Review For Data Mining Class, 2019 Spring

Author: Han Bing.

Answer Support(Orderd by Name):

Chen Wang, Chang Feng, Cao Hengkui, Han Bing, Jiang Yuxuan, Qinyu, Sun Rongyi, Wang Xinyu, Yao Qingshan, Zhao Xing.

Learning and Search:

1. Why all learning problems are inverse problems, requiring unbounded exhaustive searches, thus ill-posed?

学习问题对人类而言是先验约束下的归纳和推理（朴素贝叶斯），对机器而言是数据监督下的学习和优化（逻辑回归）。我们希望的是学到一个用数学框架表征的映射关系 f : X -> Y，搜索一个可泛化的计算模型，因此这是一个无界搜索过程，是不适定的反问题。

2. Why gradient is the key mathematical assumption that we could count on in order to search? What would be the general implications for such an assumption of continuality or locality?

a) 关键的数学假设：状态、空间的数学函数在搜索时光滑可导，如果不光滑，我们就不能应用微分，梯度。但是我们必须要用到梯度来调整参数，这也是唯一的办法。否则就不能进行搜索操作。

b) 意义：可以通过当前状态对未来的可能进行判断和预测。在搜索的时候使用梯度下降$\theta \leftarrow \theta + \alpha \frac{\partial L}{\partial \theta}$ 使用当前状态的梯度更新参数，按照一个确定的有效的准则进行搜索。

3. What is the generalizability of a mathematical process, from both expressive and inclusive point of views?

学习到的数学过程需要有一定的泛化能力，在两个方面：

Expressive，模型要有较强的表达能力，当模型输入有每一个较小的差异δ时，输出f(x)也应该存在差异。
$\forall \delta > 0, |x_1 - x_2| = \delta, \exists \gamma > 0, |f(x_1) - f(x_2)| > \gamma$

Inclusive，模型还要具有一定的包容性，模型要适用于一个较大的适用范围。数学形式：已知模型f:X->Y, |X|越大，模型的适用范围就越大，包容性也就越强。

4. What would be some of the solutions for such an ill-posed problem in order to yield at least some reasonable results?

适定问题有三个要求：解存在，解唯一，解连续依赖于初边值条件。否则就是不适定问题。

解决方案的核心思想就是获得高维复杂问题的低复杂度解，具体做法：

```
使用梯度系统求解局部性最优解
采用降维的方法得到系数的低秩解
采用正则化的方法对模型的结构进行约束和优化
将复杂不定函数分解为简单基函数的线性组合
采用深度学习的方法对问题进行结构化分解和逻辑重构
```

5. What are some of the mathematical hurdles that have prevented more generalizable solutions?

本质障碍具体表现在：

```
牛顿的正交、线性、均匀时空体系(维数灾难，搜索是无边界的)
高阶微分的局部解和奇异性(很难获得全局最优解)
多尺度问题(很难将实际问题投射到欧几里得空间中)
随机不确定问题
```

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. 为什么变量依赖(交互)会成为一个极其困难甚至不可能的问题?为这样一个奇点给出哲学、数学、物理、计算和数值的例子。

根据希尔伯特第十问题，不能通过有限步骤来判定不定方程是否存在有理整数解，且仅部分有解问题可以采用图灵机计算。对于一个可变依赖的不适定问题，首先很难判定其是否存在低复杂度解；其次，也很难确定是否存在一个具有泛化能力的图灵机可以解决该问题。

```
philosophical：真理不可证明性
mathematical：微分奇异性， 费马大定理
physical：三体问题
computational： 图灵序， 并行性
numerical：马尔科夫初值敏感(混沌现象)
```

7. Why a Euclidian-based measure would be most favored but usually impossible to obtain for a real world issue?

欧式问题中采用的是正交、线性、均匀的时空体系，即所有的测度和数值都是统一可比，具有物理含义的，因此更受欢迎。但是真实世界的问题并非都是欧式的，因为这些问题往往是复杂的、高纬度的、不成交的，很难找到一个统一的测度去度量，但这些问题可能存在着局部欧式性，所以可以对复杂问题进行分解。

8. What are some of the key requirements for a real issue to be formulated as a Euclidian problem?

要求采用正交、线性、均匀的时空体系，且所有的测度和数值都需要统一可比（最好也是统一的）

9. What would be the mathematical alternative frameworks to translate a non-Euclidian problem to mathematically appropriate solutions?

```
使用正则化对结构进行约束，将问题归一到同分布
使用PCA降维，将高纬度问题归一到低纬度正交的空间中，得到稀疏低秩解
使用kernel函数进行非线性变换，希望在升维得到的特征空间中变为欧式问题
使用概率图描述不同变量间的相互依赖关系，在此基础上进行取样和推理
使用深度学习，对原问题进行结构化分解和逻辑重构，将高阶复杂问题分解为低阶欧式问题的线性组合
```

10. Why in general the complex and high-dimensional data (the so-called big data problem, n<<p) from the same “class” tend to have a low dimensional representation?

```
人类的感知是有限的，无法感知很高维的问题。
许多特征是多余的，有些特征可以相互抵消。
只有一些特征是具有决定性的。
```

11. Why we would prefer a low complexity model for a high complex problem?

①模型复杂度超过一定阈值后，随着复杂度的上升，在训练集上的偏差会降低，但在测试机上的偏差会上升。之后的模型都会出现过拟合的现象，因此模型的复杂度一定要与数据复杂度保持统一，以此来使得模型具有较好的泛化能力。

②低复杂度模型更容易学习到主要的变化，模型具有更好的包容性。同时，由于一些问题是高维复杂的，所以我们希望找到这些复杂不定函数的简单基函数线性组合，即找到问题的低复杂度解。

12. What is the loss function? Give three examples (Least Square, Logistic, Hinge) and describe their shapes and behaviors.

Loss function是将一个或多个变量的事件或值映射到一个实数的函数。在机器学习中，损失函数是度量预测结果与真实结果之间偏差的函数。

Least Square Loss : $L\left(y_i, f\left(x_i\right)\right)=\frac{1}{2}\left(y_i-f\left(x_i\right)\right)^2$, 为凸函数和光滑函数。但是这个倾向于过度惩罚异常值，导致收敛速度相对于其他会比较慢。Behavior : 假设数据分布是高斯分布，最小化y和x的线性误差。

Logistics Loss : $L\left(y_i, f\left(x_i\right)\right)=-y_i \log f\left(x_i\right)-\left(1-y_i\right) \log \left(1-f\left(x_i\right)\right)$, 损失函数是连续的。可以采用梯度下降，收敛速度会稍快。Behavior : 假设数据分布是指数族分布。并且数据落在两个label的分布服从伯努利分布。更加robust，鲁邦。

Hinge Loss: $L\left(y_i, f\left(x_i\right)\right)=\max \left(0,1-y_i f\left(x_i\right)\right)$ 在SVM中用到，不光滑，不能采用梯度下降。Behavior : 函数只考虑点在两个boundaries中的，希望去分开它们。对于已经分开的，就不考虑。更加robust，鲁邦。

13. Using these losses to approach the linear boundary of a overlapping problem, inevitably some risks will be incurred; give two different approaches to remedy the risk using the SVM-based hinge loss as an example.

在SVM中，我们是想要最大化Support vectors中间的间隔。在这里给出SVM的简单证明，略。

Risk1 : 线性不可分。①使用kernel function, 将问题转向高维达到线性可分②使用PCA降维，降低维度后达到线性可分

Risk2: 过拟合。①使用PCA降维②使用softMargin允许某些样本不满足约束$\min_{\omega, b, \xi_i} \frac{1}{2}\|\omega\|^2+C \sum_{i=1}^N \xi_i$.

14. Describe biases and variance issue in learning, and how can we select and validate an appropriate model?

损失函数可以写成一下形式（h(x)为predict, y为truth）：

$$E_D\left[(y(x ; D)-h(x))^2\right]=\left(E_D[y(x ; D)]-h(x)\right)^2+E_D\left[(y(x, D)-E_D[y(x ; D)])^2\right]$$

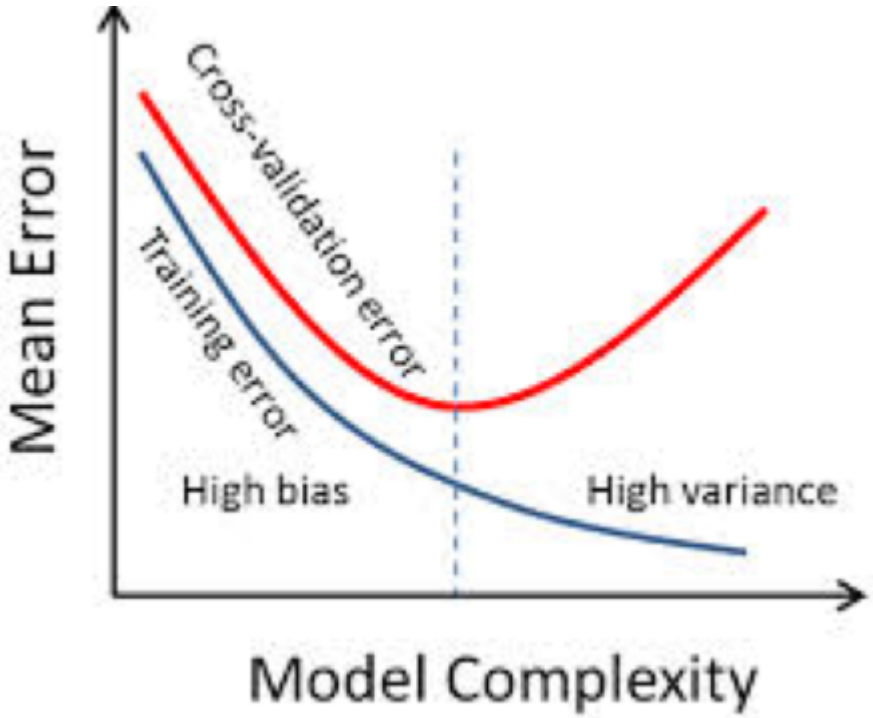
第一项就是bias偏差的平方，后一项就是variance方差。

bias：度量了学习算法的期望预测与真实结果的偏离程度（学习算法本身的拟合能力）

Variance：度量了同样大小的训练集的变动导致的学习性能的变化（刻画了数据扰动的影响）

在模型选择中：为了取得较好的泛化能力，需要使偏差较小能充分拟合数据，还要使方差较小，数据扰动产生的影响也要较小，防止过拟合。

k-fold交叉验证帮助选择模型：将数据集D划分为k个大小相似的互斥子集，每个子集Di都尽可能保持数据分布的一致性，然后每次用k-1个子集作为训练集，余下的那个子集作为测试集。进行k次训练和测试，最终返回k个测试结果的均值，做出“损失-模型复杂度”曲线。找到验证集的误差转折的点。是最理想的结果。



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?

加入正则项。一般有L1和L2 norm. $R(\theta)=\sum_{j=1}^{n}|\theta_j|$和$R(\theta)=\sum_{j=1}^{n}\theta_j^2$,同时使用λ 来控制惩罚程度

Linear regression: $L(\theta)=\frac{1}{2m}\left[\sum_{i=1}^{m}\left(f_\theta\left(x_i\right)-y_i\right)^2+\lambda R(\theta)\right]$.

Logistic regression: $L(\theta)=-\left[\frac{1}{m}\sum_{i=1}^{m}\left(y_i\log f_\theta\left(x_i\right)+\left(1-y_i\right)\log(1-f_\theta\left(x_i\right))\right)\right]+\frac{\lambda}{2m}R(\theta)$

不需要一个特定的低维模型。模型的复杂程度应该和给定的数据有关。需要根据训练集的n,p选取。如果降维太深会导致信息丢失。

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?

$y_i=\theta^T x_i+\epsilon_i$,其中$\epsilon$是随机噪声。假设误差项满足高斯分布$p\left(\epsilon_i\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$, 那么就可以转化为

$p\left(y_i|x_i;\theta\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\theta^T x_i\right)^2}{2\sigma^2}\right)$, $y_i|x_i;\theta\sim N\left(\theta^T x_i,\sigma^2\right)$

求对数似然函数得：

$$l(\theta)=\log\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\theta^T x_i\right)^2}{2\sigma^2}\right)$$
$$=\sum_{i=i}^{m}\log\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\theta^T x_i\right)^2}{2\sigma^2}\right)$$
$$=m\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(y_i-\theta^T x_i\right)^2$$

可以看到对数似然函数的最大项即为求第二项的最小值，正好就是最小二乘法。

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?

Least Square-based regression：基于最小二乘法的回归是从模型总体随机抽取n组样本观测后，最合理的参数估计量应该使得模型能最后拟合样本数据，也就是估计值和观测值之差的平方和最小。最小二乘法是从最小化损失 函数的角度建模。

Likelihood-based estimation：似然函数的思想就是什么样的参数才能使我们观测到目前这组数据的概率是最大的，需要对数据的分布有假设。似然估计法是从最大化似然函数的角度建模。

两者都是把估计问题转化成优化问题，用梯度系统的方法去搜索优化目标的最优解。当问题是凸的时候，搜索是按照梯度方向向最优解一步步渐进的，最终会找到全局最优的点；若不是凸优化问题，则会使梯度系统找到的解释局部最优解。

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?

①凸函数问题。只有是凸函数，梯度算法才可以保证找到最优解。但是对于很多高阶复杂问题，很难找到一个适合的凸函数作为优化目标。

②学习率α问题。学习率大，会在极值附近跳来跳去，影响收敛。学习率小，容易卡在局部极值，迭代速度慢等等。

③同步问题。如果有很多梯度下降的参数，如何combine在一起是一个问题。

19. What are the five key problems whenever we are talking about modeling (Existence, Uniqueness, Convexity, Complexity, Generalizability)? Why they are so important?

```
存在性：需要有解存在，只有存在，我们search才有意义。
唯一性：我们希望找到唯一解，不是唯一就会变得搜索很困难。
凸性：只有凸函数，才能保证找到全局最优。
复杂性：人工智能本质是找到高维复杂问题的低复杂度解。
泛化能力：机器学习的模型要对新样本具有一定的适应能力。没有泛化能力，学习就没意义
```

20. Give a probabilistic interpretation for logistic regression? How is it related to the MLE-based generative methods? 这是原理的证明题

a)概率学解释：逻辑回归是判别学习算法。我们希望模型能够预测出后验概率。在逻辑回归中，假设事件服从伯努利分布，令$f_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1+e^{-\theta^T x}}$, 有$P(y=1|x;\theta) = f_\theta(x)$, $P(y=0|x;\theta) = 1 - f_\theta(x)$，因此
$p(y|x;\theta) = (f_\theta(x))^y (1-f_\theta(x))^{1-y}$.

b)关系

$l(\theta) = \log L(\theta)$
$= \log \prod_{i=1}^{m} p\left(y_i|x_i;\theta\right)$
$= \sum_{i=1}^{m} y_i \log f_\theta\left(x_i\right) + \left(1-y_i\right) \log(1-f_\theta\left(x_i\right))$

通过证明可以得到对数似然方程是凹函数，在凹函数中，任何极大值就是最大值。在逻辑回归中优化的目标就是$min - l(\theta)$， 等价。

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?

Skip

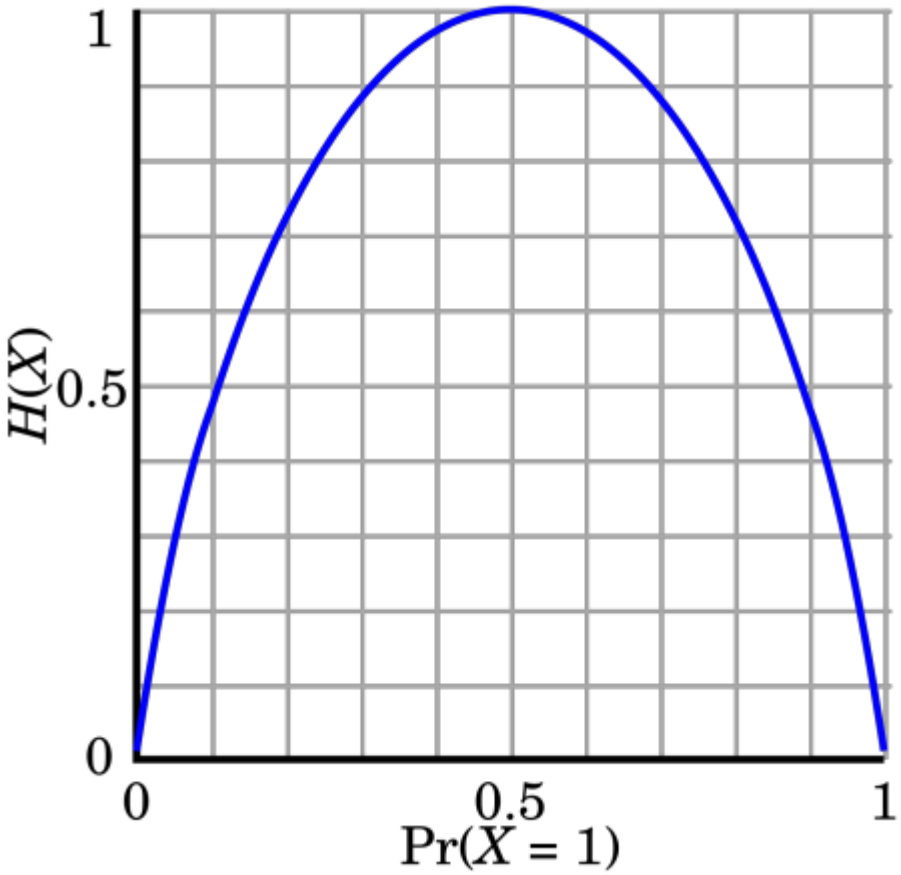
22. Can you provide a probabilistic comparison for liner and logistic regression?

线性回归处理的因变量是连续型随机变量，相反，逻辑回归处理的因变量是离散型二值的随机变量。
线性回归要求建立因变量和自变量之间的线性关系，而逻辑回归不要求因变量和自变量间存在这种线性关系
逻辑回归可以得到近似的后验概率预测，线性回归没有输出概率的能力

23. Why the log of Odd would be something related to entropy and effective information?

$P(X=1)=p$,$P(X=0)=1-p$,两者的比值被称为(对数几率)，$\frac{p}{1-p}$

在信息论中，二元信息熵函数被定义为$H(X)=H_b(p)=-p\log_2 p-(1-p)\log_2(1-p)$, 是消除不确定性所需信息量的度量。如下图，当p=0.5，信息熵最高，不确定性最大。



但我们对信息熵求导，发现$\frac{d}{dp}H_b(p)=-\log_2\left(\frac{p}{1-p}\right)$, 二元信息熵的导数是对数几率。所以对数几率是信息熵更本质的属性，指明了最小化信息熵的优化方向。

24. Why often we want to convert a liner to a logistics regression, conceptually and computationally?

conceptually：线性回归预测出的Y可能会超出(0~1)范围，逻辑回归预测出的Y都被归为0~1；逻辑回归相比较于线性回归有更好的鲁棒性，不要求因变量与自变量满足线性关系

computationally：逻辑回归更偏重于中间跳变部分的数据，而线性回归要考虑所有的数据点，因此计算的复杂度会降低。

25. Compare the generative and discriminative methods from a Bayesian point of view?

生成模型：学习得到联合概率分布$P(x,y)$，即特征x和标记y共同出现的概率，然后求条件概率分布P(x|y)。能够学习到数据生成的机制。

判别模型：学习得到后延概率$P(y|x)$，即在特征x出现的情况下标记y出现的概率。

对比：①生成方法尝试使用贝叶斯公式对联合概率分布进行建模，判别方法尝试对条件概率分布进行建模。②一般来说，生成方法计算复杂度更高。③然而，生成方法通常为我们提供了更多关于如何生成数据的洞察。④知道数据的大致分布情况，采用生成算法会得到更好的效果，因为在生成学习中用了分布的特征；对于多种分布或不确定分布，采用判别学习会得到更好的效果，因为判别学习算法更适应于一般情况。

26. What are the most important assumption for something Naïve but still very effective? For instance for classifying different documents?

朴素贝叶斯假设：对已知类别，假设所有属性相对独立；换言之，在给定y的条件下，x_i之间条件独立。基于以上假设，后验概率可重写为：$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P\left(x_i|c\right)$

可以用于文本分类任务中，具体做法如下：

假设字典的大小为|V|,我们可以构造一个特征向量，表示哪些词出现了。

当有足够多的训练样本后，由于同类样本满足同分布。计算先验概率$p\left(x^j = k|y = c\right) = \frac{\sum_{i=1}^{m} 1\left\{x_i^j = k \wedge y_i = c\right\}}{\sum_{i=1}^{m} 1\{y_i = c\}}$

遇到新的没有标记的样本x后，可以按照朴素贝叶斯公式$P(c|\mathbf{x}) = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P\left(x_i|c\right) = P(c) \prod_{i=1}^{d} P\left(x_i|c\right)$来对文本进行分类。

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?

太困难了

28. What are the key advantages of linear models? But why linear model tends not expressive?

Linear Model: ①数学上更简单，模型复杂度低，自变量和因变量是线性关系。②不会产生复杂数学问题，如不能求得全局最优解。③泛化能力好

线性模型的能力被局限在线性函数中，所以它无法理解两个输入变量间的相互作用。

29. What are the key problems with the complex Neural Network with complex integrations of non-linear model?

1. 函数不是凸函数，有很多局部解，可能找不到全局最优解。
2. 存在高阶组合的奇异性问题。
3. 很大的参数量

30. What are three alternatives to approach a constrained maximization problem?

约束问题

$$\begin{aligned} &\max f(x) \\ &\text{s. t. } g_i(x) = c_i \text{ for } i = 1, 2, \cdots, n \\ &h_j(x) \geq d_j \text{ for } j = 1, 2, \cdots, m \end{aligned}$$

1. 按照约束的f(x)作为优化目标，每次朝着可行的搜索方向(f(x)上升且不会越出可行域)进行迭代。
2. 将主优化问题转化为相应的对偶问题
3. 修改约束在进行求解，比如min变为max,去除整体的根号等等

31. What is the dual problem? What is strong duality?

$\max f(x)$
$\text{s. t. } h_i(x) = c_i \text{ for } i = 1, 2, \cdots, k$
$g_j(x) \leq 0_j \text{ for } j = 1, 2, \cdots, l$

拉格朗日算子：$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^{k} \alpha_i h_i(x) + \sum_{i=1}^{l} \beta_i g_i(x)$，定义$\theta(x) = \max_{\alpha,\beta:\alpha_i \geq 0} L(x, \alpha, \beta)$，可以得到$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$

所以主优化问题和优化问题$\min_x \theta(x) = \min_x \max_{\alpha,\beta:\alpha_i \geq 0} L(x, \alpha, \beta)$等价，这两个就是对偶问题。

当$p^* = d^* = L(x, \alpha, \beta^*)$ 就是强对偶

32. What are the KKT conditions? What is the key implication of them? Including the origin of SV.

KKT condition:

$$\begin{aligned} \frac{\partial}{\partial w_i} \mathcal{L}(w, \alpha, \beta) &= 0, \quad i = 1, \ldots, k \\ \frac{\partial}{\partial \beta_i} \mathcal{L}(w, \alpha, \beta) &= 0, \quad i = 1, \ldots, l \\ \alpha_i g_i(w) &= 0, \quad i = 1, \ldots, m \\ g_i(w) &\leq 0, \quad i = 1, \ldots, m \\ \alpha_i &\geq 0, \quad i = 1, \ldots, m \end{aligned}$$

Key implication：原优化问题等价于优化拉格朗日函数在KKT condition下。

在求解过程中，能发现满足KKT条件的数据点都位于最大间隔边界上，即支持向量。这也显示出支持向量机的一个重要性质：训练完成后，大部分的训练样本都不需保留，最终模型仅与支持向量有关。

33. What is the idea of soft margin SVM, how it is a nice example of regularization?

软间隔的核心思想：我们很难确定使得训练样本在样本空间或特征空间中线性可分的超平面是否是 过拟合的结果，因此我们要引入正则项来缓解该问题，即允许某些样本不满足支持向量机的约束条 件。于是，优化目标修改为：

$\min \frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{m} \xi_i$
$\text{s. } t \cdot y_i f(x_i) \geq 1 - \xi_i, i = 1, 2, \cdots, m$
$\xi_i \geq 0, i = 1, 2, \cdots, m$

C表示对错误的惩罚系数。参数ξ 放松了对准确划分的限制。

第一项表示结构风险，用于描述f的某些性质，第二项则是经验风险，用于描述模型与训练数据的契合程度。C用于调整二者的比重。所以软SVM就是一个带正则项的优化问题。

34. What is the idea of kernel? Why not much additional computational complexity?

a)当原始样本空间中不存在能正确划分两类样本的超平面，可将样本通过非线性变换从原始空间映射到一个更高维的特征空间中，使得样本在特征空间中可分。如果原始空间是有限维的，那么一定存在一个高维特征空间使样本可分

b)由于SVM只涉及到内积运算。所以我们假象$K(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >= \Phi(x_i)^T \Phi(x_j)$, 那么不用计算复杂的非线性变化，可以由这个函数K直接得到非线性变换的内积，不需要额外的计算。

35. What is the general idea behind the kernel? What key computation do we perform? Why is it so general in data modeling?

a)把样本空间中线性不可分的样本通过非线性核函数映射到高维特征空间，使得数据线性可分

b)在样本空间中计算核函数$K(x_i, x_j)$

c)如果数据分不开，可以用一个简单的测度的变换就可以将数据分开。因此，常常通过引入核函数来 将线性学习器拓展为非线性学习器

36. Why we often want to project a distance “measure” to a different space?

投射到高维：当数据在低维样本空间不可分的时候，通过映射到高维可以把非线性问题变为线性问题

投射到低维：当维度太高时会出现不正交的问题，通过一些降维的方法把原始数据映射到低维空间，变为欧式问题。

真实的数据一般是不能在欧式空间中表示出来的。因此，我们想要投射到真实的空间。

37. What a Turin machine can do? What some of the key computable problems a Turin machine can do?

38. What a Turin machine cannot do? What some of the key computable problems a Turin machine cannot do?

不能实现真正的并行。

39. Give your own perspectives on the Hilbert No.10 problem in the context of computational limit.

能否通过有限步骤来判定不定方程是否存在有理整数解?

40. Give your own perspectives on the Hilbert No.13 problem in the context of mathematical limit.

一般七次代数方程以二变量连续函数之组合求解的不可能性。

41. Discuss human intelligence vs. Turin machine as to whether they are mutually complementary, exclusive, or overlapping, or contained into each other one way or another.

没有answer

42. Explain Bayesian from a recursive point of view, to explain the evolution of human intelligence.

Recursion to update the prior and structure, search problem

43. What are computational evolutional basis of instinct, attention, inspiration, and imagination?

一切的基础都是prior.

44. Explain the core idea of machine learning to decompose a complex problem into an integration of individual binary problems, its mathematical and computational frame works.

二元过程的结构和逻辑重组

45. What are the limitation of Euclidian (Newtonian) basis, from the space, dimension, measure point of view?

都太理想了。

46. Why differentials of composite non-linear problems can be very complex and even singular?

47. What is the basis of Turin halting problem? And why temporal resolution (concurrency) is a key for logics and parallelism?

1.停机问题就是判断任意一个程序是否能在有限的时间之内结束运行的问题。

2.时间是本质，对于逻辑和并发来说，时间是一种隐藏的状态。

就是没有真正的并行。

48. What is mathematical heterogeneity and multiple scale?

heterogeneous 是homogeneous 的反义.

mathematical heterogeneous: we need to integrate independent problems. Multiple scale: Combine all the local scales and normalize.

数学异构:我们需要集成独立的问题。

多重尺度:将所有局部尺度结合起来进行归一化。

49. Explain convexity of composite functions? How to reduce local solutions at least in part?

复合函数单调。凸函数时单调的，所以局部最优就是全局最优。

1）使用relu作为激活函数

2) capsule : 划分为多个凸的区域。

50. Why local solution is the key difficulty in data mining, for more generalizable learning?

在采用梯度下降时会陷入局部最优解而需要花费很多功夫去拜托局部最优去寻找全局最优。

Probabilistic Graphical Model

1. Compare the graphical representation with feature vector-based and kernel-based representations.

图模型：是一个替代性的表述，没有对特征做任何的降维、合并或转换。强调的是特征之间的关系，这种关系不可分割，很复杂的（有逻辑上、因果上的关系等），不如直接用一种结构去表述。因此，在图模型问题中，最本质的问题是确定特征间的关系。

基于特征向量和核函数的表示：作相反的假设，认为特征之间互相的联系到最后都可以去掉，通过降维或者维数之间组合的关系，强化独立性，希望将它们彻底分开。那么如果特征不能分开，降维、核函数映射等的方法就会引起比较大的误差。特征向量是在一个正交归一的欧式空间表示的，用一些简单的、一般的数学方法进行测度。核函数的方法是在之前的测度无法完成的时候对之前的参数进行非线性修改达到可以分开的目的。

2. Explain why sometime a marginal distribution has to be computed in a graphical model?

上游的先验对下游的推断有影响，但上游的先验是一个混合物，例如
$P(x_i) = \sum_{x_{i-1}} \sum_{x_{i-2}} \cdots \sum_{x_1} P(x_1, x_2, \cdots, x_i)$, 假设上游有k个不同的状态，，即k个总和为1的边缘分布。对于下游来说，不知道上游有那些k对它有什么样的影响，因此要做一个边缘概率的估计。在图模型中，概率通过这样一个结构连接在一起（有先后，上下）每个上游消息传到下游都有一组可能性，它是上有概率对下游有影响的一个边缘概率分布。这个边缘概率分布是可以递归的，k的分布会发生变化；越到下游， k的分布越来越简单，越来越清晰。

3. Why class labels might be the key factor to determine if presumptively different data distributions can be indeed discriminated?

Assume labels are outstanding for each class: From a Bayesian point of view: label equals to constrain which is ideal condition or truth and it should be as universal as possible. If label is ambiguous, that means it will vary from each individual. This will make it difficult to learn.

假设每个类的标签都是突出的：

从贝叶斯的观点来看:标签等于约束，约束是理想条件还是真理，它应该尽可能的普遍。如果标签是含糊不清的，这意味着它将因人而异。这将使学习变得困难。

4. Why knowledge-based ontology (representation) be a possible solution for many prior-based inference problems?

(not for sure) P(x,y) = P(y|x)P(x)

知识本体是从理解的角度出发的一种普遍的本体。利用先验结构，可以保证数据的分布

5. Why a graphical model with latent variables can be a much harder problem?

我们需要最大化对数条件似然函数$\log p(x|\theta)$，分为以下两种情况。

1) 如果所有变量都是可以观测到的，即训练样本是完整的，那么对数似然可以分解为局部项之和
$l(\theta; D) = \log p(x, z|\theta) = \log p\left(z|\theta_z\right) + \log p\left(x|z, \theta_x\right)$,可以通过一般的监督学习方法来学习参数$\theta$.

2) 如果存在不可观测的隐变量，所有的参数通过边际化被耦合在一起。此时可以通过对隐变量计算 期望，来最大化已观测数据的对数"边际似然" $l(\theta; D) = \log \sum_z p(x, z|\theta) = \log \sum_z p\left(z|\theta_z\right) p\left(x|z, \theta_x\right)$,此时用EM算法来进行参数估计。

6. What is the key assumption for graphical model? Using HMM as an example, how much computational complexity has been reduced because of this assumption?

核心假设：马尔科夫假设，即某一时刻的状态只和前一时刻的状态有关，某一状态到另一状态不遂时间的推移而改变。

在HMM中，马尔科夫的具体表现形式为：

每一个可观测状态仅与当前隐状态有关。$P\left(o_t|i_t, o_{t-1}, i_{t-1}, \cdots, o_1, i_1\right) = P\left(o_t|i_t\right)$

每一个隐状态仅与前一个隐状态有关。$P\left(i_t|o_{t-1}, i_{t-1}, \cdots, o_1, i_1\right) = P\left(i_t|i_{t-1}\right)$.

假设马尔科夫链长度为T，共有k个状态，那么复杂度从$O(k^T)$降为$O(Tk^2)$.

7. Why does EM not guarantee a global solution? What is a simple proof for that?

EM算法的两个步骤是：

- **E步**：以当前参数θ^t推断隐变量分布$P(Z|X,\theta^t)$，并计算对数似然$LL(\theta|X,Z)$关于Z的期望

  $Q(\theta|\theta^t) = \mathbb{E}_{Z|X,\theta^t} LL(\theta|X,Z)$

- **M步**：寻求参数最大化期望似然，即

  $\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t)$

  即EM算法优化的目标是对数似然函数关于Z的期望$Q(\theta|\theta^t)$，这是一个非凸优化问题，因此很容易陷入局部最优解。当修改初始条件时，往往会得到不同的最终收敛结果（简单证明）

正式证明如下：

- 对数似然函数可以被拆解为

  $\log p(X|\theta) = \log p(X,Z|\theta) - \log p(Z|X,\theta)$

  $= \sum_Z p(Z|X,\theta^t) \log p(X,Z|\theta) - \sum_Z p(Z|X,\theta^t) \log p(Z|X,\theta)$

- 应用**Jesen's Inquality**和**条件信息熵**

  $\log p(X|\theta) \geq Q(\theta|\theta^t) + H(\theta|\theta^t)$

  当且仅当$\theta = \theta^t$时，上述等式成立，此时EM算法收敛

- 由上述推导可以看出，我们优化的对数似然函数期望$Q(\theta|\theta^t)$只是对数似然函数$\log p(X|\theta)$的一个下界，无法保证得到全局最优解

8. Why is K-mean only an approximate and local solution for clustering?

> 1．从n个数据对象任意选择k个对象作为初始聚类中心
> 2．根据每个样本x_j与各均值向量与各均值向量的距离，根据距离最近的均值向量确定x_j的簇标记，将样本x_j划入相应的簇．
> 3．重新计算每个簇的均值向量
> 4．若函数收敛，则算法终止；否则，返回步骤2

k-mean算法实际上是想最小化平方误差$E = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2$，找到它的最优解需要考察所有簇的划分，这是一个NP难的问题，因此我们采用贪心策略，通过迭代优化求近似解。kmeans 只使用了聚类中心的信息，隐含假设数据是高斯分布的，只能得到局部最优解，修改初值条件往往会导致不同的收敛结果。

k-means实际是EM算法，所以不是最优的。

9. How to interpret the HMM-based inference problem from a Bayesian perspective, using the forward/backward algorithm?

一个隐马尔可夫模型可以由以下三组参数定义：

- 状态转移概率：模型在各个状态间转换的概率，通常记为矩阵$A = [a_{ij}]_{N\times N}$

  $a_{ij} = P(y_{t+1} = s_j | y_t = s_i)$
- 输出观测概率：模型根据当前状态获得各个观测值的概率，通常记为矩阵$B = [b_{ij}]_{N\times M}$

  $b_{ij} = P(x_t = o_j | y_t = s_i)$
- 初始状态概率：模型在初始时刻各状态出现的概率，通常记为$\pi = (\pi_1, \pi_2, \cdots, \pi_N)$

  $\pi_i = P(y_1 = s_i), \quad 1 \le i \le N$

1. 在前向算法中，给定HMM模型$\theta = [A, B, \pi]$，定义到时刻t部分观测序列为$o_1, o_2, \cdots, o_t$且状态为i的概率为前向概率$\alpha_i(t)$，本质上是一个似然概率，记作：

   $\alpha_i(t) = P(O_1^t, q_t = j|\theta) = \sum_{\mathbf{q}} p(O_1^t, \mathbf{q}, q_0 = 1, q_{T+1} = N|\theta)$

   有了这个定义，我们可以递推地求得前向概率

   $\alpha_j(t) = b_j(o_t) \sum_{i=1}^{N-1} a_{ij}\alpha_i(t-1)$
2. 在后向算法中，给定HMM模型$\theta = [A, B, \pi]$，定义在时刻t状态为i的前提下，从$t+1$到T的部分观测序列为$o_{t+1}, o_{t+2}, \cdots, o_T$的概率为后向概率$\beta_t(i)$，本质上也是一个似然概率，记作：

   $\beta_i(t) = P(O_{t+1}^T, q_t = j|\theta) = \sum_{\mathbf{q}} p(O_{t+1}^T, \mathbf{q}, q_0 = 1, q_{T+1} = N|\theta)$

   同样的，我们也可以递推地求得后向概率

   $\beta_j(t) = \sum_{i=1}^{N-1} b_i(o_{t+1})a_{ji}\beta_i(t+1)$

结合上述推导过程，可以通过前向算法和后向算法推断观测序列的概率：

$p(O|\theta) = \alpha_N(T+1) = \beta_1(0) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t)$

10. Show how to estimate a given hidden state for a given series of observations using the alpha and beta factors.

参考上题。给定观测序列O和隐藏状态i，在时间t处于该状态的概率为：
$\gamma_i(t) = p\left(q_t = i|O, \theta\right) = \frac{P(q_t=i, O|\theta)}{P(O|\theta)} = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N}\alpha_j(t)\beta_j(t)}$，那么在观测O下状态i出现的期望为：$\sum_{t=1}^{T}\gamma_i(t)$. 随着地推的进行，先验逐渐清晰，此时后验的影响很小。因此使用viterbi算法估计隐藏状态序列时，仅使用了前向过程的信息。

11. How a faster inference process would be constructed, given a converging network?

对一个收敛的HMM模型$\theta = [A, B, \pi]$，给定观测序列$O = [o_1, o_2, \cdots, o_T]$，可以使用Viterbi算法来快速推断出隐藏的状态。我们想要递推的找到：$\hat{\mathbf{q}}_1^T = \arg\max_{\mathbf{q}_1^T} p\left(O_1^T, \mathbf{q}_1^T, q_0 = 1, q_{T+1} = N|\theta\right)$. 我们定义
$\phi_j(t) = \max_{\mathbf{q}_1^t} p\left(O_1^t, \mathbf{q}_1^{t-1}, q_0 = 1, q_t = j|\theta\right)$ 表示时间t处于状态j时观测结果最优可能对应的状态序列的概率。然后按照下列更新：
$$\begin{aligned}\phi_j(1) &= b_j\left(o_1\right)\cdot\pi_j \\ \phi_j(t) &= b_j\left(o_t\right)\cdot\max_{i\in S}\left(a_{ij}\cdot\phi_i(t-1)\right)\end{aligned}$$

然后最优的隐藏状态序列可以通过$q_j^{\max}(t) = \arg\max_{1<i<N} a_{ij}\phi_i(t-1), \hat{q}_{t-1} = q_{\hat{q}_t}^{\max}(t)$得到。

12. How can an important node (where inference can be significantly and sensitively affected) be detected using an alpha and a beta process?

参考9. 可以通过公式$p\left(q_t=i|O,\theta\right)=\frac{P(q_i=i,O|\theta)}{P(O|\theta)}=\frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N}\alpha_j(t)\beta_j(t)}$, 计算时间t处于状态i的概率。记$p_t^{\max}=\max_{1\leq 1\leq N}p\left(q_t=i|O,\theta\right)$. 表示时间t处于某一状态的最大概率。如果$p_t^{max}$与$p_{t-1}^{max}$相比发生了跳变，则可以认为隐藏状态qt节点很重要，因为提供了一个很强的先验，约束了后续的搜索范围。

13. Why often an alpha process (forward) is more important than beta(backward)?

因为前向算法的过程是一个推理、判断和排除的过程。通过不断的迭代，隐马尔科夫链中的显眼关系越来越清晰，先验概率的比重逐渐变大，可以据此缩减穷举的范围。

14. What are the key differences between an alpha and a beta process from human and machine intelligence point of views?

前向算法：与人类智慧类似，是一个先验约束下的搜索过程，通过迭代不断优化先验概率 后向算法：与机器智能类似，是一个通过穷举学习后验的过程，通过迭代不断优化似然概率

15. How data would contribute to the resolution of an inference process from a structural point of view?

HMM模型中不是所有的结构都重要，需要看有哪些节点重要或者有哪一片节点重要。数据可以帮助我们确定上下文结构关系，去除一些不重要的依赖关系，来缩减搜索的范围，从而得到一个最佳的稀疏解。

16. For a Gaussian graphical model, what is the implication of sparsity for such a graphical model? How is such sparsity achieved computationally?

a) 参数之间的相关性应尽可能被定义在较少的维数里，抓住主要的参数即可，否则任何两两之间的参数都可能有相关性 b) 稀疏性强调要抓住有主要关系的节点，同时有一些边需要根据位置来评价。可以引入Frobenius范数来进行正则化，其不仅约束了边的值，还约束了图的秩，可以得到比较理想的稀疏解

17. Explain the objective function of Gaussian graphical model? Thus the meaning of MLE?

minimizes the summation of all the covariance matrix. inverse covariance matrix

18. How a row or colum-based norm (L1 or L2) can be used to construct hub-based models? And why this might be highly applicable?

Every single node supposed to be connected to many other nodes. The rows is either big or small, and when it's small, it is zero. Row or column based norm encourages sparsity. So we can have everything row based. The rows will give us a different network called hub based network, which captures the major connectivity of the feature network. The hub is important from the stability, control and dynamics point of view.

19. How Gaussian graphical model be used to model a temporally progressive problem? Why still a simple 2d matrix compression problem can still be very demanding computationally?

Pass，听不懂

20. Why a spectral or nuclear norm would be better than Frobenius norm to obtain a sparse matrix?

Spectral 范数，即矩阵最大的奇异值。 Nuclear 范数， 即矩阵奇异值的和。 Frobenius 范数，即矩阵元素绝对值的平方和再开平方。 spectral 范数与 nuclear 范数都是可以用于约束低秩矩阵的正则项。在协方差矩阵中，奇异值就是特征值（yb自己说的）。特征值描述了矩阵稀疏性，一定程度上实现了对gaussian graphical model的低秩近似

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.

- PCA
  1. 对所有样本进行中心化：$x_i \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$
  2. 计算样本的协方差矩阵XX^T
  3. 对协方差矩阵XX^T做特征值分解
  4. 取最大的d'个特征值所对应的特征向量构成投影矩阵$W = (w_1, \cdots, w_{d'})$
  5. 通过$W^T x_i$将样本映射到低维空间
- SVD

  假设矩阵A是一个$m \times n$的矩阵，那么我们定义矩阵A的SVD为

  $A_{m\times n} = U_{m\times m} \sum_{m\times n} V^T_{n\times n}$

  其中$\sum$除了主对角线上元素以外全为0，主对角线上的每个元素都称为奇异值；U和V满足$U^TU = I, V^TV = I$
- SVD应用于PCA

  在奇异值矩阵$\sum$中，奇异值是σ_i按照从大到小的顺序排列的，设矩阵XX^T的特征值λ_i，存在关系$\sigma_i = \sqrt{\lambda_i}$，可以根据奇异值来选出最大的$d$个特征值；同时，右奇异矩阵$V$就是$XX^T$的特征向量张成的。

  所以我们可以通过SVD得到协方差矩阵XX^T最大的d个特征向量张成的矩阵，但是有一些SVD实现算法可以不必求出协方差矩阵XX^T，也能求出右奇异矩阵V。也就是说，PCA算法可以不用做特征值分解，这在样本量很大的时候能大大减少运算量

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

Regular PCA: 做完SVD之后直接做硬切割，选出最大的d个特征值对应的特征向量，把方差变化比较大的部分流了下来，直接丢弃了所有的可能性。

Low-ranked PCA:对秩进行了压缩，做玩SVD后对奇异值的进行了约束。弱的可变性没有被丢弃，而是通过一定的变换压缩到了低秩的空间中。依据的原理是，我们可以用最大的 个奇异值和对应的左右奇异向量来近似描述矩阵，即: $A_{m\times n} = U_{m\times m} \sum_{m\times n} V^T_{n\times n} \approx U_{m\times k} \sum_{k\times k} V^T_{k\times n}$ 然后通过$V^T x_i$ 可以将样本映射到k维空间。

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

矩阵可以被看作一种线性变换，而这种线性变换效果与基的选择有关。一个线性变换的作用可以包含旋转、缩放和投影三种类型的效应。奇异值分解正是对线性变换这三种效应的一个析构，而特征值分解其实是对旋转缩放两种效应的归并。特征值分解和奇异值分解都是给一个矩阵(线性变换)找一组特殊的基，特征值分解找到了特征向量这组基，在这组基下该线性变换只有缩放效果。而奇异值分解则是找到另一组基，这组基下线性变换的旋转、缩放、投影三种功能独立地展示出来了。 SVD可以将一个比较复杂的矩阵用更小更简单的几个子矩阵的相乘来表示，这些小矩阵描述的是矩阵的重要的特性。得到的对角矩阵对角线上的值就是奇异值，如果奇异值越大，则表明左奇异向量中的对应行和右奇异向量中对应列的重要性越高。

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

深度学习认为问题本身是由每部分之间结构的组合，所以在分解和组合的过程中学习的是各部分结构以及它们之间的关系。深度学习首先将特征分解，然后学习如何重构特征，是一个从底层，到局部，到整体的一个过程，模仿的是人类对知识的理解.

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

Function	ADVANTAGE	DISADVANTAGE
Sigmoid	smooth and expressiveness	gradient vanish; lower convergence speed
Relu	encourages sparsity; higher convergence speed; less likelihood of gradients vanish	less expressiveness

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

矩阵分解的目的：矩阵分解就是将一个高维矩阵降解成几个低维矩阵来简化数据的表示。在我们的日常生活中，接触到的很多数据，都是存在线性相关的，很高的维度也不利于数据处理和分类计算。矩阵分解将高维高稀疏的表示向量转化为低维相对不那么稀疏的矩阵，进而方便后面的相似计算。 一个高维度矩阵问题，我们将其分解为多个相对低维度的矩阵，复杂度会降低，一个整体被分为了几个小部分。这样还能控制矩阵的稀疏度。这样还能降低网络层之间的关联度，使网络尽可能稀疏。 从复杂度来讲，分解实际上是有帮助的，因为分解后的组合复杂度会降低，但有些人也可能会有更多的开销。 但是降维并不是一定好的，有的时候会损失部分信息，让模型的表现不那么好

7. Discuss convexity of composite functions, what mathematical strategies might be used to at least reduce local solutions.

在复合函数中，都是线性函数相乘组合在一起的，不断的乘，所以这个函数是非常复杂的，一般也不是凸函数，所以复合函数拥有许多的局部解。

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?

一般在input layers使用L2 norm， 在hidden layer中使用L1 norm，在output layer中使用sigmoid. 在图中也可以看出，L2 norm更容易获得平滑的weight。L2norm是向量的各元素的平方和再求平方根，使这个值变小，可以使每个weight都接近0，但又不等于0。这样就可以平滑各个feature的weight，防止出现很大的数字。这样就能让输入的数据变得平滑，相对差异变小。所以input layer一般使用L2 norm，这样我们就能进行一个相对于各个feature更平衡的学习，否则就会因为数字不同有一个偏差。 在hidden layer中一般使用L1 norm，因为我们在实际学习中，更希望获得一个尽可能稀疏的网络，L1相比较而言更aggressive。在两层相邻的layer中，只有重要的connection 才能得到保存。其他的都被置0。所以在hidden layer更倾向于L1 norm，在最小化经验误差的情况下，可以让我们选择解更简单（趋向于0）的解，降低模型的复杂度。 sigmoid是二分类问题上的probability，在output中 layer中，可以直接输出分类的概率，这UI与很多需要利用概率辅助决策的任务非常有用。所以要用sigmoid。另外一个是因为对数几率函数是任意阶可导的凸函数，有着很好的数学性质，很多数值优化算法都可以直接用于求取最优解，求导容易。

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.

a)真正的特征是特征经过分解后的一系列二分元问题和它们的重新组合（结构）

做脸的位置检测时，一个脸可能涵盖很多像素点，可以用一个几万维的向量表示。但是，也可以 把这个向量降维成三维，表示脸的空间位置信息，这样更能表示问题的本质，可能比其他的结构 效果好 考虑一个编码问题，假如有 个命题变元，那么一共可以有 种取值结果，可以采用一个8 维的one-hot向量来代表每种取值。但是，也可以用一个3维的向量来表示取值方案，命题变元为 真时对应位取1，否则取0。这样进行预测时，可以对作用sigmoid函数，对每

一维取值单独进行 预测，大大减少模型的复杂度

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?

这些特征是人为提取出的，特征分解是将这个对象的特征分解到最低层，在一层层的学习中学习这些基本单元的组合方式（局部，整体）和分布，最终学习到的是这个对象的基本构成和关系，这往往是最本质的问题。并且在学习时不是对某一个特征进行学习，而是在多层的结构中学习组合关系，从统计学上说比学习单独的特征要强得多，而且降低了噪音，更有鲁棒性。

1. 高阶非线性导数相互作用过于复杂，因此深度学习将问题简化为线性组合关系
2. 由于问题非凸，且参数间有很强的耦合性，会存在很多的局部解
3. 基于马尔可夫假设，只考虑了同层的线性组合关系，对于不同层的相互关系没有考虑

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

模型的复杂性和样本的复杂性要匹配，模型越复杂所需要的样本数量就越多，在匹配的情况下模型的复杂性要尽可能低。选特征时的维数和模型的维数及相互之间结构上的关系都应该匹配，否则无法很好地刻画特征与特征之间的关系，导致实验结果和实际结果误差很大。具体来说：

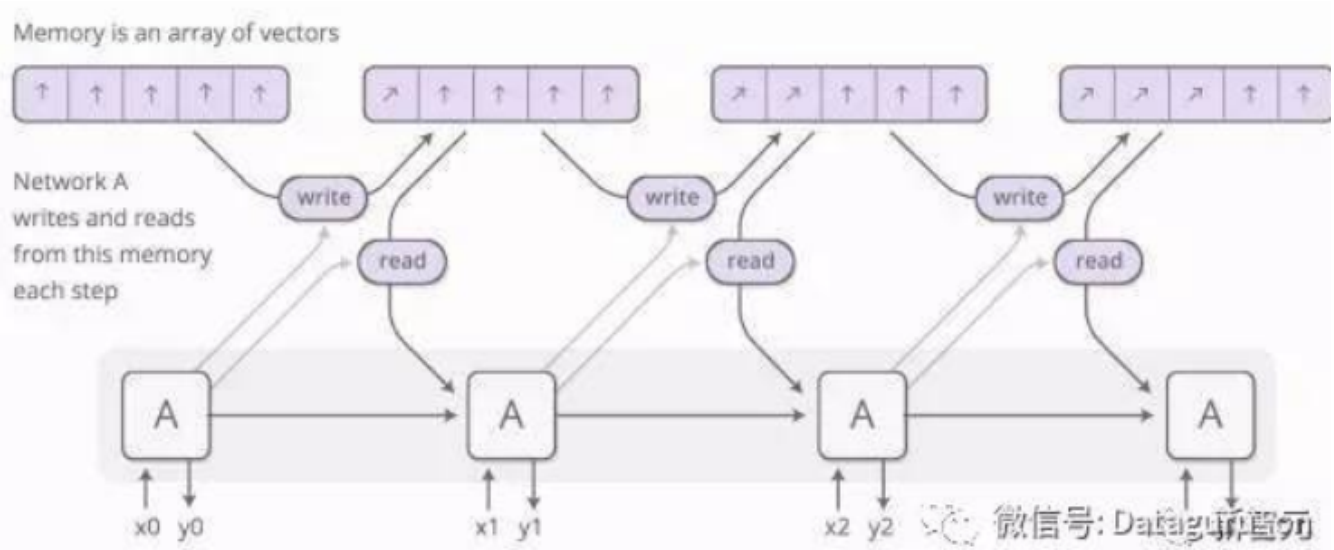
结构上：特征的维数和特征之间的关系需要和选择的模型匹配。 统计量上：如果特征的维数较大，我们需要的样本容量也需要大。

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

存在这样一种特殊的结构，上面的一些节点表示问题的组合、逻辑关系、推理关系等。对于高维复杂问题，我们试图最小化这个结构框架，即得到问题最关键的独立的元素及其相互关系（逻辑，因果等），这样可以在最低的纬度抓住问题本质的机制及机制的相互作用。在包含大部分信息的情况下进行降维。

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

NTM(Netural Turing Machine), 在一个很高的层面上构建神经计算模型，作为图灵机的实现。核心思想就是在RNNs的基础上加上augment记忆模块。如图：



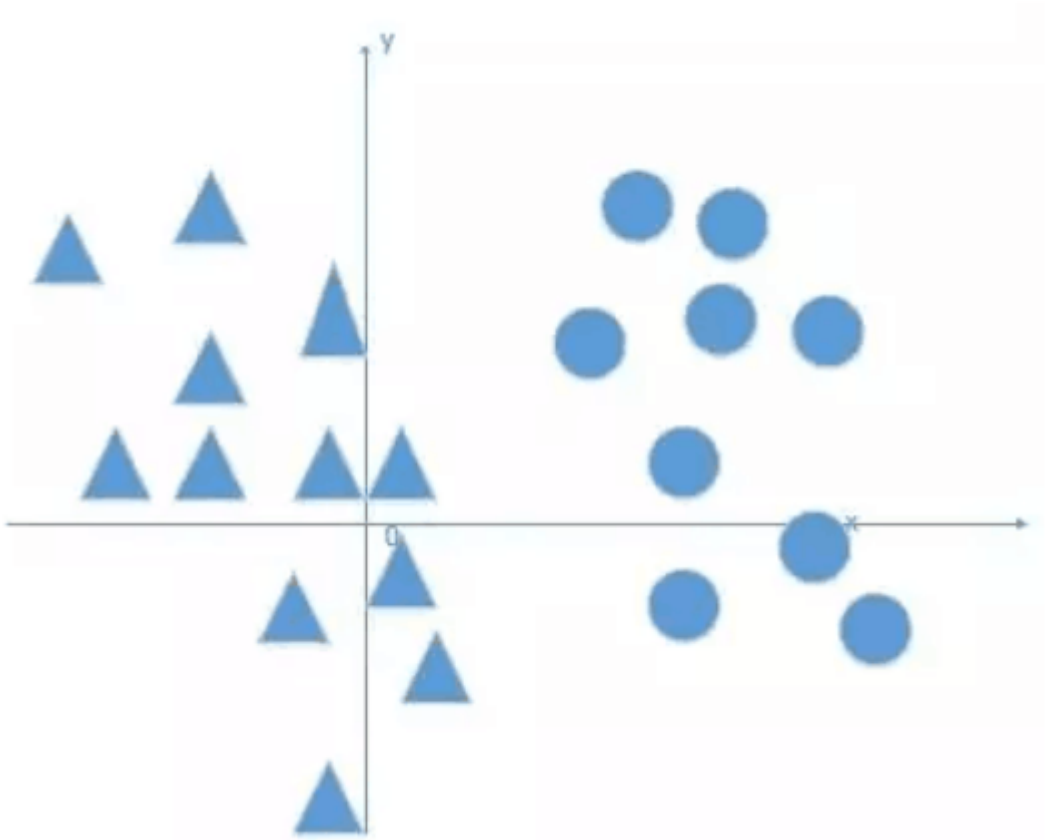
不断的读入记忆模块中的数据，处理后，在写到模块中，再读入。关键是引入“写入注意力分布”。这些模型可以执行的许多任务 - 例如学习数的加法 - 客观上并不是很困难。 RNN能做CNN不能做：RNN可以有用于描述时间上连续状态的输出（样本出现的时间顺序对于自然语言处理、语音识别、手写体识别等应用非常重要。），有记忆功能，并且RNN中，神经元的输出可以在下一个时间段直接作用到自身，但CNN就只能用于静态输出。

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

Pass

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

加入激活函数，输出就变成了g(wx+b)，在这个式子中，w是表示分类的线的方向绝对分割平面的方向所在。而b则表示，竖直平面沿着垂直于直线方向移动的距离。 如果不加偏执，那么分割的线就只能过原点。对于下图这种就无法进行分类。只有加了偏置，这个式子的表示能力才能拓展到任意一个分割平面。做到更好的分类效果



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

Pass

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

Pass，强化学习，不要求

Value based的方法主要是求 V(s) 和 Q(s,a)。计算的是各个state或者action的值。在求得这些值以后，在计算策略的时候，使用的 a = argmax Q(s,a)，即对于当前状态s, 遍历其所有的Q(s,a)获得当前状态下应该采用的a值。

1. Value based 方法只适合离散的action情况，因为要取argmax_a Q(s,a)

2. Value based 方法只适合 deterministic policy的情况，policy based 方法适合 stochastic policy 也适合于 deterministic policy. 另外对于value based 的方法，针对action增加 exploration 的效果是采用的 e greedy 方法来做， 如下面这段代码。而policy based 方法，本身就带有随机探索的功能。

Policy based 方法是直接 maximize the expected return. Policy-based methods find the optimal policy by directly optimizing for the long term reward. How you choose to optimize the policy parameters then becomes the different branches within policy based methods

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

Pass，强化学习，不要求

19. Explain an expectation-based objective function?

Pass，强化学习，不要求

20. Explain Haykin's universal approximation theorem.

Pass，强化学习，不要求

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?

①数据越多越好，模型越简单越好。 ②最终目标是减小泛化误差(generalization error)，即算法在未知数据集上的表现。复杂的 模型可能在训练集上表现很好，但存在high variance的问题，面对全新的测试集，容易出 现过拟合现象。 ③当训练数据量较少，而所选模型比数据本身复杂时，会造成指数级误差。 ④在模型选择时，可以采用交叉验证CV, AIC, BIC等方法

来自课堂的补充内容： 我们需要考虑的关键问题： 模型的复杂度和数据的复杂度需要匹配。 并且模型复杂度应尽可能降低来提高模型的泛化程度，数据的量要充足

数据侧的一些工作总称为特征工程，主要包括数据预处理（无量纲化，二值化，哑编码，缺失值计算，数据变换），特征选择，数据降维（PCA，LDA）模型侧的工作目的是在同样的数据上，选取更好的模型。更好指的是模型泛化性（generalizability）更好，或者说模型鲁棒性（robustness）更好 。

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?

分类问题中，我们试图将标签为0和1的数据点进行划分（partition）；聚类问题种，我们试图从不同类别的边界性（marginal）角度解决分类问题；HMM中，我们用条件概率下的后验（posterior）对数据作出预测；graphical modeling中我们使概率推断复杂化，用结构的（structural）手段试图改变概率中的条件项。

从上面的描述中我们不难看出，为了使模型更具表达性，我们逐渐从划分的角度，到边界的角度，最终到推断（inference）的角度来解决建模的问题。模型表达性的增强伴随着模型复杂度的提高，和模型逻辑性的增强，一定程度上，这还帮助我们提高了模型的可解释性。

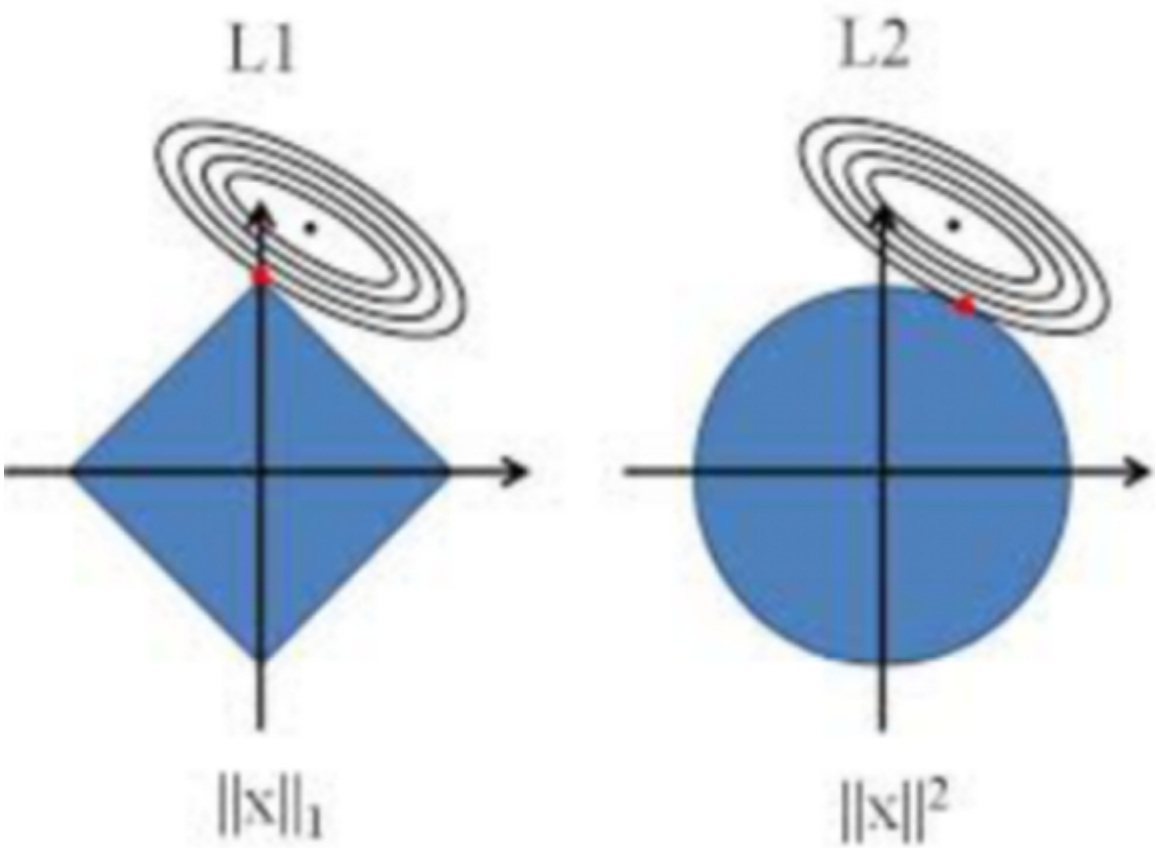
3. What are the potential risks we could take when trying to perform a logistic regression for classification using a sparsity-based regularization?

在逻辑回归时，我们倾向于认为一部分特征是冗余的，因此使用sparsity-based正则化项，以实现降低维度（dimension reduction）的效果。在引入正则项时，如果我们使用L-0 norm，将会导致优化过程成为NP-hard问题，因此我们用L-1 norm替代L-0 norm，既保证算法复杂度合理，也近似了L-0 norm降低维度的作用。但由于L-1 norm并不是真正表达稀疏性的范数（L-0 norm），使用它将可能导致模型中一些重要的特征被错误地去除，甚至进而可能导致模型欠拟合（under-fitting）

4. Give five different structural constrains for optimization with their corresponding scalars.

依据老师的是以下 L-1 norm, L-2 norm, Frobenius范数，核范数，谱范数。（其实有很多种范数）

此外，老师提示我们应该在回答中画下图。并对下面两种范数是否稀疏做出说明。



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.

- 局部性 (Locality)　　梯度系统 (Gradient System)
- 凸问题 (Convexity)　　全局解 (Global Solution)
- 线性组合 (Linearity)　　相互作用 (Interaction)
- 稀疏低秩解 (Sparsity)　　降维 (Dimension Reduction)
- 正则化 (Regularization)　　结构约束 (Structural Prior)
- 二元问题 (Binary)　　结构化分解和逻辑重构 (Structural Decomposition & Logics Reconstruction)
- 贝叶斯 (Bayesian)　　有约束的搜索 (Constrained Search)
- 最大期望 (Expectation)　　局部梯度 (Local Solution)
- 马尔科夫假设 (Markov)　　递归函数 (Recursion)
- 点积与测度 (Inner Product)　　非欧氏问题 (Non-Euclidian Measure)
- 概率图模型 (Graphical Model)　　取样和推理 (Sampling)

图片取自ppt第一章。

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

因为数据的表征能够为好的建模提供以下几点帮助：好的维度，好的算法，好的结构。

在上面几点的基础上，我们就能使数据发挥最大的作用。因此，数据表征与实际建模同样重要。

7. How does the multiple-layer structure (deep learning) become attractive again?

深度学习的一个核心观点是将一个复杂的非线性过程解构（decompose）成线性的组合。正是这种解构的策略，使得深度学习受到了欢迎。个人认为应该还涉及到计算力的提升。

8. Discuss Turin Completeness and the limit of data mining

Pass。不要求。

9. Discuss general difæculties of using gradient for composite functions or processes.

复杂函数的梯度存在大量局部极小点（local minimum），和奇异点（singularity），因此梯度算法通常很难找到全局最优解。

10. What is the trend of machine learning for the next 5-10 years?

应该包含 机器学习的未来--融合人与机器智能