**Week 1 Quiz - Bird recognition in the city of Peacetopia (case study)**

1. 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
2. If you had the three following models, which one would you choose?
   * Test Accuracy 98%
   * Runtime 9 sec
   * Memory size 9MB
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.
   *  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, 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.
4. 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
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?

* +  True
  +  False

1. 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.
2. 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.

1. 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)

1. 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.
2. 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.

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.

1. 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.)

* + It is now harder to measure avoidable bias, thus progress will be slower going forward.
    - If the test set is big enough for the 0,05% error estimate to be accurate, this implies Bayes error is ≤0.05

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.
2. 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.
   * 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.
3. 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.)
   * If 100,000,000 examples is enough to build a good enough Cat detector, you might be better of 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.
   * 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.