data is being tested on a new type of data.



You have only 1,000 images of the new species of bird. The city expects a better system from you within the next 3 months. Which of these should you do first? O Put the 1,000 images into the training set so as to try to do better on these birds. Try data augmentation/data synthesis to get more images of the new type of bird. O: Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split. Use the data you have to define a new evaluation metric (using a new dev/test set) taking into account the new species, and use that to drive further progress for your team.

15. The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your 0.75 / 1 point work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly $useful\ aren't\ they.)\ Because\ of\ years\ of\ working\ on\ Cat\ detectors, you\ have\ such\ a\ huge\ dataset\ of\ 100,000,000\ cat$ images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.) ☐ Needing two weeks to train will limit the speed at which you can iterate.

Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.