Structuring Machine Learning Projects

Week 1: Bird Recognition in the City of Peacetopia (Case Study)

1 Version 1: You are delighted because this list of criteria will speed development and provide guidance on how to evaluate two different algorithms. True/False?

Answer: False.

Comment: More than one metric expands the choices and tradeoffs you have to decide for each with unknown effects on the other two.

1 Version 2: 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.

Answer: True.

2 Version 1: Given models with different accuracies, runtimes, and memory sizes, how would you choose one?

Answer: Find the subset of models that meet the runtime and memory criteria. Then choose the highest accuracy.

Comment: Once you meet the runtime and memory thresholds, accuracy should be maximized.

2 Version 2: If you had the three following models. which one would you choose?

Answer: 98%, 9 sec, 9MB.

Comment: 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 < 10 sec.

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

Answer: Accuracy is an optimizing metric; running time and memory size are satisfying metrics.

4 With 10,000,000 data points, what is the best option for train/dev/test splits?

Answer: 95, 2.5, 2.5.

Comment: The size of the data set allows for bias and variance evaluation with smaller data sets.

5 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?

Answer: False.

Comment: 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 citizens' data images proportionately to the train/dev/test sets. You object because:

Answer: If we add the images to the test set then it won't reflect the distribution of data expected in production.

Comment: Using the data in the training set could be beneficial, but you wouldn't want to include such images in your test set as they are not from the expected distribution of data you'll see in production.

7 You train a system, and its error are as follows: training set error 4.0, Dev set error 4.5. 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?

Answer: No, because there is insufficient information to tell.

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 0.3 error, Bird watching expert 2 0.5 error, Normal person 1 1.0 error, Normal person 2 1.2 error. If your goal is to have "human-level performance" be a proxy (or estimate) for Bayes error, how would you define "human-level performance"?

Answer: 0.3 (accuracy of expert 1).

9 Which of the following statements do you agree with?

Answer: A learning algorithm's performance can be better than human level performance but it can never be better than Bayes error.

10 You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so you define that as "human-level performance." After working further on your algorithm, you end up with the following: Human level performance 0.1, Training set error 2.0, 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.)

Answer: Train a bigger model to try to do better on the training set; Try decreasing regularization.

11 You also evaluate your model on the test set, and find the following: Human level performance 0.1, Training set error 2.0, Dev set error 2.1, Test set error 7.0. What does this mean? (Check the two best options.)

Answer: You have overfit to the dev set; You should try to get a bigger dev set.

12 After working on this project for a year, you finally achieve: Human level performance 0.1, Training set error, 0.05, Dev set error 0.05. What can you conclude? (Check all that apply)

Answer: If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤ 0.05; 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. You and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! Still, when Peacetopia tries out both systems, they conclude they 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?

Answer: Brainstorm with your team to refine the optimizing metric to include false negatives as they further develop the model.

Comment: The target has shifted so an updated metric is required.

14 You've handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the 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 model is being tested on a new type of data. There are only 1,000 images of the new species. The city expects a better system from you within the next 3 months. Which of these should you do first?

Answer: 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 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.)

Answer: Buying faster computers could speed up your teams' iteration speed and thus your team's productivity; Needing two weeks to train will limit the speed at which you can iterate; If 100,000,000 examples is enough to build a good enough Cat detector, you might be better off training with just 10,000,000 examples to gain a $\approx$ 10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it's trained on less data.