**Learning and Search:**

1. **W**hy all learning problems are inverse problems, requiring unbounded exhaustive searches, thus ill-posed?
   1. 穷举，每一个问题都是无界的
2. What is the key mathematical assumption that we count on in order to search? What would be the general implications for such an assumption of continuality or locality?
   1. 利用路径，寻找最好的路径，唯一的路径：**梯度（微分）**
   2. 全局搜索，局部光滑和可导
3. What is the generalizability of a mathematical process, from both expressive and inclusive point of views?
   1. 所有的变化和差异都可以响应出来，线性、光滑的范围
   2. 较大适用的范围和空间，所有的x都有对应的y
4. What would be some of the solutions for such a ill-posed problem to yield at least some reasonable results?
   1. 约束、正则化、降维、梯度、稀疏、凸问题、模型，将一个复杂的问题低复杂
5. What are some of the mathematical hurdles that have prevented more generalizable solutions?
   1. 大部分问题不是欧式问题，不在欧氏空间中（空间维数，距离，测度）
   2. 微分 三阶非线性出现的奇异性和局限性
6. Why variable dependences (interactions) could become an extremely difficult and even impossible problem? Give philosophical, mathematical, physical, computational, and numerical examples for such a singularity.
   1. 真理不可证明
   2. 微分的奇异性、费马大定理
   3. 三体问题
   4. 图灵序，并行（没有真正的并行）
   5. 初值敏感
7. Why a Euclidian problem would be most favored but usually impossible to obtain for a real world issue?
   1. 真实问题高维变量不正交、测度不唯一不统一，不可比，不是欧式问题
8. What are some of the key requirements for a real issue to be formulated as a Euclidian problem?
   1. 正交的、可比的 --> 数学问题
9. What would be the mathematical alternative frameworks to translate a non-Euclidian problem to mathematically appropriate solutions?
   1. 二范数smooth一下，约束一下值的大小
   2. 降维，如用PCA保证正交
   3. 升维，非线性kernel变换
   4. 没有办法变成欧式问题，概率图，图模型来describe变量之间的联系
   5. feature分解后重新组合
10. Why we would prefer a low complexity model for a high complex problem?
    1. 模型过于复杂时，training error降低，但testing error会升高
    2. 低复杂度模型抓住了主要的东西
11. What is the loss function? Give three examples (Least Square, Logistic, Hinge) and describe their shapes and behaviors;
    1. Least square: assume 高斯
    2. Logisitic：对数据分布不敏感
    3. Hinge loss: c
12. Using these losses to approach the actual linear boundary, inevitably some risks will be incurred; give two different approaches to remedy the risk using the SVM-based hinge loss as an example;
    1. Kernel
    2. 把boundary变大
13. *How many possible models are there given a set of training data? What is the key assumption of PAC learning for model selection?*
14. Describe biases and variance issue in learning, and how can we select and validate an appropriate model?
    1. 多次试验找到testing error最小
15. How to control model complexity in linear and logistic regression? Are there supposed to be a unique low-dimensional model for a given high dimensional problem?
    1. 降维，一范数（参数稀疏）、二范数（光滑）
    2. 不是，跟p、n和数据分布是有关的，降得太多模型就失真了
16. Using the Least Square as the objective function, we try to find the best set of parameters; what is the statistical justification for the Lease Square if the underlying distribution is Gaussian?
17. Could you describe the convexity as to how it would facilitate a search? Using the Least Square-based regression and Likelihood-based estimation as the examples?
    1. Global solution梯度搜索逼近
18. Gradient Decent has a number of different implementations, including SMO, stochastic methods, as well as a more aggressive Newton method, what are some of the key issues when using any Gradient-based searching algorithm?
19. What are the five key problems whenever we are talking about a learning process (Existence, Uniqueness, Convexity, Complexity, Generalizability)? Why are they so important?
20. Give a probabilistic interpretation for logistic regression, how is it related to the MLE-based generative methods?
    1. 概率分布
21. What are the mathematical bases for the logics regression being the universal posterior for the data distributed in any kinds of exponential family members?
22. Can you provide a probabilistic comparison for liner and logistic regression?
    1. Logistic 边界
23. Why the log of Odd would be something related to entropy and effective information?
24. Why often we want to convert a liner to a logistics regression, conceptually and computationally?
25. Compare the generative and discriminative methods from a Bayesian point of view?
26. What are the most important assumption for something Naïve but still very effective? For instance for classifying different documents?
27. What would be the most effective way to obtain a really universal prior? And what would be the most intriguing implications for human intelligence?
28. //For the regular and multinomial Naïve Bayes, what are their key assumptions? Why the multinomial method can be more context sensitive?
29. What are the key advantages of linear models? What are the key problems with the complex Neural Network?
    1. 数学上的简单，无奇异性、边界更直观
    2. 局部解、高阶奇异性
30. What are three alternatives to approach a constrained maximization problem?
31. What is the dual problem? What is strong duality?
    1. 对偶问题
    2. 跟顺序、路径无关
32. What are the KKT conditions? What is the key implication of them? Including the origin of SV?
33. What is the idea of soft margin SVM, how it is a nice example of regularization?
    1. 降维
34. What is the idea of kernel? Why not much additional computational complexity?
35. What is the general idea behind the kernel? What key computation do we perform? Why is it so general in data modeling?
    1. 线性不可分->线性可分
36. Why we often want to project a distance “measure” to a different space?
    1. 变成欧式问题
    2. 平衡所有的feature

**Probabilistic graphical model:**

1. Compare the graphical representation with feature vector-based and kernel-based representations;
   1. Graphical把复杂的因果逻辑关系表示出来
   2. vector-based and kernel-based把参数分开，只考虑怎么分布、边界
2. Explain why sometime a marginal distribution has to be computed in a graphical model;
3. Why a graphical model with latent variables can be a much harder problem?
   1. 只能保证局部最佳，不能全局最佳
   2. 非凸，有很多的极值点
4. What is the key assumption for graphical model? Using HMM as an example, how much computational complexity has been reduced because of this assumption?
   1. 马尔可夫链
   2. 有马尔可夫链假设，复杂度 没有复杂度
5. Why does EM not guarantee a global solution? What is a simple proof for that?
   1. EM marginalized-estimation global solution需要穷举组合可能性
   2. Likelihood期望最大化
6. Why is K-mean only an approximate and local solution for clustering?
   1. Local解，也是EM算法
7. How to interpret the HMM-based inference problem from a Bayesian perspective, using the forward/backward algorithm?
   1. Forward t->t-1 先验 后验变成后一个的先验 human的方法
   2. Machine
8. Show how to estimate a given hidden state for a given series of observations using the alpha and beta factors;
9. How a faster inference process would be constructed, given a converging network?
10. How can an important node be detected using an alpha and a beta process?
11. Why often an alpha process can be more important than beta?
12. What are the key differences between an alpha and a beta process from human and machine intelligence point of views?
    1. Alpha是对likelihood的优化，beta是对data-likelihood的优化
13. How data would contribute to the resolution of a inference process from a structural point of view?
14. For a Gaussian graphical model, what is the implication of sparsity for such a graphical model? How is such sparsity achieved computationally?
    1. 高斯模型聚类->PCA
15. Why a Frobenius norm would be better than L1 or L2 norms?

**Dimension reduction and feature representation:**

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;
   1. PCA推导过程

2) Compare regular PCA with the low-ranked PCA, what would be advantage using the low-ranked PCA and how it is formulated?

3) For a low rank-regularized PCA, what would be the limit of dimension reduction for a given p and n of your data?

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

5) What would be the true features of an object modeling problem? 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?

这样可以使component分开，

deep learning over-fitting

1. 局部解，参数的收敛不是全局的，而是局部的，所有参数耦合在一起是impossible的
2. 无法讨论更多层参数的耦合，（马尔可夫链）
3. 高阶微分的奇异性，初值敏感 🡪 将初值定为线性

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

model selection

feature selection feature是问题的维数，维数太高需要降维，用更少的feature describe问题

p和n的关系

7) What is the key motivation behind a kernel-based method in data representation?

Kernel-based是overfitting，效果很好，但不generalizable

（**Kernel与其他方式比较，优缺点）**

8) What would be the ultimate and best representation for a high dimensional and complex problem?

拓扑结构，结点表示问题的组合，minimize结构，抓住主要的部分

9) Give two examples to highlight the importance of selecting appropriate dimensions for feature representations.

脸的位置，Pixel从几万维降到三维，就是三维重构的问题

encoder problem： 2^3 truth table 3种参数组合产生

10) For a typical big data problem (p>>n), what considerations we will have to take when trying to select an appropriate model (for instance, //to perform a SVM)?

不需要推导，视具体的p和n而定

范数、点积度量n之间的差别是否足够大，得到有效的n

generalize

**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?

数据要降维

1. Why all machine learning problems are ill-posed?

既要有表达的能力，又要有泛化的能力

1. Describe from the classification, to clustering, to HMM, to more complex graphical modeling, what we are trying to do for a more expressive model?
2. What are the potential risks we could take when trying to perform a logistic regression for classification using a sparsity-based regularization?
3. What are the potential risks we could take when trying to perform a linear regression using a sparsity-based regularization?
4. Give five different structural considerations a search can be constrained with corresponding simple scalars;

什么假设，什么结构

一范数降维、二范数smooth input

1. 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?

Locality

尽可能用线性的东西

对凸的追求

点积聚类，代表距离

1. Why data representation is at least equally as important as the actual modeling, the so-called representation learning?

挑选出真正的feature

1. How does the multiple-layer structure (deep learning) become attractive again?
2. What is the trend of AI research and development for the next 5-10 years?

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?

数学语言描述

SVM是一个点积测度的问题，LR的x是在统一的距离测度下，

1. kernel kernel的风险
2. 用c控制 升维

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?

降低局部信号的强度，但提升了整个结构的强度

通过回归做范式的方法，得到的特征不是真正的特征

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?

把kernel和NN具体写出来，Kernel把非线性问题通过映射升维，把数据人为变化，NN试图找出泛化，kernel动x，NN动空间

局部解、马尔可夫假设、高维作用关系

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).

杨森不等式

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?

深度，浅说明问题比较简单，深说明数据没有起到什么作用，不匹配，模型和问题本身要comparable

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?