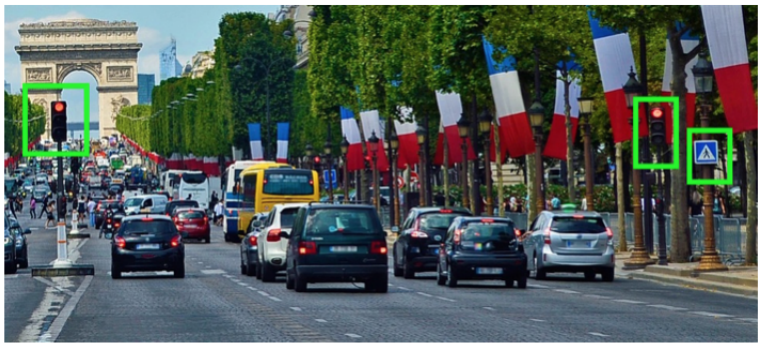
Q.

To help you practice strategies for machine learning, in this week we'll present another scenario and ask how you would act. We think this "simulator" of working in a machine learning project will give a task of what leading a machine learning project could be like!



$$y^{(i)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{matrix} \text{“stop sign”} \\ \text{“pedestrian crossing sign”} \\ \text{“construction ahead sign”} \\ \text{“red traffic light”} \\ \text{“green traffic light”} \end{matrix}$$

You are employed by a startup building self-driving cars. You are in charge of detecting road signs (stop sign, pedestrian crossing sign, construction ahead sign) and traffic signals (red and green lights) in images. The goal is to recognize which of these objects appear in each image. As an example, the above image contains a pedestrian crossing sign and red traffic lights
Your 100,000 labeled images are taken using the front-facing camera of your car. This is also the distribution of data you care most about doing well on. You think you might be able to get a much larger dataset off the internet, that could be helpful for training even if the distribution of internet data is not the same. You are just getting started on this project. What is the first thing you do? Assume each of the steps below would take about an equal amount of time (a few days).

- Spend a few days checking what is human-level performance for these tasks so that you can get an accurate estimate of Bayes error.
- Spend a few days getting the internet data, so that you understand better what data is available.
- Spend a few days collecting more data using the front-facing camera of your car, to better understand how much data per unit time you can collect.
- Spend a few days training a basic model and see what mistakes it makes

View Answer

1. **Explanation :**

```
As discussed in sections, applied ML is a highly iterative process. If you train a basic model and carry out error analysis (see what mis
takes it makes) it will help point you in more promising directions.
```

Q.

Your goal is to detect road signs (stop sign, pedestrian crossing sign, construction ahead sign) and traffic signals (red and green lights) in images. The goal is to recognize which of these objects appear in each image. You plan to use a deep neural network with ReLU units in the hidden layers. For the output layer, a softmax activation would be a good choice for the output layer because this is a multi-task learning problem. True/False?

- True
- False

View Answer

1. **Explanation :**

```
Softmax would be a good choice if one and only one of the possibilities (stop sign, speed bump, pedestrian crossing, green light and red
light) was present in each image.
```

Q.

You are carrying out error analysis and counting up what errors the algorithm makes. Which of these datasets do you think you should manually go through and carefully examine, one image at a time?

- 500 randomly chosen images
- 10,000 randomly chosen images
- 10,000 images on which the algorithm made a mistake
- 500 images on which the algorithm made a mistake

View Answer

1. **Explanation :**

```
Focus on images that the algorithm got wrong. Also, 500 is enough to give you a good initial sense of the error statistics. There's proba
bly no need to look at 10,000, which will take a long time.
```

Q.

After working on the data for several weeks, your team ends up with the following data: 100,000 labeled images taken using the front-facing camera of your car. 900,000 labeled images of roads downloaded from the internet. Each image's labels precisely indicate the presence of any specific road signs and traffic signals or combinations of them. For example, $y^{(i)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ means the image contains a stop sign and a red traffic light. Because this is a multi-task learning problem, you need to have all your $y^{(i)}$ vectors fully labeled. If one example is equal to $\begin{bmatrix} 0 \\ ? \\ 1 \\ 1 \\ ? \end{bmatrix}$ then the learning algorithm will not be able to use that example. True/False?

- True
- False

View Answer

1. **Explanation :**

```
As seen in the sections on multi-task learning, you can compute the cost such that it is not influenced by the fact that some entries hav
en't been labeled.
```

Q.

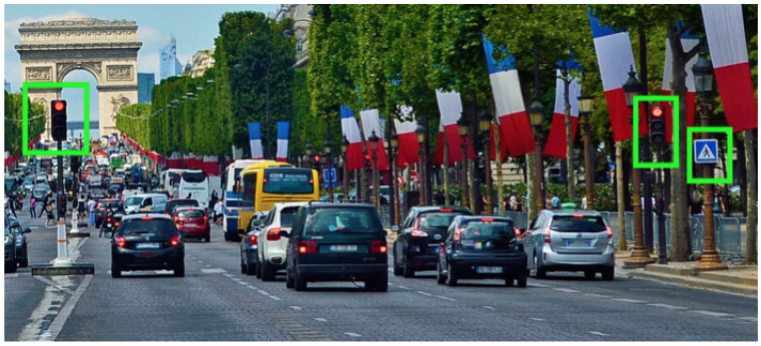
The distribution of data you care about contains images from your car's front-facing camera; which comes from a different distribution than the images you were able to find and download off the internet. How should you split the dataset into train/dev/test sets?

- Mix all the 100,000 images with the 900,000 images you found online. Shuffle everything. Split the 1,000,000 images dataset into 600,000 for the training set, 200,000 for the dev set and 200,000 for the test set.
- Mix all the 100,000 images with the 900,000 images you found online. Shuffle everything. Split the 1,000,000 images dataset into 980,000 for the training set, 10,000 for the dev set and 10,000 for the test set.
- Choose the training set to be the 900,000 images from the internet along with 20,000 images from your car's front-facing camera. The 80,000 remaining images will be split equally in dev and test sets.
- Choose the training set to be the 900,000 images from the internet along with 80,000 images from your car's front-facing camera. The 20,000 remaining images will be split equally in dev and test sets.

View Answer

Q.

Assume you've finally chosen the following split between of the data:



$$y^{(i)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{matrix} \text{“stop sign”} \\ \text{“pedestrian crossing sign”} \\ \text{“construction ahead sign”} \\ \text{“red traffic light”} \\ \text{“green traffic light”} \end{matrix}$$

You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Which of the following are True?

- You have a large variance problem because your training error is quite higher than the human-level error.
- You have a large data-mismatch problem because your model does a lot better on the training-dev set than on the dev set
- You have a large variance problem because your model is not generalizing well to data from the same training distribution but that it has never seen before.

Q.

Assume you've finally chosen the following split between of the data:

Dataset:	Contains:	Error of the algorithm:
Training	940,000 images randomly picked from (900,000 internet images + 60,000 car's front-facing camera images)	8.8%
Training-Dev	20,000 images randomly picked from (900,000 internet images + 60,000 car's front-facing camera images)	9.1%
Dev	20,000 images from your car's front-facing camera	14.3%
Test	20,000 images from the car's front-facing camera	14.8%

You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Which of the following are True?

- Your algorithm overfits the dev set because the error of the dev and test sets are very close.
- You have a large avoidable-bias problem because your training error is quite a bit higher than the human-level error. ✔

Q.

Based on table from the previous question, a friend thinks that the training data distribution is much easier than the dev/test distribution. What do you think?

- Your friend is right. (I.e., Bayes error for the training data distribution is probably lower than for the dev/test distribution.)
- Your friend is wrong. (I.e., Bayes error for the training data distribution is probably higher than for the dev/test distribution.)
- There's insufficient information to tell if your friend is right or wrong.

View Answer

Q.

You decide to focus on the dev set and check by hand what are the errors due to. Here is a table summarizing your discoveries:

Overall dev set error	14.3%
Errors due to incorrectly labeled data	4.1%
Errors due to foggy pictures	8.0%
Errors due to rain drops stuck on your car's front-facing camera	2.2%
Errors due to other causes	1.0%

In this table, 4.1%, 8.0%, etc.are a fraction of the total dev set (not just examples your algorithm mislabeled). I.e. about 8.0/14.3 = 56% of your errors are due to foggy pictures. The results from this analysis implies that the team's highest priority should be to bring more foggy pictures into the training set so as to address the 8.0% of errors in that category. True/False?

- True because it is the largest category of errors. As discussed in lecture, we should prioritize the largest category of error to avoid wasting the team's time.
- False because this would depend on how easy it is to add this data and how much you think your team thinks it'll help.
- True because it is greater than the other error categories added together (8.0 > 4.1+2.2+1.0).
- False because data augmentation (synthesizing foggy images by clean/non-foggy images) is more efficient.

View Answer

Q.

You can buy a specially designed windshield wiper that help wipe off some of the raindrops on the front-facing camera. Based on the table from the previous question, which of the following statements do you agree with?

- 2.2% would be a reasonable estimate of the maximum amount this windshield wiper could improve performance.
- 2.2% would be a reasonable estimate of the minimum amount this windshield wiper could improve performance.
- 2.2% would be a reasonable estimate of how much this windshield wiper will improve performance.
- 2.2% would be a reasonable estimate of how much this windshield wiper could worsen performance in the worst case.