**1.**This example is adapted from a real production application, but with details disguised to protect confidentiality.



You are a famous researcher in the City of Peacetopia. The people of Peacetopia have a common characteristic: they are afraid of birds. To save them, you have **to build an algorithm that will detect any bird flying over Peacetopia** 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 labeled:

* y = 0: There is no bird on the image
* y = 1: There is a bird on the image

Your goal is to build an algorithm able to classify new images taken by security cameras from Peacetopia.

There are a lot of decisions to make:

* What is the evaluation metric?
* How do you structure your data into train/dev/test sets?

**Metric of success**

The City Council tells you that they want an algorithm that

1. 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 different security cameras.

**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

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

FALSE

**You meet with them and ask for just one evaluation metric. True/False?**

FALSE

### 2.The city revises 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 10 sec to classify a new image.”
* “We want the model to fit in 10MB of memory.”

**Given models with different accuracies, runtimes, and memory sizes, how would you choose one?**

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

**If you had the three following models, which one would you choose?**

|  |  |  |
| --- | --- | --- |
| Test Accuracy | Runtime | Memory size |
| 98% | 9 sec | 9MB |







### 2. The city asks for your help in further defining the criteria for accuracy, runtime, and memory. How would you suggest they identify the criteria?

Suggest to them that they define which criterion is most important. Then, set thresholds for the other two.

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

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

### 3.The essential difference between an optimizing metric and satisficing metrics is the priority assigned by the stakeholders. True/False?

FALSE

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

Accuracy is an optimizing metric; running time and memory size are a satisficing metrics.

**3. Which of the following best answers why it is important to identify optimizing and satisficing metrics?**

Identifying the metric types sets thresholds for satisficing metrics. This provides explicit evaluation criteria

### 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?**

* + Train 9,500,000 Dev 250,000 Test 250,000

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

train - 95%, dev - 2.5%, test - 2.5%

### 5. Now that you’ve set up your train/dev/test sets, the City Council comes across another 1,000,000 images from social media and offers them to you. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. Which of the following is the best use of that additional data?



Add it to the training set.

### 5.After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the “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 your 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."

FALSE

**5.After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the “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 your algorithm.**

**You should not add the citizens’ data to the training set, because this will cause the training and dev/test set distributions to become different, thus hurting dev and test set performance. True/False?**

FALSE

**5.Now that you’ve set up your train/dev/test sets, the City Council comes across another 1,000,000 images from social media and offers them to you. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm. You should add the citizens’ data to the training set. True/False?**

FALSE

**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:**

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

**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 to the test set. You object because:**

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

### 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 to the test set. You object because:

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.

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

### 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 to the dev set. You object because: (Choose all that apply)

- The dev set no longer reflects the distribution of data (security cameras) you most care about..

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.

### 7.You train a system, and its errors are as follows (error = 100%-Accuracy):

|  |  |
| --- | --- |
| 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?



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 (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 “human-level performance”?



0.3% (accuracy of expert #1)

### 8.You want to define what human-level performance is to the city council. Which of the following is the best answer?



The performance of their best ornithologist (0.3%).

### 9.Which of the below shows the optimal order of accuracy from worst to best?

Human-level performance -> the learning algorithm’s performance -> Bayes error.

**9. Which of the following statements do you agree with?**

* + A learning algorithm’s performance can be better human-level performance but it can never be better than Bayes error.

### 10.After working on your algorithm you have to decide the next steps. Currently, human-level performance is 0.1%, training is at 2.0% and the dev set is at 2.1%. Which statement below best describes your thought process?



Address bias first through a larger model to get closest to human level error.



Decrease regularization to boost smaller signals.

1. **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.)**

* + Try decreasing regularization.
  + Train a bigger model to try to do better on the training set.

### 10.Which of the following best expresses how to evaluate the next steps in your project when your results for human-level performance, train, and dev set error are 0.1%, 2.0%, and 2.1% respectively?

Based on differences between the three levels of performance, prioritize actions to decrease bias and iterate

1. **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.)**

* + You should try to get a bigger dev set.
  + You have overfit to the dev set.

### 11.You’ve now also run your model on the test set and find that it is a 7.0% error compared to a 2.1% error for the dev set. What should you do? (Choose all that apply)

Try increasing regularization to reduce overfitting to the dev set.

Increase the size of the dev set.

### 12.After working on this project for a year, you finally achieve:

|  |  |
| --- | --- |
| Human-level performance | 0.10% |
| Training set error | 0.05% |
| Dev set error | 0.05% |

What can you conclude? (Check all that apply.)

If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is ≤0.05≤0.05

It is now harder to measure avoidable bias, thus progress will be slower going forward.

### After working on this project for a year, you finally achieve: Human-level performance, 0.10%, Training set error, 0.05%, Dev set error, 0.05%. Which of the following are likely? (Check all that apply.)

Pushing to even higher accuracy will be slow because you will not be able to easily identify sources of bias.

The model has recognized emergent features that humans cannot. (Chess and Go for example)

### 12.After working on this project for a year, you finally achieve: Human-level performance, 0.10%, Training set error, 0.05%, Dev set error, 0.05%. Which of the following are true? (Check all that apply.)

You are close to Bayes error and possible overfitting.

All or almost all of the avoidable bias has been accounted for.

### 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?

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

1. **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?**
   * Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

### 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 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?

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. You have a huge dataset of 100,000,000 cat images. Training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

Accuracy should exceed the City Council’s requirements but the project may take as long as the bird detector because of the two week training/iteration time.

You could consider a tradeoff where you use a subset of the cat data to find reasonable performance with reasonable iteration pacing.

Given a significant budget for cloud GPUs, you could mitigate the training time.

### 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.)

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 ≈≈10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it’s trained on less data.

### 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. You have a huge dataset of 100,000,000 cat images. Training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

Lowering the number of images will reduce training time and likely allow for an acceptable tradeoff between iteration speed and accuracy.

This significantly impacts iteration speed.