This document is view-only accessible via http://www.comp.nus.edu.sg/~cs3244/AY2122S1/faq.html. To ask a question, please ask your #tg leader, or write directly in the component channel (#project, #assignment, #assessment, #lectures, #projects, #tutorials)

CS3244 AY21/22 Sem 1

Archived (not so frequent) Question and Answers

#projects
#lectures
- Week 01 – ML Pipeline
- Week 02 – Paradigms of ML and kNN
- Week 03 – Decision Trees
- Week 04 – Linear Models
- Week 05 – Bias and Variance
- Week 06 – Regularization and Validation

#tutorials
#assignments
#assessment

This document archives the pertinent questions from the class Slack. This document is fair and admissible material to search and use for assessments.

Last updated: 26 Sep 2021 – #assessments Qs 1–9
23 Sep 2021 – Week 06, Q #2.

#projects

1. **Q**: Would I be able to form a project group with a friend that is not in the same tutorial group?
   A: Yes. The tutorial groups are not restrictions.

2. **Q**: Do all members of a team have to fill out the luminus sub group declaration survey? Or is it sufficient for 1 member to submit this?
   A: One member to fill out the form. If you need to modify it, please have the same member edit their submission. 😎

3. **Q**: I was wondering if we must join sub group of 3? (I don't really have friends taking this module..)
   A: You can join as a subgroup of between 1 (just yourself) to 3 students. You can join this morning's meet-up activity during our lecture slot at 11:00-12:00 to bump into other classmates looking for project partners.

4. **Q**: May I know how detailed are we expected to answer these 2 questions? Since we haven't really started working on the project, I'm afraid we cannot make a good estimate of the schedule and the resource we will be using.
   A: It's OK if you don't have a very good idea of your answers to this question. These questions are there also to guide you along the necessary criteria to think about for your projects. If you have a sentence or two, it should be sufficient.

   For the schedule, most group in the past specify a rough weekly schedule where each

member or subgroup is assigned duties to do. Don't worry, your peers and your instructing staff won't be holding you to it, but it is useful to try to take the project and break it down to smaller milestones, so that your group isn't surprised at the amount of work at the last minute and so that your to-be-assigned project mentor can help check your progress against your estimate (and perhaps help you refine your schedule).

Of course, if you have more detail on either question, that's fine as well. Some teams may not have additional resources that they want to add, and that's fine, but please justify why.

5. **Q**: To add on to this question, may I know what does it mean by "(including peer review)"? Is it only 1 or 2 member from our team will be doing the peer review? Also, may I know whether what other deliverables we have aside from slides and proposal (e.g. do we need to submit our codes, do we need to make a video etc?)
   A: We hope that all group members will be involved the the peer review process, but will not be enforcing it. It is good practice (where your group dynamics allow) to balance having all members take part in each part of the project and in specializing in just one component.

   Aside from the project proposal, you will be asked to make your final project video presentation (and the respective slides) available. There will be no project report deliverable. If you have code that you wish to submit for our review, you may.

6. **Q:** Can we ask what do you mean for Qn8: "A high-level description of the general approach you'll use to address the questions. Survey on the current progress on the problem/task. Sketch out what evidence you are planning to gather (e.g., how you can answer the questions through experiments on the data)."? Does Survey on the current progress mean a literature review on what has been done in the field already? Also by "what evidence you are planning to gather" do you mean how we are planning to measure success/failure of our models?
   A: Thanks for your question. Yes, the survey on the current progress is a short summary on what has been tackled on your problem / dataset already, and not what your group has done up to this point. We don't expect you to have started much besides getting background on your project 😅
   By what evidence you are planning to gather, we mean what data or experimental results you plan to do to answer the questions stated in the previous Question 7. Evaluation is part of the following Question 9, but in Question 8 we are asking for the sources of evidence to use for the evaluation.

7. **Q**: Are there any limitations to the library that we can use?
   A: There shouldn't be. If in doubt, please ask us directly or through your project mentor (yet to be assigned; hopefully will be assigned this week) or your TG leader.
   General ML and data science toolkits are fine.

8. **Q**: Can I ask whether we will receive responses on our submitted proposal? If yes, when will we receive it?
   A: You will receive three sets of feedback:
   - From your peer teams (from the assignments and form that were disseminated to you early morning today)
   - From your project mentor
   - From your an assigned instructor (either @[Instructor] Brian Lim or @[Instructor] Kan Min Yen)

   We hope to get these to you via LumiNUS Gradebook but may also try to disseminate these by email since formatting in Gradebook is problematic. You'll get this feedback around late Week 07 or 08.

9. **Q**: Just to be sure, for our project work, can we code in any language (i.e. R) other than python? [*Appears in #lectures*]
   A: Answered in AMA on 9 Sep (Week 05B).
   (TL;DR) Yes.

10. **Q**: Does the project difficulty matters for the grading, e.g. taking the ez-est dataset VS doing CV/RL? Would we be graded based on the novelty or "importance" of the use case/problem our group comes up with? [*Merges two separate questions; appears in #lectures*]
    A: We aren't looking at technical complexity when grading the project. So the choice of CV vs. RL, or Vision vs. NLP is inconsequential for the most part.
    - What we are looking at is the learning that comes out of engaging in the project. We want to see you twist your mind and come up with interesting approaches to the problem of choice
    - We also do not place heavy emphasis on metrics like accuracy, log loss, precision, recall, FID score, etc.. We'd care more about questions like "why did you use XYZ Metric over ABC Metric for this problem?". Getting a +0.5 accuracy boost doesn't matter as much as why you chose to do ABC Technique that brought about that accuracy boost in the first place
    - We care more about how you communicate your findings to us in an interesting way like your analysis, your wins, your losses (pun not intended), etc.. That way, it shows us that you gained valuable experience from this project that you can apply to future projects
    - You are allowed to explore models not covered in class at your own discretion. We only teach you the fundamentals in hopes of making you comfortable with the math/concepts involved. Beyond that, you can look at more complex models and techniques not taught in CS3244 for your projects. But again, complexity is not the focus, communication and understanding the 2W1H (why, what, how) are.
    - There isn't any true "novelty" in these projects per se. They are popular benchmarks found in the real world with increasing difficulty of use. We want

you to have your own unique spin to these solutions (please don't copy-paste/plagiarise someone else's code from online LOL) and present them in a way you and I (ie. the teaching staff) understand.

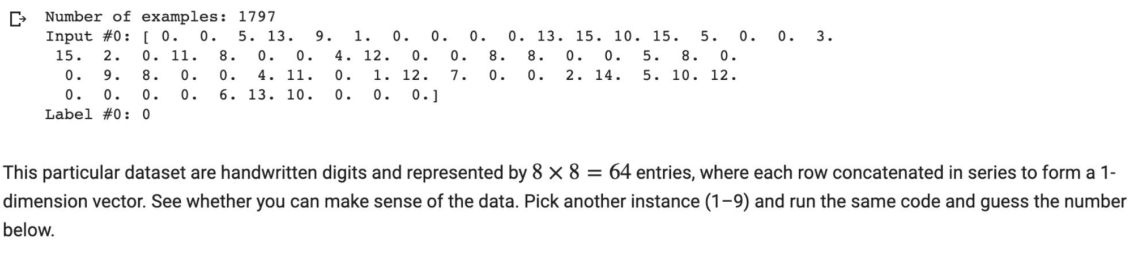
These projects are for you in the long run, not us. Hope this helps.

#lectures

We are archiving the questions in AMA with their respective topic.

Week 01 – ML Pipeline

1. **Q**: Hi everyone, regarding the Colab [example] we went through, how can I tell that the data set is 8 * 8 based on the output ? I am confused, why not a 1 * 64 1d array?



```
Number of examples: 1797
Input #0: [ 0.  0.  5. 13.  9.  1.  0.  0.  0.  0. 13. 15. 10. 15.  5.  0.  0.  3.
 15.  2.  0. 11.  8.  0.  0.  4. 12.  0.  0.  8.  8.  0.  0.  5.  8.  0.
  0.  9.  8.  0.  0.  4. 11.  0.  1. 12.  7.  0.  0.  2. 14.  5. 10. 12.
  0.  0.  0.  0.  6. 13. 10.  0.  0.  0.]
Label #0: 0
```

This particular dataset are handwritten digits and represented by $8 \times 8 = 64$ entries, where each row concatenated in series to form a 1-dimension vector. See whether you can make sense of the data. Pick another instance (1−9) and run the same code and guess the number below.

A: `print(x_digits[j].shape)` to obtain the shape of the array.
In this case the image has been flattened (preprocessed) from 8, 8 to (64).
Sklearn returned a numpy array, which you can find out by printing `type(x_digits[j])`. If you want to eyeball the output to figure out the shape, you have to pay attention to the brackets. A (2,2) array would be [[0, 1], [2, 3]] whereas a flattened (4) array would be [0, 1, 2, 3]

Week 02 – Paradigms of ML and kNN

1. **Q:** For the L1 distance metric, if the images are tinted or shifted, wouldn't the values of the pixels change? If so, why would they still have the same distance to the original? (Slide 66)
A: Sorry this wasn't so clear. They are all copies of the original but the right 3 are changed in a minor way to make the L1 pixel different exactly the same amount. The boxed one has clear differences in the boxes; the shifted shifts the image very slightly downwards but is otherwise identical to the original; and the tinted one changes all the pixels with a bluish tint.

2. **Q**: We mentioned in the lesson about Hypothesis set, from my understanding, we see a simple problem and say oh, we can define
`H = the set of all simple linear regression models available in the form of y = mx + c where m and c are parameters to be learned (?)`

We went through KNN with us, but it is not immediately clear to me (at least), what can be our hypothesis set...? I am a bit lost on what function f from the H approximates the dataset?
A: H could be a set of all possible KNN candidates, which means 1-NN, 2-NN, 3-NN etc.. And we believe the true function will reside in this space. Hope that clarifies.
k is a hyperparameter of kNN, so the hypothesis set is dependent on this as well as the distance metric. kNN's decision surface also highly depends on the data itself (as you know).

3. **Q**: Why does a high dimensional space not work for kNN so well?
   A: The problem is that in high dimensions, distances between points are generally far apart from each other and hard to discriminate between them. This *Curse of Dimensionality* is real and a problem and makes distance metrics in high dimensions generally less useful.

4. **Q**: Why is Manhattan distance (L1) better than Euclidean distance (L2) in our credit assignment problem?
   A: We have lots of answers from our class so you can check it out on the thread there. Generally, it's clear that since the dimensions (features) in the credit assignment problem are very different from each other, it's not very proper to do comparison across dimensions, which is what Euclidean distance does. It's more proper to a comparison of each features separately and combine them together afterwards (more akin to L1).
   And don't worry if you don't understand L1, L2 or Lp in general, these are just different (precise) math notations for the class of Lp norms:
   https://en.wikipedia.org/wiki/Lp_space. The Wikipedia page is too formal for what we need, but we'll cover the essential elements of this later Week 06A (Regularization and Validation).

Week 03 – Decision Trees

1. **Q:** When we decide what's the best question in decision trees, can we choose the questions in the order that will make the tree shortest?
   A: This is possible but expensive to do. Searching for the shortest (fewest tests for any path needed) is a combinatorially expensive problem (as any of the 2^n orderings could possibly lead to the shortest.
   For this reason the decision tree learning (DTL) algorithm uses a greedy search, using **information gain** as the criterion.
   This is also related to the reason why we build DTs fully before pruning them, instead of growing them to a desired size / depth. Can you guess why ? 🤔 [Hint: you can look up the *horizon effect* which applies to these types of general search problems]

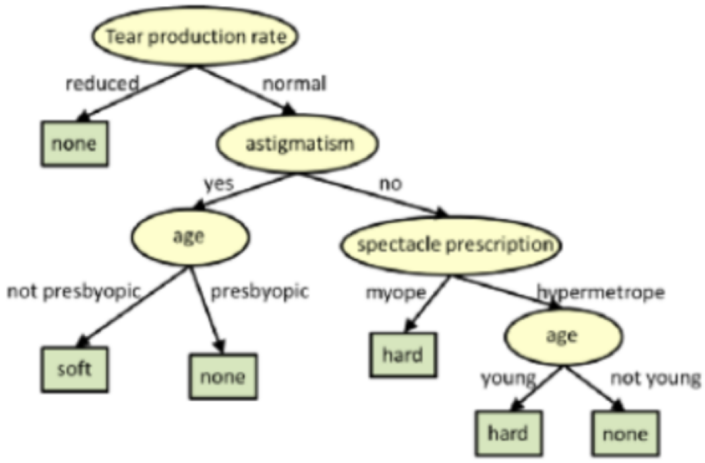
Q': Ahh I see. I think using this greedy algorithm makes the tree suffer from potential horizon effect. So when we build a full tree and back trace from the leaves, I guess we can use some of the info to re-evaluate whether the greedy choice is a good choice. So sort of mitigate horizon effect?
A': Yes, that's correct!

2. **Q**: I was wondering since you mentioned that we can't have negative values for information gain i.e. no info loss. Since we derived it from entropy , does this tie to the 2nd law of thermodynamics.
   A: Not really. We can observe this as picking a criterion for a test is akin to partitioning the (current) input dataset for the decision stump's problem. A criterion splits the original set into C_i subsets. Overall the sum of the entropy of the subsets must either conserve or lessen the entropy.
   Quick question check for you, though. Can a split subset have higher entropy than its parent?

3. **Q**: May I clarify what does the MODE() mean in line 3?

   if *examples* is empty **then return** *default*
   else if all *examples* have the same classification **then return** the classification
   else if *attributes* is empty **then return** MODE(*examples*)

   I saw some explanation online which seems treat this case as *None*. But what will happen if there is a test data go into this branch?



Original Tree

   A: This refers to the statistical Mode of the examples, i.e., the majority label, in contrast to other "averages", such as Mean or Median.

4. **Q**: Regarding the time complexity of the decision stump, I'm not sure how binary search works(hence how it reaches n * log(m)), since splitting in one way does not tell you which direction the split should go?

A: Indeed we don't know which direction improves the decision stump's threshold but this is an easy constant test. Once you know a direction improves or worsens performance you can do binary search to find an optimum value.
Note this is in the case of a binary target (y) value and not for native multiclass output.

5. **Q**: For the Test cost mentioned in the lecture, how do we calculate them in reality? I mean in most cases, I guess we are not able to have all the possible values for x.

A: You're absolutely right! You will never have all possible values of x . However, you can "estimate" test cost using validation.
Sometimes we may be lucky enough to have a probability distribution estimate of types of x that we will come across. For example, in surveys of the public we know that seniors are usually underrepresented, and that young working adults are usually overrepresented. In this way, we can sometimes have an idea of the distribution of possible x s, our \scriptX.

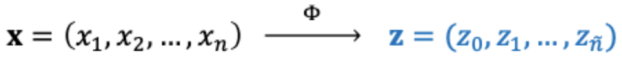
6. **Q**: Regarding ensembles, is it safe to think of each report ht(x) as a basis for the vector space h(x) and how does report ht(x) exactly differ from h(x) in usage and definition? Thanks
A: *Answered in AMA on 9 Sep (Week 05B).*
(TL;DR) [Not exactly. Sometimes the ensemble can pick out hypotheses that are not in the original \scriptH hypothesis space of \scriptA]

Week 04 – Linear Models

1. **Q**: Are there any textbook/reference materials for lectures 4, 5, and 6?
A: There are plenty of other good books on ML, but most go heavy on the math since the professors and lectures are in the math or stats department. These include the classics:
    - Hastie, Tibsharani and Friedman, The Elements of Statistical Learning
    - Bishop, Pattern Recognition and Machine Learning

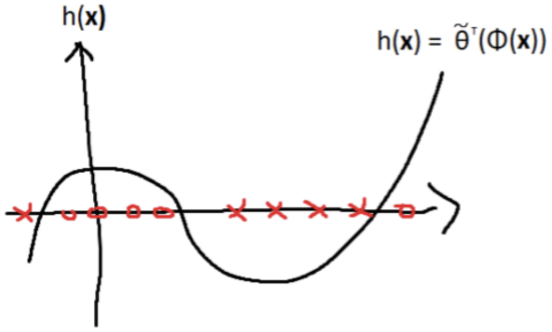
   I don't have the aligned chapters offline for the lectures but can look into that if someone wants. As an aside, the "simpler" version of the famous Hastie, Tibsharani and Friedman, The Elements of Statistical Learning is called *An Introduction to Statistical Learning*, it is free and written by the same authors (https://www.statlearning.com/)

2. **Q**: Hi everyone, for the relationship between n~ and n, I know that n~ is not necessarily equal to n but can n~ > n? If n~ is more than n will some of the features in Z space be linearly dependent on each other?



A: The point of non-linear mapping is to transform the decision boundary non-linearly!

You choose the mappings to make. If they were linearly dependent, it means you should just "save" on that transformed variable and remove it since it can be represented by a linear combination of the other features. Consider your data as just 1 dimensional represented by the red points. You could transform it to a (1, x, x^2, x^3) feature space and form a such a decision boundary. All of the features here are linearly independent. If for some reason you transform it to (1, x, 2*x, x^2, x^3), you'd might as well remove the 2*x feature.



n does not have to be the same is *n~* and can be either larger or smaller.
The example is one where we take one point *x* and map to many (1, x^2, x^3) .
In lecture, an example of MNIST (x_1, ..., x_256) being mapped to (z_1, z_2) where these two variables are intensity and horizontal symmetry.

3. **Q**: May I ask regarding the slides for linear regression, why is it that we transpose the weight instead of the vector of the gradient? (Line 3) Thank you!

$$\Delta L_{train} = L_{train}\big(\theta(t+1)\big) - L_{train}\big(\theta(t)\big)$$
$$= L_{train}(\theta(t) + \eta v) - L_{train}\big(\theta(t)\big)$$
$$= \eta \nabla L_{train}(\theta(t)^{\top}) v + O(\eta^2)$$
$$\geq -\eta \left\| \nabla L_{train}\big(\theta(t)\big) \right\|$$

A: Indeed there is no difference. We need to just multiply the gradient of L_train (a direction and hence represented by a vector) against the v (which represents the direction of the step we want to take). We minimize the value (i.e., on the final line the inequality becomes equality, when we set v to be exactly the value on the final line; opposite the direction of the gradient). I will issue a replacement slide for this in lecture slide update later and announce it.

4. **Q:** I have some questions regarding the math in the tutorial. For example the f hat here refers to the learned hypothesis or does it refer to the average hypothesis that is introduced in the lecture?
A: The ^ on a variable in statistics usually refers to approximation. So in tutorial, this f^hat symbol means an approximation of the function f . So this is a synonym for h(x) that we used in our lecture up to now.

Week 05 – Bias and Variance

1. **Q**: We know bias as how well our hypotheses set can approximate the target function. This implies that high bias leads to *underfitting*, since our model isn't able to capture the appropriate features in the data. However, when talking about deterministic noise, we only talk about how it leads to *overfitting* together with stochastic noise. We say that the higher the deterministic noise, the more likely we are to overfit. During the previous lecture, you introduced deterministic noise as equivalent to bias. Doesn't this contradict each other?
   A: This is a common confusion and the root of the problem is the use of the word bias. Bias can refer to the modeling power of the model, or to the amount of error due to the model.
   Models with high bias have a lot of parameters (think \scriptH_10 compared with \scriptH_2.). Because they have lots of parameters they fit the signal well, possibly too well, and so they have little error due to their high bias.
   I.e., high bias models exhibit low bias error, but high variance error.
   As a corollary, low bias models (e.g., \scriptH_2 when compared with \script_H10) exhibit high error due to bias (underfitting error; or deterministic error).
   Deterministic noise here is the same as noise due to underfitting, where the model that we use is not powerful enough (that is, its average hypothesis h, does not give a good approximation to the signal). This is not due to the sampled dataset, since even if we pull a new dataset from the true distribution \scriptX.
   So on [Slide ~61, Week 05a], deterministic noise here is being equated with the associate bias error. A high bias model (with lots of capability to fit well), will exhibit low bias error (deterministic noise).

2. **Q**: As below:

This document is view-only accessible via http://www.comp.nus.edu.sg/~cs3244/AY2122S1/faq.html. To ask a question, please ask your #tg leader, or write directly in the component channel (#project, #assignment, #assessment, #lectures, #projects, #tutorials)

> Hi guys, guest here, as usual, please ignore if not important, but I have been "stuck" in the Bias-Variance Decomposition for a week now. Hope to reaffirm my understanding here.
>
> Consider the general regression setup where we are given a random pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$. We also assume we know the true/target function f which establishes the true relationship between X and Y. Assume X and the stochastic noise ϵ is independent.
>
> With a fixed point x_q (the test point is univariate and have one single point only), can I understand the following:
>
> > **Variance: My understanding**
> >
> > > Imagine I built multiple hypothesis h_i using multiple datasets $\mathcal{D}_i$ sampled from the same distribution $\mathcal{P}$ say a uniform distribution $[0, 1]$, and given a fixed point x_q which is from the test set (unseen test point), then can I understand intuitvely that the **variance** of the model over that fixed point x_q to be the **total sum of the average mean squared deviation** of each h_i from the average hypothesis $\bar{h}$, where the latter is the mean prediction made by $h_i(\mathrm{x}_q)$, further divided by the number of hypothesis we have (since we are taking two expectations here.
>
> ---
>
> > **Bias: My understanding**
> >
> > > Same setting as above, **bias** of the model over a fixed point x_q to be the **squared error** of $f(\mathrm{x}_q)$ and $\bar{h}(\mathrm{x}_q)$. In particular, if x_q is has m samples, then we need to sum the squared error of each individual test points and divide by the number of test points.

A: Yes, that's right. In lecture, we went over the bias and variance decomposition in this specific easy case of squared error with Gaussian noise of mean 0, standard deviation 1. Your analysis does the same.
Bias is the distance from the best (average) hypothesis, h-bar, to the target function and variance is the [expected] distance from the actual learned h from the average hypothesis h-bar.
Aside: This is related to the reason why variance is reduced by factor 1/sqrt(K) when we use a larger validation set. We have two expectations D_k and test point x to integrate over, but because points are drawn i.i.d., the cross terms drop (See related question from Week 06 below).

Week 06 – Regularization and Validation

1. **Q**: Can I just double confirm for the k-fold validation: so when we have 10 folds, we leave 1 fold out as test set and use the remaining 9 folds for training & validation then repeat for all folds?
   A: Yes, that is correct. For k-fold, it usually connotes that each fold is completed separated from the other folds; i.e., the dataset is completely partitioned into k parts of equal size.
   We repeat the experiment k times where each time we use one fold for validation and the remaining k-1 folds of data for training our h-minus classifier.
2. **Q**: should h_D(x) be h_D_k(x) in the summation?
   Imagine many, many data sets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ drawn:

$$\bar{h}(x) \approx \frac{1}{K} \sum_{k=1}^{K} h_{\mathcal{D}}(x)$$

   A: Yes, good catch! That's a 🐛 in the slide.

3. **Q**: I don't understand how variance gets reduced by a factor of 1/K (Slide 34)
   A: See https://math.stackexchange.com/questions/1708266/why-square-a-constant-when-determining-variance-of-a-random-variable TL;DR: When you multiply your data by a constant (here, K), the deviations also get multiplied by that constant, so the squared deviations get multiplied by the square of that constant.

#tutorials

Currently there are no questions in the general tutorial channel. TAs, please feel free to add your own questions and answers about tutorials here.

#assignments

For our one-and-only Assignment #1, please refer to the pinned post in Slack in the #assignments channel: 📣 https://cs3244-2110.slack.com/files/T024V4R9J7L/F02DBL7U0UU

#assessment

1. **Q:** I would hope to ask regarding the rule that we are allowed to access "annotated notes." I make some pop-up notes (if not moving mouse over it, it is shown as a small yellow note symbol) on the lecture slides which is viewed using Adobe Reader, and hope to ask if viewing those pop-up notes I made on lecture slides is allowed during the exam and is considered annotated notes?
   A: Yes this is fine!

2. **Q:** I hope to ask if hand-written notes(on foolscap paper) and printed or electronic PDF notes like tutorial and past year papers are all allowed to be used for the exams, or some of them is allowed? This is because I am not sure if "pre-compiled notes" means only what written by oneself is allowed, and the exam is not open-book but open lecture notes..
   A: Yes this is allowed too.

3. **Q**: Is backward navigation allowed during the midterm?
   A: Yes.

4. **Q**: I hope to ask if hand-written notes(on foolscap paper) and printed or electronic PDF notes like tutorial and past year papers are all allowed to be used for the exams, or some of them is allowed? This is because I am not sure if "pre-compiled notes" means only what written by oneself is allowed, and the exam is not open-book but open lecture notes.

A: Yes this is allowed.

5. **Q**: Are we allowed to use any programming tools during midterm (e.g. Can I write a python program on my computer to calculate, lets say, entropy, for me)?
   Q': Are we allowed to have a terminal / colab window open with `numpy` imported to do the matrix multiplication questions from the paper
   A: No, this is not allowed. Your midterm will not feature calculation questions that require significant computation beyond what is needed for a standard scientific calculator.

6. **Q**: Hi Prof, are we allowed to use electronic devices like iPad to view annotated lecture notes?
   A: Unfortunately the answer is no, we need to be able to see and record your screen. You can only use physical printouts or your primary test taking device to take the midterm. You may use the ipad to view your annotated lecture notes if and only if you are taking the exam on your iPad and you are recording your screen. If you need to refer to your notes, please print them out or view them on your primary test taking device.

7. **Q**: In light of the recent 26 Sep announcements, is the physical midterm examination venue still allowed?
   A: Yes. Our use of the i3 Auditorium for the midterm assessment is still within the current bounds. So those who have opted for the physical exam are still allowed.

8. **Q**: Are calculators/graphic calculators allowed for the midterms?
   A: Yes, physical calculators that don't have any other function (phones and tablets are not allowed) are allowed. To be clear, the midterm will not have questions that require much calculation, so it should be comfortable to take by hand.

9. **Q:** Are we allowed to use PDF textbooks during the exam too?
   A: Yes, only on your primary test-taking device.