# Quiz for Neural Network Optimization

Friday, May 3, 2019 2:33 AM

If you have 10,000,000 examples, how would you split the train/dev/test set?

- [x] 98% train . 1% dev . 1% test
- [ ] 60% train . 20% dev . 20% test
- [ ] 33% train . 33% dev . 33% test

## Question 2

1
point

**2. Question 2**

The dev and test set should:

- [x] Come from the same distribution
- [ ] Come from different distributions
- [ ] Be identical to each other (same (x,y) pairs)
- [ ] **Have the same number of examples**

## Question 3

1
point

**3. Question 3**

If your Neural Network model seems to have high bias, what of the following would be promising things to try? (Check all that apply.)

- [ ] Get more training data
- [x] Increase the number of units in each hidden layer
- [ ] Add regularization
- [ ] Get more test data
- [x] Make the Neural Network deeper

## Question 4

1
point

**4. Question 4**

You are working on an automated check-out kiosk for a supermarket, and are building a classifier for apples, bananas and oranges. Suppose your classifier obtains a training set error of 0.5%, and a dev set error of 7%. Which of the following are promising things to try to improve your classifier? (Check all that apply.)

- [x] Increase the regularization parameter lambda
- [ ] Decrease the regularization parameter lambda
- [x] Get more training data
- [ ] Use a bigger neural network

## Question 5

1
point

**5. Question 5**

What is weight decay?

- [ ] The process of gradually decreasing the learning rate during training.
- [x] A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration.
- [ ] A technique to avoid vanishing gradient by imposing a ceiling on the values of the weights.
- [ ] Gradual corruption of the weights in the neural network if it is trained on noisy data.

## Question 6

1
point

**6. Question 6**

What happens when you increase the regularization hyperparameter lambda?

- [x] Weights are pushed toward becoming smaller (closer to 0)
- [ ] Weights are pushed toward becoming bigger (further from 0)
- [ ] Doubling lambda should roughly result in doubling the weights
- [ ] Gradient descent taking bigger steps with each iteration (proportional to lambda)

## Question 7

1
point

**7. Question 7**

With the inverted dropout technique, at test time:

- [ ] You apply dropout (randomly eliminating units) and do not keep the 1/keep_prob factor in the calculations used in training
- [ ] You apply dropout (randomly eliminating units) but keep the 1/keep_prob factor in the calculations used in training. ✗
- [ ] You do not apply dropout (do not randomly eliminate units), but keep the 1/keep_prob factor in the calculations used in training. ✗
- [x] You do not apply dropout (do not randomly eliminate units) and do not keep the 1/keep_prob factor in the calculations used in training

## Question 8

1
point

**8. Question 8**

Increasing the parameter keep_prob from (say) 0.5 to 0.6 will likely cause the following: (Check the two that apply)

- [ ] Increasing the regularization effect
- [x] Reducing the regularization effect
- [ ] Causing the neural network to end up with a higher training set error
- [x] Causing the neural network to end up with a lower training set error

## Question 9

1
point

**9. Question 9**

Which of these techniques are useful for reducing variance (reducing overfitting)? (Check all that apply.)

- [x] Dropout
- [ ] Xavier initialization
- [ ] Exploding gradient
- [x] L2 regularization
- [ ] Vanishing gradient
- [ ] Gradient Checking
- [x] Data augmentation

## Question 10

1
point

**10. Question 10**

Why do we normalize the inputs $x$?

- [ ] Normalization is another word for regularization--It helps to reduce variance ✗
- [ ] It makes it easier to visualize the data
- [ ] It makes the parameter initialization faster
- [x] It makes the cost function faster to optimize

**1. Question 1**

Which notation would you use to denote the 3rd layer's activations when the input is the 7th example from the 8th minibatch?

- [x] a^{[3]\{8\}(7)}*a*[3]{8}(7)
- [ ] a^{[3]\{7\}(8)}*a*[3]{7}(8)
- [ ] a^{[8]\{7\}(3)}*a*[8]{7}(3)
- [ ] a^{[8]\{3\}(7)}*a*[8]{3}(7)

Question 2

1

point

**2. Question 2**

Which of these statements about mini-batch gradient descent do you agree with?

- [ ] Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.
- [ ] You should implement mini-batch gradient descent without an explicit for-loop over different mini-batches, so that the algorithm processes all mini-batches at the same time (vectorization).
- [x] One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.

Question 3

1

point

**3. Question 3**

Why is the best mini-batch size usually not 1 and not m, but instead something in-between?

- [ ] If the mini-batch size is 1, you end up having to process the entire training set before making any progress.
- [ ] If the mini-batch size is m, you end up with stochastic gradient descent, which is usually slower than mini-batch gradient descent.
- [x] If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.
- [x] If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.

Question 4

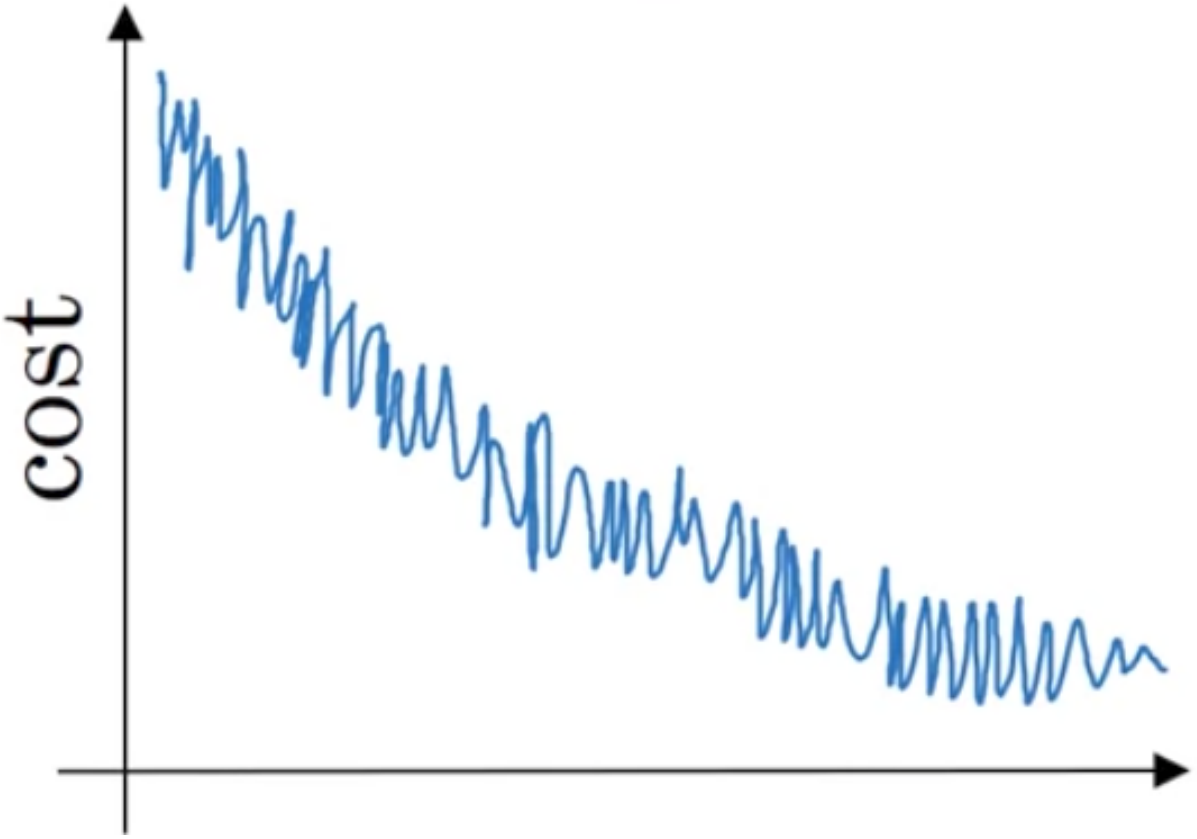
Question 4

1

point

**4. Question 4**

Suppose your learning algorithm's cost $J$, plotted as a function of the number of iterations, looks like this:




Which of the following do you agree with?

- [ ] Whether you're using batch gradient descent or mini-batch gradient descent, something is wrong.
- [x] If you're using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.
- [ ] If you're using mini-batch gradient descent, something is wrong. But if you're using batch gradient descent, this looks acceptable.
- [ ] Whether you're using batch gradient descent or mini-batch gradient descent, this looks acceptable.

Question 5

1

point

**5. Question 5**

Suppose the temperature in Casablanca over the first three days of January are the same:

Jan 1st: \theta_1 = 10^o C$\vartheta 1 = 10^{o}C$

Jan 2nd: \theta_2 10^o C$\vartheta 2 10^{o}C$

(We used Fahrenheit in lecture, so will use Celsius here in honor of the metric world.)

Say you use an exponentially weighted average with \beta = 0.5$\beta = 0.5$ to track the temperature: v_0 = 0$v_0 = 0$, v_t = \beta v_{t-1} +(1-\beta)\theta_t$v_t = \beta v_{t-1} + (1-\beta)\vartheta_t$. If v_2$v_2$ is the value computed after day 2 without bias correction, and v_2 ^{corrected}$v_2^{corrected}$ is the value you compute with bias correction. What are these values? (You might be able to do this without a calculator, but you don't actually need one. Remember what is bias correction doing.)

- [x] v_2 = 7.5$v_2 = 7.5$, v_2^{corrected} = 10$v_2^{corrected} = 10$
- [ ] v_2 = 7.5$v_2 = 7.5$, v_2^{corrected} = 7.5$v_2^{corrected} = 7.5$
- [ ] v_2 = 10$v_2 = 10$, v_2^{corrected} = 10$v_2^{corrected} = 10$
- [ ] v_2 = 10$v_2 = 10$, v_2^{corrected} = 7.5$v_2^{corrected} = 7.5$

Question 6

1

point

**6. Question 6**

Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number.

- [ ] \alpha = \frac{1}{1+2*t} \alpha_0$\alpha = \frac{1}{1+2*t}\alpha_0$
- [ ] \alpha = 0.95^t \alpha_0$\alpha = 0.95^t\alpha_0$
- [x] \alpha = e^t \alpha_0$\alpha = e^t\alpha_0$
- [ ] \alpha = \frac{1}{\sqrt{t}} \alpha_0$\alpha = \frac{1}{\sqrt{t}}\alpha_0$

Question 7

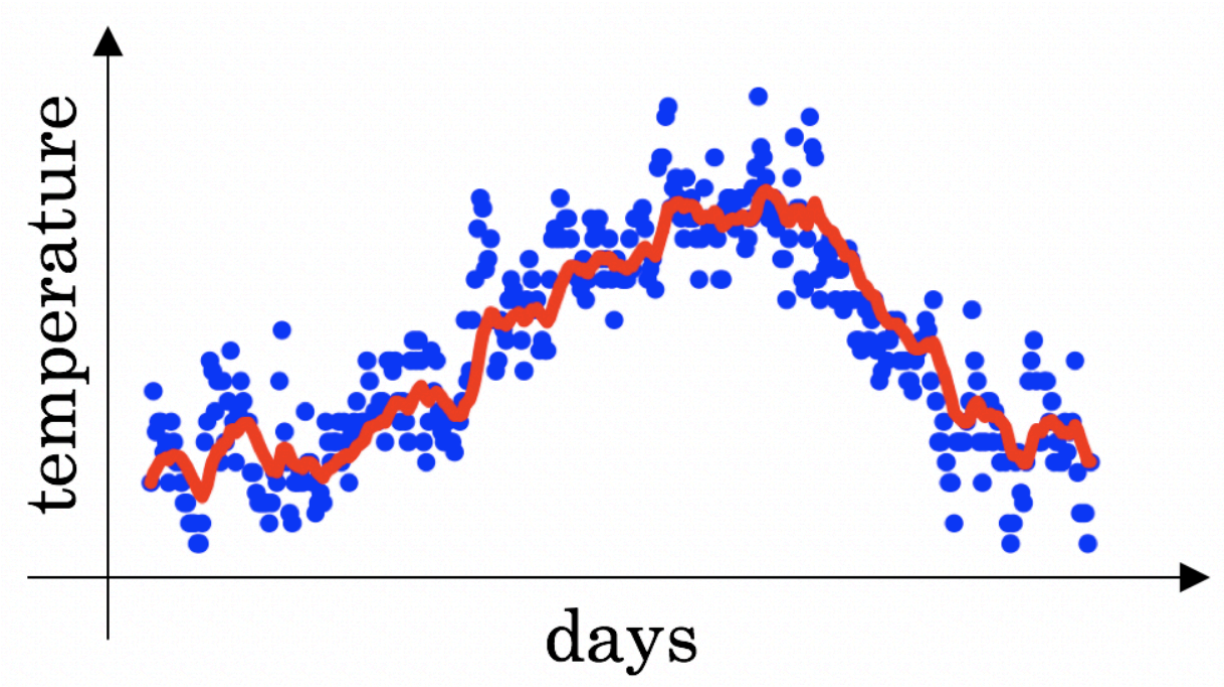
1

point

**7. Question 7**

You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: v_{t} = \beta v_{t-1} + (1-\beta)\theta_t$v_t = \beta v_{t-1} + (1-\beta)\vartheta_t$. The red line below was computed using \beta

= 0.96=0.9. What would happen to your red curve as you vary \beta? (Check the two that apply)




- [ ] Decreasing \beta$\beta$ will shift the red line slightly to the right.
- [x] Increasing \beta$\beta$ will shift the red line slightly to the right.
- [x] Decreasing \beta$\beta$ will create more oscillation within the red line.
- [ ] Increasing \beta$\beta$ will create more oscillations within the red line.

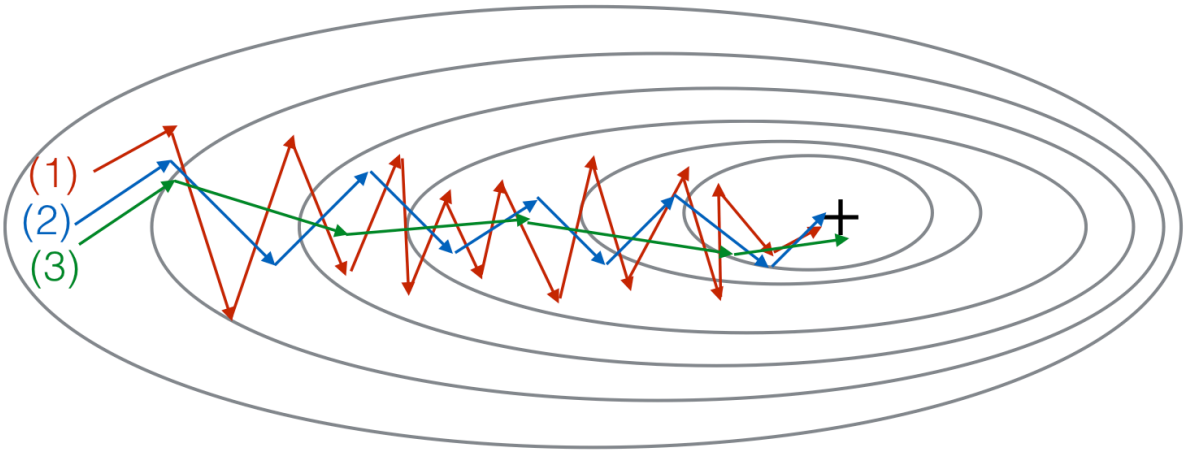
Question 8

1

point

**8. Question 8**

Consider this figure:




These plots were generated with gradient descent; with gradient descent with

momentum (\beta$\beta$ = 0.5) and gradient descent with momentum (\beta$\beta$ = 0.9). Which curve corresponds to which algorithm?

- [ ] (1) is gradient descent with momentum (small \beta$\beta$). (2) is gradient descent. (3) is gradient descent with momentum (large \beta$\beta$)
- [x] (1) is gradient descent. (2) is gradient descent with momentum (small \beta$\beta$). (3) is gradient descent with momentum (large \beta$\beta$)
- [ ] (1) is gradient descent with momentum (small \beta$\beta$), (2) is gradient descent with momentum (small \beta$\beta$), (3) is gradient descent
- [ ] (1) is gradient descent. (2) is gradient descent with momentum (large \beta$\beta$) . (3) is gradient descent with momentum (small \beta$\beta$)

Question 9

1

point

**9. Question 9**

Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function $\mathcal{J}(W_{[1]},b_{[1]},...,W_{[L]},b_{[L]})$. Which of the following techniques could help find parameter values that attain a small value for $\mathcal{J}$? (Check all that apply)

- [x] Try using Adam
- [x] Try tuning the learning rate \alpha$\alpha$
- [x] Try better random initialization for the weights
- [x] Try mini-batch gradient descent
- [ ] Try initializing all the weights to zero

Question 10

1

point

**10. Question 10**

Which of the following statements about Adam is False?

- [ ] The learning rate hyperparameter \alpha$\alpha$ in Adam usually needs to be tuned.
- [ ] We usually use “default” values for the hyperparameters \beta_1, \beta_2$\beta$1,$\beta$2 and $\varepsilon$ in Adam (\beta_1 = 0.9$\beta$1=0.9, \beta_2 = 0.999$\beta$2=0.999, $\varepsilon=10^{-8}$)
- [ ] Adam combines the advantages of RMSProp and momentum
- [x] Adam should be used with batch gradient computations, not with mini-batches.

**1. Question 1**

If searching among a large number of hyperparameters, you should try values in a grid rather than random values, so that you can carry out the search more systematically and not rely on chance. True or False?

- [ ] True
- [x] False

Question 2

1

point

**2. Question 2**

Every hyperparameter, if set poorly, can have a huge negative impact on training, and so all hyperparameters are about equally important to tune well. True or False?

- [ ] True
- [x] False

Question 3

1

point

**3. Question 3**

During hyperparameter search, whether you try to babysit one model ("Panda" strategy) or train a lot of models in parallel ("Caviar") is largely determined by:

- [ ] Whether you use batch or mini-batch optimization
- [ ] The presence of local minima (and saddle points) in your neural network
- [x] The amount of computational power you can access
- [ ] The number of hyperparameters you have to tune

Question 4

1

point

**4. Question 4**

If you think \beta$\beta$ (hyperparameter for momentum) is between on 0.9 and 0.99, which of the following is the recommended way to sample a value for beta?

- [ ]
```
r = np.random.rand()
beta = r*0.09 + 0.9
```

- [x]
```
r = np.random.rand()
beta = 1-10**(- r - 1)
```

```
beta = 1-10**(- r - 1)
```

```
1
2
```

```
r = np.random.rand()
beta = 1-10**(- r + 1)
```

```
1
2
```

```
r = np.random.rand()
beta = r*0.9 + 0.09
```

Question 5

1

point

**5. Question 5**

Finding good hyperparameter values is very time-consuming. So typically you should do it once at the start of the project, and try to find very good hyperparameters so that you don't ever have to revisit tuning them again. True or false?

- [ ] True
- [x] False

Question 6

1

point

**6. Question 6**

In batch normalization as presented in the videos, if you apply it on the l/th layer of your neural network, what are you normalizing?

- [ ] b^{[l]}$b[l]$
- [x] z^{[l]}$z[l]$
- [ ] W^{[l]}$W[l]$
- [ ] a^{[l]}$a[l]$

Question 7

1

point

**7. Question 7**

In the normalization formula $z(i)norm=z(i)-\mu\sigma 2+\varepsilon\sqrt{}$, why do we use epsilon?

- [x] To avoid division by zero
- [ ] To speed up convergence
- [ ] In case \mu$\mu$ is too small
- [ ] To have a more accurate normalization

Question 8

1

point

**8. Question 8**

Which of the following statements about \gamma$\gamma$ and \beta$\beta$ in Batch Norm are true?

- [x] They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent.
- [ ] There is one global value of $\gamma \in \mathbb{R}$ and one global value of $\beta \in \mathbb{R}$ for each layer, and applies to all the hidden units in that layer.
- [ ] \beta$\beta$ and \gamma$\gamma$ are hyperparameters of the algorithm, which we tune via random sampling.
- [x] They set the mean and variance of the linear variable z^[l]$z^{[l]}$ of a given layer.
- [ ] The optimal values are $\gamma = \sqrt{\sigma^2 + \varepsilon}$, and \beta = \mu$\beta = \mu$.

Question 9

1

point

**9. Question 9**

After training a neural network with Batch Norm, at test time, to evaluate the neural network on a new example you should:

- [ ] Skip the step where you normalize using \mu$\mu$ and \sigma^2$\sigma^2$ since a single test example cannot be normalized.
- [ ] Use the most recent mini-batch's value of \mu$\mu$ and \sigma^2$\sigma^2$ to perform the needed normalizations.
- [ ] If you implemented Batch Norm on mini-batches of (say) 256 examples, then to evaluate on one test example, duplicate that example 256 times so that you're working with a mini-batch the same size as during training.
- [x] Perform the needed normalizations, use \mu$\mu$ and \sigma^2$\sigma^2$ estimated using an exponentially weighted average across mini-batches seen during training.

Question 10

1

point

**10. Question 10**

Which of these statements about deep learning programming frameworks are true? (Check all that apply)

- [x] Even if a project is currently open source, good governance of the project helps ensure that the it remains open even in the long term, rather than become closed or modified to benefit only one company.
- [ ] Deep learning programming frameworks require cloud-based machines to run.

☑ A programming framework allows you to code up deep learning algorithms with typically fewer lines of code than a lower-level language such as Python.

**1. Question 1**

# Problem Statement

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 labelled:

- 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

False

Question 2

1

point

**2. Question 2**

After further discussions, the city narrows down 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 10sec to classify a new image."
- "We want the model to fit in 10MB of memory."

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

| Test Accuracy | Runtime | Memory size |
|---|---|---|
| 97% | 1 sec | 3MB |

| Test Accuracy | Runtime | Memory size |
|---|---|---|
| 99% | 13 sec | 9MB |

| Test Accuracy | Runtime | Memory size |
|---|---|---|
| 97% | 3 sec | 2MB |

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

Question 3

1

point

**3. Question 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.

Question 4

1

point

**4. Question 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 | Dev | Test |
|---|---|---|
| 3,333,334 | 3,333,333 | 3,333,333 |

| Train | Dev | Test |
|---|---|---|

| | | |
|---|---|---|
| 6,000,000 | 1,000,000 | 3,000,000 |

| Train | Dev | Test |
|---|---|---|
| 9,500,000 | 250,000 | 250,000 |

| Train | Dev | Test |
|---|---|---|
| 6,000,000 | 3,000,000 | 1,000,000 |

Question 5

1

point

**5. Question 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

Question 6

1

point

**6. Question 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.

The 1,000,000 citizens’ data images do not have a consistent x-->y mapping as the rest of the data (similar to the New York City/Detroit housing prices example from lecture).❌

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

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

Question 7

1

point

**7. Question 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?

- [ ] Yes, because having 4.0% training error shows you have high bias.
- [ ] Yes, because this shows your bias is higher than your variance.
- [ ] No, because this shows your variance is higher than your bias.
- [x] No, because there is insufficient information to tell.

Question 8

1

point

**8. Question 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”?

- [ ] 0.0% (because it is impossible to do better than this)
- [x] 0.3% (accuracy of expert #1)
- [ ] 0.4% (average of 0.3 and 0.5) ✗
- [ ] 0.75% (average of all four numbers above) ✗

Question 9

1

point

**9. Question 9**

Which of the following statements do you agree with?

Which of the following statements do you agree with?

- [x] A learning algorithm's performance can be better than human-level performance but it can never be better than Bayes error.
- [ ] A learning algorithm's performance can never be better than human-level performance but it can be better than Bayes error.
- [ ] A learning algorithm's performance can never be better than human-level performance nor better than Bayes error.
- [ ] A learning algorithm's performance can be better than human-level performance and better than Bayes error.

Question 10

1

point

**10. Question 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.)

- [ ] Try increasing regularization.
- [x] Train a bigger model to try to do better on the training set.
- [ ] Get a bigger training set to reduce variance.
- [x] Try decreasing regularization.

Question 11

1

point

**11. Question 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.)

- [ ] You should get a bigger test set.
- [x] You have overfit to the dev set.
- [ ] You have underfit to the dev set.
- [x] You should try to get a bigger dev set.

Question 12

1

point

**12. Question 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.)

- [x] If the test set is big enough for the 0.05% error estimate to be accurate, this implies Bayes error is $\leq 0.05$
- [ ] This is a statistical anomaly (or must be the result of statistical noise) since it should not be possible to surpass human-level performance.
- [ ] With only 0.09% further progress to make, you should quickly be able to close the remaining gap to 0%
- [x] It is now harder to measure avoidable bias, thus progress will be slower going forward.

Question 13

1

point

**13. Question 13**

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?

- [ ] Look at all the models you've developed during the development process and find the one with the lowest false negative error rate. ✗
- [ ] Ask your team to take into account both accuracy and false negative rate during development ✗

development.

- [x] Rethink the appropriate metric for this task, and ask your team to tune to the new metric.
- [ ] Pick false negative rate as the new metric, and use this new metric to drive all further development.

Question 14

1

point

**14. Question 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?

- [x] 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.
- [ ] Put the 1,000 images into the training set so as to try to do better on these birds.
- [ ] Try data augmentation/data synthesis to get more images of the new type of bird.

- [ ] Add the 1,000 images into your dataset and reshuffle into a new train/dev/test split.❌

Question 15

1

point

**15. Question 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.)

- [ ] Having built a good Bird 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.
- [x] 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 \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.
- [x] Needing two weeks to train will limit the speed at which you can iterate.
- [x] Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.

1

point

**1. Question 1**

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!

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

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

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 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 getting the internet data, so that you understand better what data is available.
- [x] Spend a few days training a basic model and see what mistakes it makes.
- [ ] Spend a few days checking what is human-level performance for these tasks so that you can get an accurate estimate of Bayes error.

Question 2

1

point

**2. Question 2**

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
- [x] False

Question 3

1

point

**3. Question 3**

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

Question 4

1

point

**4. Question 4**

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)}

$y(i) = \lceil \lfloor | | | | 10010 \rceil \rfloor | | | |$ 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)}

$y(i)$ vectors fully labeled. If one example is equal to $\lceil \lfloor | | | | 0?11? \rceil \rfloor | | | |$ then the learning algorithm will not be able to use that example. True/False?

- [ ] True
- [x] False

Question 5

1

point

**5. Question 5**

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

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.

- [x] 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.
- [ ] 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.
- [ ] 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.

Question 6

1

point

**6. Question 6**

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? (Check all that apply).

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

- [ ] You have a large variance problem because your training error is quite higher than the human-level error.

Question 7

1

point

**7. Question 7**

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.)
- [x] There's insufficient information to tell if your friend is right or wrong.

Question 8

1

point

**8. Question 8**

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.✗
- [ ] True because it is greater than the other error categories added together (8.0 > 4.1+ 2.2+1.0).
- [x] 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.

- [ ] False because data augmentation (synthesizing foggy images by clean/non-foggy images) is more efficient.

Question 9

1

point

**9. Question 9**

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?

- [x] 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.

Question 10

1

point

**10. Question 10**

You decide to use data augmentation to address foggy images. You find 1,000 pictures of fog off the internet, and "add" them to clean images to synthesize foggy days, like this:


Which of the following statements do you agree with?

- [ ] There is little risk of overfitting to the 1,000 pictures of fog so long as you are combing it with a much larger (>>1,000) of clean/non-foggy images.❌
- [ ] Adding synthesized images that look like real foggy pictures taken from the front-facing camera of your car to training dataset won't help the model improve because it will introduce avoidable-bias.

- [x] So long as the synthesized fog looks realistic to the human eye, you can be confident that the synthesized data is accurately capturing the distribution of real foggy images (or a subset of it), since human vision is very accurate for the problem you're solving.

Question 11

1

point

**11. Question 11**

After working further on the problem, you've decided to correct the incorrectly labeled data on the dev set. Which of these statements do you agree with? (Check all that apply).

- [x] You should also correct the incorrectly labeled data in the test set, so that the dev and test sets continue to come from the same distribution
- [ ] You should correct incorrectly labeled data in the training set as well so as to avoid your training set now being even more different from your dev set.
- [ ] You should not correct the incorrectly labeled data in the test set, so that the dev and test sets continue to come from the same distribution
- [x] You should not correct incorrectly labeled data in the training set as it does not worth the time.

Question 12

1

point

**12. Question 12**

So far your algorithm only recognizes red and green traffic lights. One of your colleagues in the startup is starting to work on recognizing a yellow traffic light. (Some countries call it an orange light rather than a yellow light; we'll use the US convention of calling it yellow.) Images containing yellow lights are quite rare, and she doesn't have enough data to build a good model. She hopes you can help her out using transfer learning.

What do you tell your colleague?

- [x] She should try using weights pre-trained on your dataset, and fine-tuning further with the yellow-light dataset.
- [ ] If she has (say) 10,000 images of yellow lights, randomly sample 10,000 images from your dataset and put your and her data together. This prevents your dataset from "swamping" the yellow lights dataset.
- [ ] You cannot help her because the distribution of data you have is different from hers, and is also lacking the yellow label.
- [ ] Recommend that she try multi-task learning instead of transfer learning using all the

data.

Question 13

1

point

**13. Question 13**

Another colleague wants to use microphones placed outside the car to better hear if there're other vehicles around you. For example, if there is a police vehicle behind you, you would be able to hear their siren. However, they don't have much to train this audio system. How can you help?

- [ ] Transfer learning from your vision dataset could help your colleague get going faster. Multi-task learning seems significantly less promising.
- [ ] Multi-task learning from your vision dataset could help your colleague get going faster. Transfer learning seems significantly less promising.
- [ ] Either transfer learning or multi-task learning could help our colleague get going faster.
- [x] Neither transfer learning nor multi-task learning seems promising.

Question 14

1

point

**14. Question 14**

To recognize red and green lights, you have been using this approach:

- **(A)** Input an image (x) to a neural network and have it directly learn a mapping to make a prediction as to whether there's a red light and/or green light (y).

A teammate proposes a different, two-step approach:

- **(B)** In this two-step approach, you would first (i) detect the traffic light in the image (if any), then (ii) determine the color of the illuminated lamp in the traffic light.

Between these two, Approach B is more of an end-to-end approach because it has distinct steps for the input end and the output end. True/False?

- [ ] True
- [x] False

Question 15

1

point

**15. Question 15**

Approach A (in the question above) tends to be more promising than approach B if

you have a _____ (fill in the blank).

- [x] Large training set
- [ ] Multi-task learning problem.
- [ ] Large bias problem.
- [ ] Problem with a high Bayes error.