CS129: Applied Machine Learning
Welcome to Hogwarts

Case Study: Crow Detection at Hogwarts

This case study is adapted from a real production application, but with details disguised to protect confidentiality.

You are a famous wizard in Hogwarts. The residents of Hogwarts are afraid of crows because they believe they are spies sent by Voldemort. To save them, you have to build an algorithm that detects any crow flying over Hogwarts and alerts the population.

The City Council gives you a dataset of 10,000,000 images of the sky above Hogwarts, taken from the city's security cameras. They are labeled:

$$y = \begin{cases} 0 & \text{if no crow is present in the image,} \\ 1 & \text{if a crow is present in the image.} \end{cases}$$

Your goal is to build an algorithm able to classify new images taken by security cameras from Hogwarts. There are many decisions to make:

- What is the evaluation metric (i.e., how do you decide your model is good)?
- How do you structure your data into train/dev/test sets?
- How do you go about tuning your hyperparameters?
- And much more ...

1 Metric of Success

The City Council tells you that they want an algorithm that satisfies the following three criteria:

1. It has high accuracy.
2. It runs quickly and takes only a short time to classify a new image.
3. It can fit in a small amount of memory, so that it can run on a small processor attached to many different security cameras.

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.

Is the above statement **True** or **False**?
True

2 Designing the Right Evaluation Metric

After further discussion, the City Council provides additional information. Crows appear in only 0.5% of images. A false negative (failing to detect a crow) is 100 times more costly than a false positive (raising an alarm when there is no crow). The system automatically triggers an alarm whenever a crow is detected.

(a) Explain why classification accuracy may be a misleading metric in this setting.

Accuracy can be misleading because the classes are highly imbalanced (only 0.5% positives). A classifier can achieve very high accuracy by almost always predicting "no crow," while still failing at the core objective (detecting crows).

(b) Suppose your model predicts "no crow" for every image. What would its accuracy be? Is this model useful? Explain briefly. If the model always predicts $y = 0$, then it is correct on the 99.5% of images with no crow, so its accuracy is 99.5%. This model is not useful because it detects no crows (TP = 0 and recall = 0), which is unacceptable given the high cost of missed detections.

(c) Propose an evaluation metric that better reflects the City Council's objective. You may express it using quantities from the confusion matrix (TP, FP, FN, TN). Justify your choice. A suitable metric is a cost-weighted error that reflects the asymmetric costs:

$$\text{Cost} = 100 \cdot \text{FN} + 1 \cdot \text{FP},$$

optionally normalized by the number of examples m:

$$\text{AvgCost} = \frac{100 \cdot \text{FN} + \text{FP}}{m}.$$

This directly aligns evaluation with the stated objective that false negatives are far more costly than false positives.

(d) Suppose your classifier outputs probabilities $P(Y = 1 \mid X)$. Should you use a default decision threshold of 0.5? Explain how the asymmetric cost affects the optimal threshold. No. With asymmetric costs, the optimal threshold is generally much lower than 0.5. Predict $y = 1$ when the expected cost of predicting 1 is lower than predicting 0:

$$C_{\text{FP}}(1 - p) < C_{\text{FN}} p \quad \Longleftrightarrow \quad p > \frac{C_{\text{FP}}}{C_{\text{FP}} + C_{\text{FN}}}.$$

With $C_{\text{FN}} = 100$ and $C_{\text{FP}} = 1$, the threshold is

$$p > \frac{1}{101} \approx 0.0099,$$

i.e., around 1% rather than 50% (assuming probabilities are reasonably calibrated and the costs are accurate).

3 Choosing a Model Under Constraints

After further discussions, the city narrows down its criteria to the following:

1. "We need an algorithm that can let us know a crow is flying over Hogwarts as accurately as possible."
2. "We want the trained model to take no more than 10 seconds to classify a new image."
3. "We want the model to fit in 10MB of memory."

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

Model	Test Accuracy	Runtime	Memory size
A	97%	1 sec	3MB
B	99%	13 sec	9MB
C	97%	3 sec	2MB
D	98%	9 sec	9MB

D

4 Structuring Your Data

Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of the following splits do you think is the best choice?

A) Train: 6,000,000 Dev: 3,000,000 Test: 1,000,000

B) Train: 9,500,000 Dev: 250,000 Test: 250,000

C) Train: 6,000,000 Dev: 1,000,000 Test: 3,000,000

D) Train: 3,333,334 Dev: 3,333,333 Test: 3,333,333

B

5 Citizens' Data and Distribution Mismatch

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 Hogwarts are so scared of crows that they volunteered to take pictures of the sky and label them, thus contributing these additional 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.

Is the above statement **True** or **False**?

False Check if it helps or not. It may not help, but maybe it will help your model learn more about crows.

6 Where to Put the Citizens' Data

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:

A) This would cause the dev and test set distributions to become different. This is a bad idea because you are not aiming where you want to hit.

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

C) A bigger test set will slow down the speed of iterating because of the computational expense of evaluating models on the test set.

D) The 1,000,000 citizens' data images do not have a consistent $x \Rightarrow y$ mapping as the rest of the data.

A,B

7 Diagnosing Bias vs. Variance from Errors

You train a system, and its errors are as follows (Error = 100% − Accuracy):

$$\text{Training set error} = 4.0\% \qquad \text{Dev set error} = 4.5\%.$$

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?

A) Yes, because having 4.0% training error shows you have high bias.

B) Yes, because this shows your bias is higher than your variance.

C) No, because this shows your variance is higher than your bias.

D) No, because there is not enough information to tell.

D

8 Defining Human-Level Performance

Note: Bayes error is the lowest possible error rate for any classifier of a random outcome and is analogous to irreducible error. Human-level performance is the best outcome for any human being.

You ask a few people to label the dataset so as to estimate human-level performance (i.e., what accuracy a human would get at spotting crows). You find the following error rates:

1. Crow watching expert #1: 0.3% error
2. Crow watching expert #2: 0.5% error
3. Normal person #1: 1.0% error

4. 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"?

A) 0.0% (because it is impossible to do better than this)

B) 0.3% (error of expert #1)

C) 0.4% (average of 0.3% and 0.5%)

D) 0.75% (average of all four numbers above)

B

9 Human-Level Performance vs. Bayes Error

Which of the following statements do you agree with?

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

B) A learning algorithm's performance can never be better than human-level performance but it can be better than Bayes error.

C) A learning algorithm's performance can never be better than human-level performance nor better than Bayes error.

D) A learning algorithm's performance can be better than human-level performance and better than Bayes error.

A

10 Next Steps from Human-Level, Train, and Dev Errors

You find that a team of ornithologists debating and discussing an image gets an even better 0.1% error rate, 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 options seem the most promising to try? (*Check two.*)

A) Train a more complex model to try to do better on the training set.

B) Get a bigger training set to reduce variance.

C) Try increasing regularization.

D) Try decreasing regularization.

A,D

11 Interpreting Test Error Much Worse Than Dev Error

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.*)

A) You have overfit to the dev set.

B) You should try to get a bigger dev set.

C) You have underfit to the dev set.

D) You should get a bigger test set.

A,B

12 Surpassing Human-Level Performance

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.*)

A) This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.

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

C) With only 0.05% further progress to make, you should quickly be able to close the remaining gap to 0%.

D) If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is $\leq 0.05\%$.

B,D

13 Choosing the Right Metric When Errors Matter Differently

It turns out Hogwarts 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. When Hogwarts tries out your and your competitor's 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 crow is in the air).

What should you do?

A) Look at all the models you have developed during the development process and find the one with the lowest false negative error rate.

B) Ask your team to take into account both accuracy and false negative rate during development.

C) Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

D) Pick false negative rate as the new metric, and use this new metric to drive all further development.

C

14 Adapting to a New Crow Species

Your system is now deployed in Hogwarts. Over the last few months, a new species of crow 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. You have only 1,000 images of the new species of crow, and the city expects a better system from you within the next 3 months.

Which of the following should you do first?

A) 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.

B) Put the 1,000 images into the training set so as to try to do better on these crows.

C) Try data augmentation/data synthesis to get more images of the new type of crow.

D) Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split.

A

15 Iteration Speed on Large Datasets

The City Council thinks that having more cats in the city would help scare off crows. They are so happy with your work on the crow detector that they also hire you to build a cat detector. Because of years of working on cat detectors, you have a huge dataset of 100,000,000 cat images, and training on this data takes about two weeks.

Which of the following statements do you agree with? (*Check all that apply.*)

A) Needing two weeks to train will limit the speed at which you can iterate.

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

C) Having built a good crow detector, you should be able to take the same model and hyperparameters and just apply it to the cat dataset, so there is no need to iterate.

D) 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 an $\approx 10\times$ improvement in how quickly you can run experiments, even if each model performs a bit worse because it is trained on less data.

A,B,D