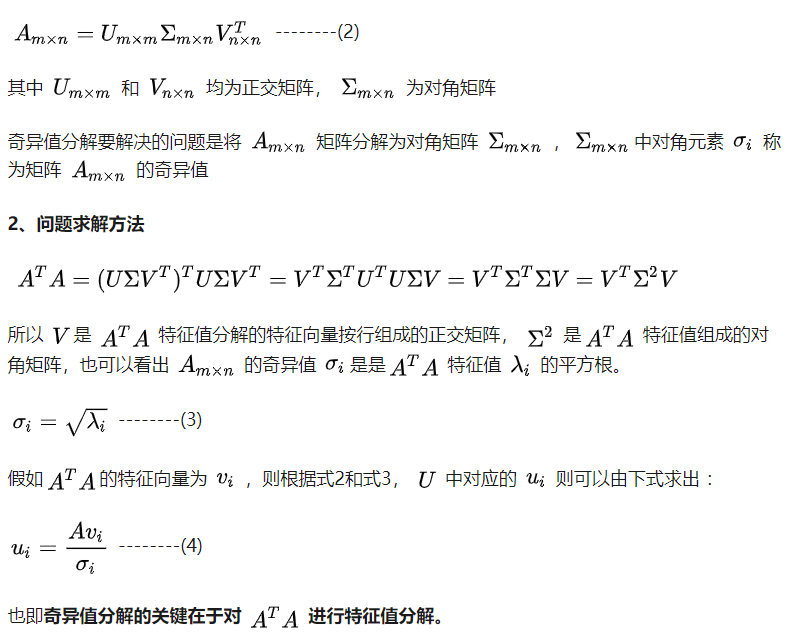
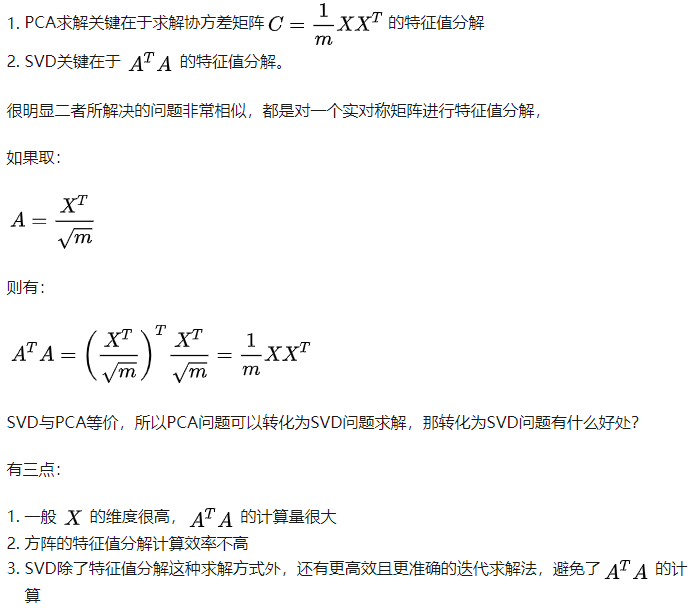
**Review Questions for IEEE Data Mining 2019 Spring (Part Two)**

1) PCA is an example of dimensional reduction method; give a full derivation of PCA with respect to its eigenvectors; explain SVD and how it is used to solve PCA

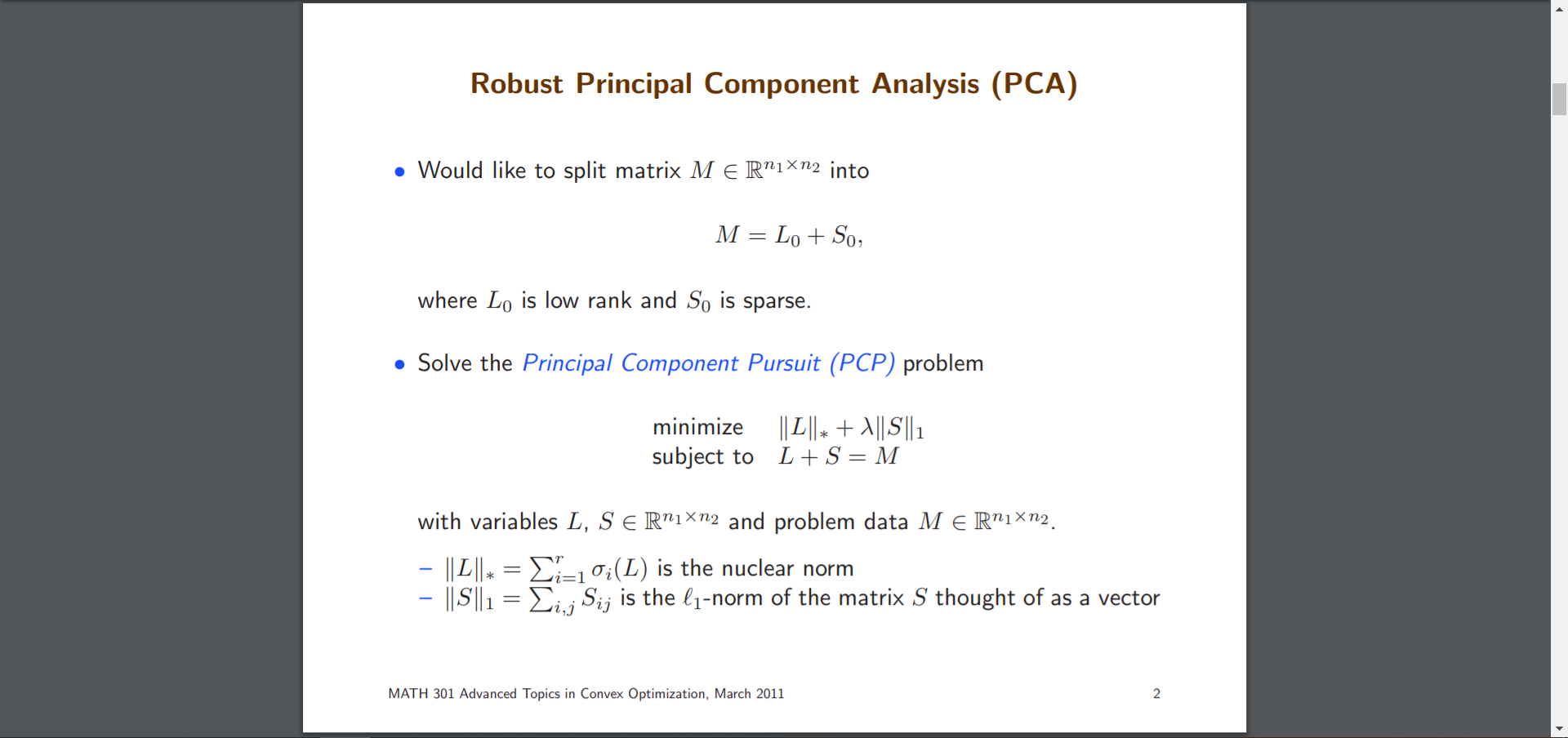
* PCA推导
  + 西瓜书（P230-231）
  + 南瓜书（<https://datawhalechina.github.io/pumpkin-book/#/chapter10/chapter10>）
* SVD（<https://zhuanlan.zhihu.com/p/58064462>）

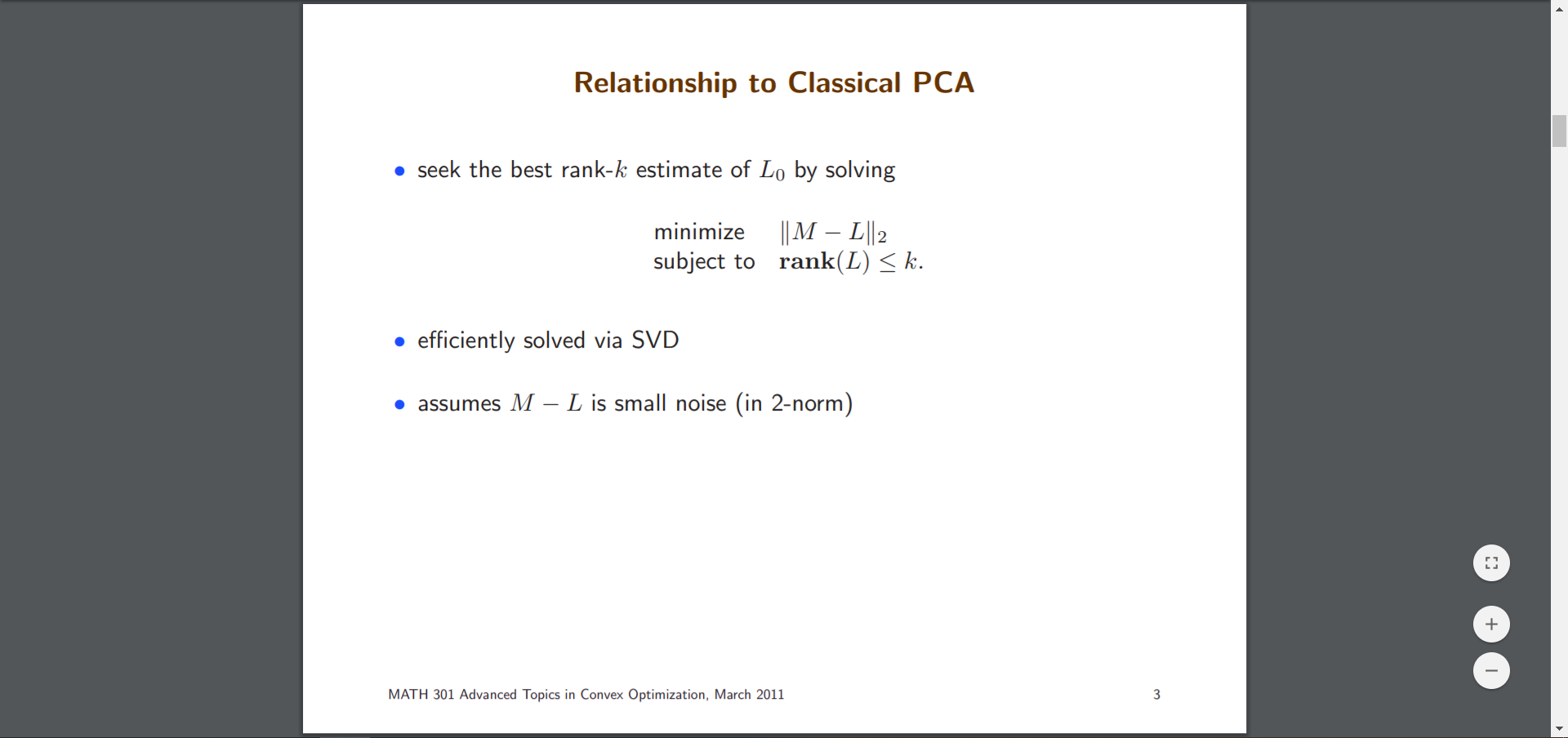


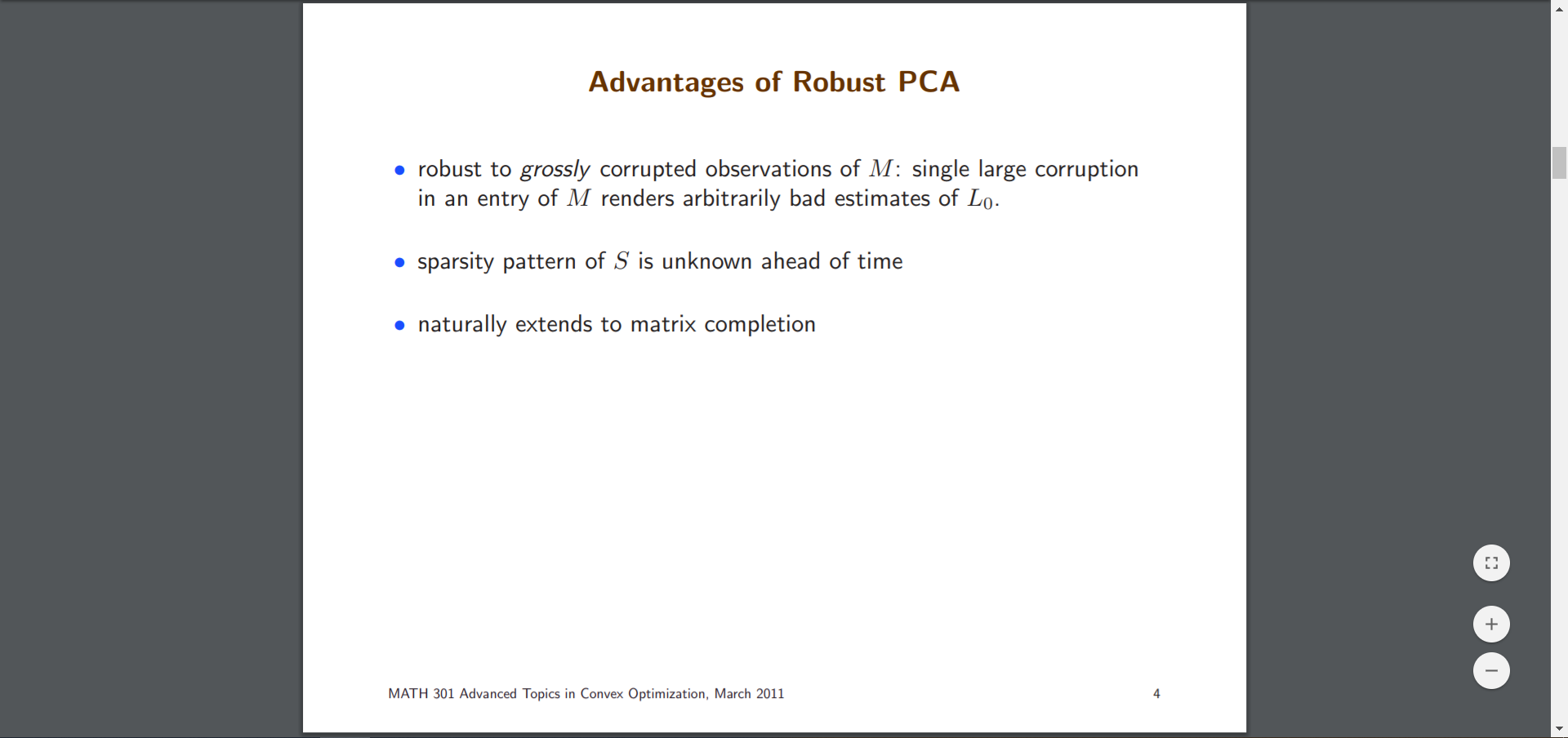
* SVD和PCA的关系（<https://zhuanlan.zhihu.com/p/58064462>），m为样本总数



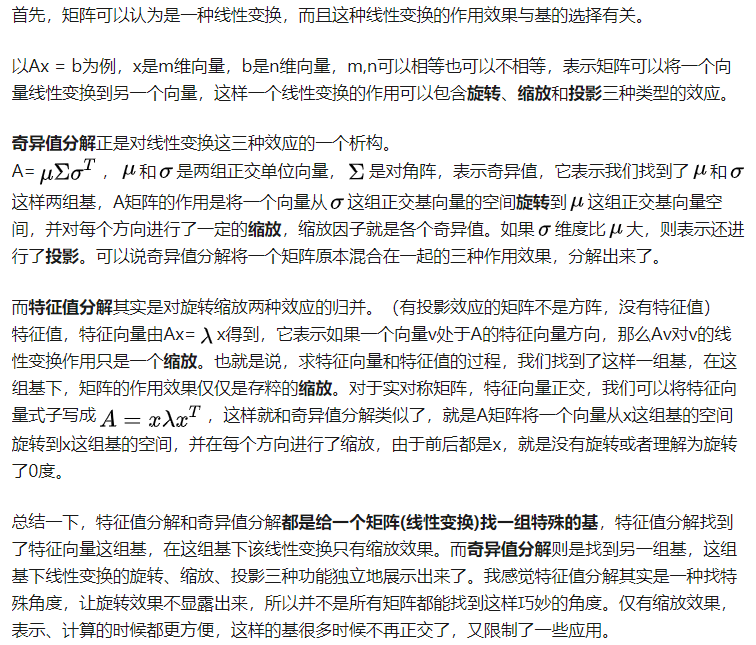
2) Compare regular PCA with the low-ranked PCA, what would be advantage using the low-ranked PCA and how it is formulated? ( ref: <https://statweb.stanford.edu/~candes/math301/Lectures/rpca.pdf> )







3) What is the difference between a singular value and its Eigen value? Explain the resulting singular values of an SVD for how the features were originally distributed;



4) What is the key motivation (and contribution) behind deep learning, in terms of data representation?

5) Compare the advantage and disadvantage of using either sigmoid or ReLu as an activation function?

ref: <https://www.jiqizhixin.com/graph/technologies/1697e627-30e7-48a6-b799-39e2338ffab5>

* ReLU:
  + 优点：
    - 相比起Sigmoid和tanh，ReLU在SGD中能够快速收敛。
    - Sigmoid和tanh涉及了很多很expensive的操作（比如指数），ReLU可以更加简单的实现。
    - 有效缓解了梯度消失的问题。
    - 在没有无监督预训练的时候也能有较好的表现。
    - 提供了神经网络的稀疏表达能力。
  + 缺点：
    - 随着训练的进行，可能会出现神经元死亡，权重无法更新的情况。如果发生这种情况，那么流经神经元的梯度从这一点开始将永远是0。也就是说，ReLU神经元在训练中不可逆地死亡了。
* Sigmoid
  + 优点：
    - 映射在(0,1)之间，单调连续，输出范围有限，优化稳定，可以用作输出层。
    - 求导容易。
  + 缺点：
    - 1.由于其软饱和性，容易产生梯度消失，导致训练出现问题。
    - 2.其输出并不是以0为中心的。

6) Discuss matrix decomposition as a strategy to solve a complex high dimensional problem into a hierarchy of lower dimensional combinations?

8) Why normally we use L2 for the input layers and L1 for the actual modeling? Explain why still sigmoid activations are still used for the output layers?

* 问题一：
  + L2 smooth the input, make the input not too different to each other (reduce bias);
  + L1 sparsity, only important dimensions maintain;
* 问题二：sigmoid re-normalize re-categorize all outputs into something zeros and ones (probabilistic); by the way, tanh is better.

9) What would be the true features of an object modeling problem? Give two examples to highlight the importance of selecting appropriate dimensions for feature representations;

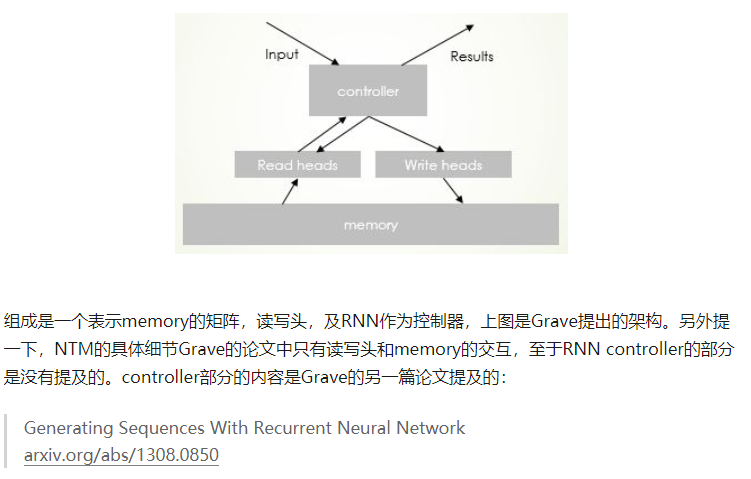
10) Why does the feature decomposition in deep learning then a topological recombination could make a better sampling? What would be the potential problems making deep learning not a viable approach?

11) Explain the importance of appropriate feature selection being compatible with model selection in the context of model complexity;

12) What would be the ultimate and best representation for a high dimensional and complex problem? How this might be possibly achieved?

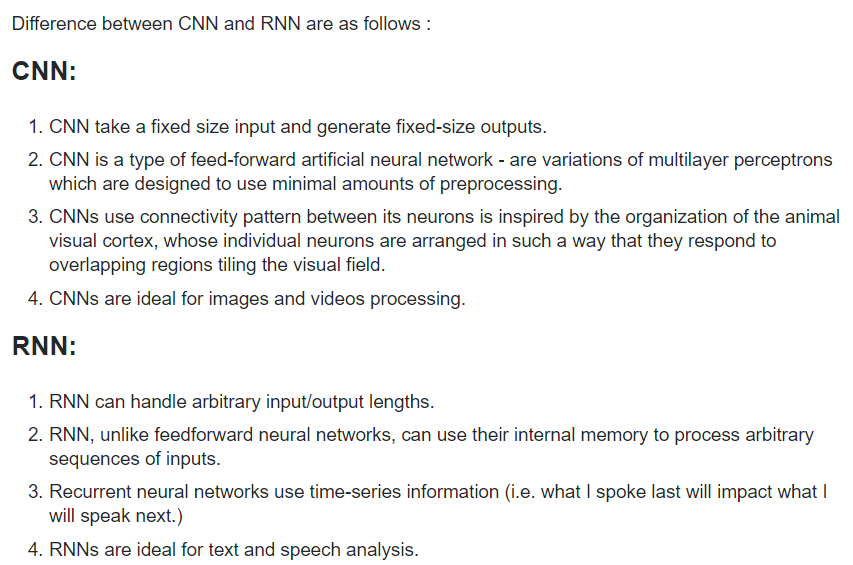
13) How RNN can be expanded our learning to fully taking advantage of Turing machine? What RNN can whereas CNN cannot do?

* 神经图灵机（<https://www.zhihu.com/question/42029751>）
  + 图灵机就是一种简单的计算机模型。正如现代计算机一样，其思想中也包含了一个外部存储器和某种处理器。本质上，图灵机包含上面写有指令的磁带和能够沿着磁带读取的设备。根据从磁带上读取到的指令，计算机能够决定在磁带上不同的方向上移动以写入或者擦除新符号等等。
  + 神经图灵机的本质是一个使用外部存储矩阵进行attentive interaction机制的RNN，由于定义的RNN各个部分都是可导的，使得输入训练数据通过机器学习（back propagation加gradient descent）训练“程序”成为了可能。架构如图：

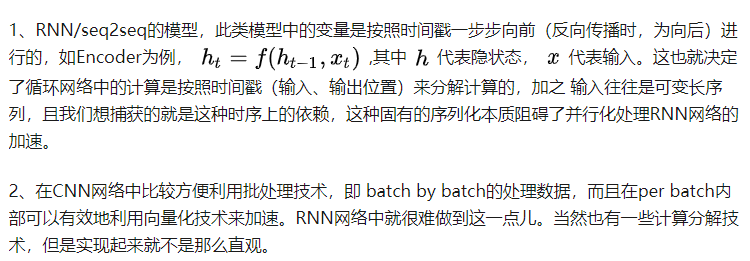


* RNN can handle sequential data while CNN cannot.

可参考<https://datascience.stackexchange.com/questions/11619/rnn-vs-cnn-at-a-high-level>



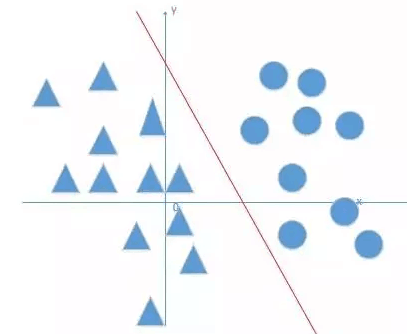
14) What is the central additional difficulty of RNN compared to CNN?



都会存在梯度消失问题：CNN因为网络层数太多导致，RNN因为时间迭代次数导致，都是因为链式求导次数太多。

15) In the activation function, there is a constant term “b” to learn, why it is important?

为什么线性模型要加 bias？答案很简单，不加 bias 你的分类线(面)就必须过原点，这显然是不灵活的。有了bias我们就可以上下左右移动我们的线了。神经网络是一样的道理。



16) LSTM integrate short and long term processes, what is the central issue to address to achieve at least some success?

17) The difference between value-based vs. policy-based gradients?

ref: <https://ai.stackexchange.com/questions/6196/what-is-the-relation-between-q-learning-and-policy-gradients-methods>

18) Why dynamical programming might not be a good approach to select an optimal strategy?

19) Explain an expectation-based objective function?

20) Explain Haykin’s universal approximation theorem;

**General problems:**

1. In learning, from the two key aspects, data and model, respectively, what are the key issues we normally consider in order to obtain a better model?
2. Describe from the classification, to clustering, to HMM, to more complex graphical modeling, what we are trying to do for a more expressive model?
3. What are the potential risks we could take when trying to perform a logistic regression for classification using a sparsity-based regularization?
4. Give five different structural constrains for optimization with their corresponding scalars;
5. Give all universal, engineering, and computational principles that we have learned in this course to obtain both conceptually low-complexity model and computationally tractable algorithms?
6. Why data representation is at least equally as important as the actual modeling, the so-called representation learning?
7. How does the multiple-layer structure (deep learning) become attractive again?
8. Discuss Turin Completeness and the limit of data mining;
   1. 图灵完备性：<https://en.wikipedia.org/wiki/Turing_completeness>
   2. 数据挖掘的极限：<https://content.wisestep.com/data-mining-purpose-characteristics-benefits-limitations>
9. Discuss general difficulties of using gradient for composite functions or processes;
10. What is the trend of machine learning for the next 5-10 years?

Part Three - Previous Exam:

1) SVM is a linear classifier with a number of possible risks to be incurred, particularly with very high dimensional and overlapping problems. Use a simple and formal mathematics to show and justify (a) how a margin-based liner classifier like SVM can be even more robust than Logistic regression? (b) how to control the overlapping boundary?

2) Why a convolution-based deep learning might be a good alternative to address the dilemma of being more selective towards the features of an object, while remaining invariant toward anything else irrelevant to the aspect of interests? Why a linear regression with regulations would result in features which are usually conceptually and structurally not meaningful?

3) There are a number of nonlinear approaches to learn complex and high dimensional problems, including kernel and neural networks. (a) please discuss the key differences in feature selection between these two alternatives, and their suitability; (b) what are the major difficulties using a complex neural network as a non-linear classifier?

4) For any learning problems, (a) why a gradient-based search is much more favorable than other types of searches? (b) what would be the possible ramifications of having to impose some kinds of sequentiality in both providing data and observing results?

5) Please use linear regression as the example to explain why L1 is more aggressive when trying to obtain sparser solutions compared to L2? Under what conditions L1 might be a good approximation of the truth, which is L0?

6) What is the key difference between a supervised vs. unsupervised learnings (where we do not have any ideas about the labels of our data)? Why unsupervised learning does not guaranty a global solution? (use mathematical formulas to discuss).

7) For HMM, (a) please provide a Bayesian perspective about the forwarding message to enhance an inference (using a mathematical form to discuss); how to design a more generalizable HMM which can still converge efficiently?

8) Using a more general graphical model to discuss (a) the depth of a developing prior-distribution as to its contribution for a possible inference; (b) how local likelihoods can be used as the inductions to facilitate the developing inference?

9) Learning from observation is an ill-posed problem, however we still work on it and even try to obtain convex, linear, and possibly generalizable solutions. Please discuss what key strategies in data mining we have developed that might have remedied the ill-posed nature at least in part? Why in general linear models are more robust than other more complex ones?

10) Using logistic regression and likelihood estimation for learning a mixture model (such as the Gaussian Mixture Model), please using Bayesian perspective to discuss the differences and consistencies of the two approaches; why logistic function is a universal posterior for many mixture models?