Review for IEEE Data Mining Class, 2019 Spring (Part One, 50%, A Total of 140 Questions)

Learning and Search:

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

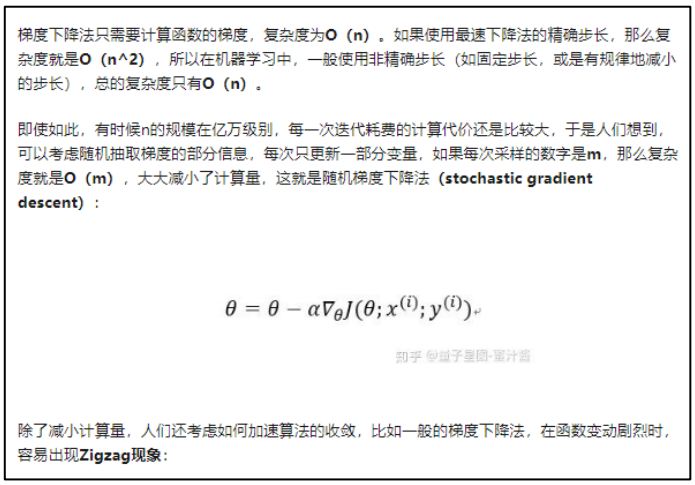
[1] The main goal of learning from examples is to infer an estimator, given a finite sample of data drawn according to a fixed but unknown probabilistic input-output relation. The desired property of the selected estimator is to perform well on new data, i.e. it should generalize. The key to obtain a meaningful solution to the above problem is to control the complexity of the solution space. In such a context to avoid undesired oscillating behavior of the solution we have to restrict the solution space. (ref: <Learning, Regularization and Ill-Posed Inverse>)

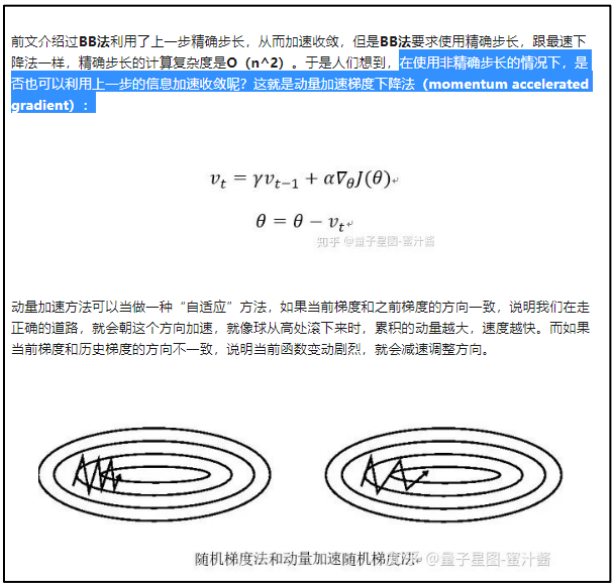
[2] A careful analysis shows that a rigorous mathematical connection between learning theory and the theory of ill-posed inverse problems is not straightforward since the settings underlying the two theories are different. In fact, learning theory is intrinsically probabilistic whereas the theory of inverse problem is mostly deterministic. (ref: < Learning from Examples as an Inverse Problem >)

2) Why gradient is the key mathematical assumption that we could count on in order to search? What would be the general implications for the existence of at least some continuality or locality?

[1] 无论是一般的梯度下降法，还是最速下降法，牛顿法，随机梯度下降法，动量加速梯度下降法，都与梯度相关。(ref: https://blog.csdn.net/Timingspace/article/details/50963564)

[2] 函数在某一点的梯度是这样一个向量，它的方向与取得最大方向导数的方向一致，而它的模为方向导数的最大值。 这里注意三点：1）梯度是一个向量，即有方向有大小；2）梯度的方向是最大方向导数的方向；3）梯度的值是最大方向导数的值。 既然在变量空间的某一点处，函数沿梯度方向具有最大的变化率，那么在优化目标函数的时候，自然是沿着负梯度方向去减小函数值，以此达到我们的优化目标。





3) What is the generalizability of a mathematical process, from both expressive (smoothness) and inclusive (capacity) point of views?

E-way: encode/decode many inputs and outputs readily.
C-way: not only given data, but also all possible x&y.

[1] 所有的变化和差异都可以响应出来，线性、光滑的范围; $x \xrightarrow{f(x)} y, \Delta x \xrightarrow{f(x)} \Delta y$

[2] 较大适用的范围和空间，所有的 x 都有对应的 y; enable to encode as many inputs as possible.

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

cross validation: find a model that generalizes well to testing data
soft margin: overlapping class
kernel method, regularization

reasonable results?

[1] 如图所示

To Obtain Low Complexity Solutions

• 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)

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] 非线性微分方程的奇异解

[3] 局部解问题

[4] 计算复杂度

[5] curse of dimensionality：在样本量一定的情况下，维度越高，样本在空间中的分布越呈现稀疏性。随着维度的增加，覆盖度指数级下降。这种分布的稀疏性带来 2 个不好的影响：

i. 模型参数难以正确估计（例如：样本不够时，得出的决策边界往往是过拟合的）。随着维度的增加，理论上需要指数增长的样本数量覆盖到整个样本空间上时，才能保证模型能有效的估计参数。而对于那些具有非线性决策边界的分类器（如：神经网络，决策树，KNN）来说，如果样本不够多时，往往很容易过拟合。

ii. 样本分布位于中心附近的概率，随着维度的增加，越来越低；而样本处在边缘的概率，则越来越高。通常来讲，样本的特征位于边缘时，比位于中心区域附近更难分类，因为边缘样本的特征取值范围变化太大。想象二维空间下的两个同心圆，假设 r1=0.5，r2=1，那么面积之比为 1/4；如果半径不变，在三维空间中，体积之比变成 1/8；到了 8 维空间下，超球体的体积之比为 1/256，仅仅占到 2%。当维数趋于无穷时，位于中心附近的概率趋于 0。这种情况下，一些度量相异性的距离指标（如：欧式距离）效果会大大折扣，从而导致一些基于这些指标的分类器在高维度的时候表现不好。

[6] 总之，no mathematical methods guarantee to obtain the best performance

6) Why variable dependences (interactions) could become an extremely difficult and even an impossible problem? Give philosophical, mathematical, physical, computational, and numerical examples for such a singularity. 真理不可证明　费马大定理　三体　停机问题　初值敏感

[1] interaction 的定义：In statistics, an interaction may arise when considering the relationship among three or more variables, and describes a situation in which the effect of one causal variable on an outcome depends on the state of a second causal variable (that is, when effects of the two causes are not additive). Although commonly thought of in terms of causal relationships, the concept of an interaction can also describe non-causal associations. Interactions are often considered in the context of regression analyses or factorial experiments.

[2] 真理不可证明

i. 真理是人们对于客观事物及其规律的正确反映。

ii. 哥德尔不完备定理鼎鼎大名：存在一个真命题，但是我们没法证明。

[3] 微分的奇异性、费马大定理

i. 费马大定理内容：整数 n >2 时，关于 x, y, z 的方程 x^n + y^n = z^n 没有正整数解。

[4] 三体问题

i. 牛顿三体运动没有解析解的原因是因为任何微小的扰动都会导致在一段时间后都会有极大的误差，而观测的精度总是有限的。能计算出来解，但是结果并不具有周期性而且趋向于混沌。

[5] 图灵机，并行（没有真正的并行）：

i. 祖--冯诺依曼计算机就是个串行机，冯氏机器的理论基础是图灵机，图灵机本质上也是个串行机，图灵机的精髓在于计算的过程而不是计算的本身。

ii. 如果说人们一直在努力实现并行计算， 最起码在基于图灵机的机器上，这种努力是徒劳的，那些所谓的锁机制，缓存一致性等等都只是治标不治本的方法。如果想实现真正的并行计算，就必须将计算本身而不是计算的过程作为一个重点

iii. 按照 lambda 演算方式设计出来的计算机是可以实现并行计算的，这种机器的计算粒度非常小，而不像在冯诺依曼机器上执行的最小粒度必须是线程，线程代表了一个图灵过程，这是图灵机必须的，并行计算的第一步必须为计算实体分配线程，可见这种方式及其拙劣，lambda 演算的机器就不必有线程的概念，因为它代表的是计算本身，而不是过程。

features are heterogeneous and arbitrary; not independent and not measurable. we can use methods to make features measurable for 欧式空间 but not general enough

[6] 初值敏感：许多非线性系统在一定条件下都具有初始条件敏感性，它是非线性动力系统的普遍行为，是确定性系统内在随机性的反映。

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

[1] 欧氏距离定义：（∑（Xi-Yi）^2）^1/2，目的是计算其间的整体距离。欧氏空间可以理解为几何空间的度量性在线性空间推广的结果，是一种具有了内积的线性空间。

[2] 真实问题高维变量：multi-scale/arbitrary/not consistently measurable/ heterogeneous。

[3] 我们熟悉的欧氏距离虽然很有用，但也有明显的缺点。它将样品的不同属性（即各指标或各变量）之间的差别等同看待，这一点有时不能满足实际要求。例如，在教育研究中，经常遇到对人的分析和判别，个体的不同属性对于区分个体有着不同的重要性。因此，有时需要采用不同的距离函数。

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

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

[1] L2 范数，使特征平滑，在统一测度中保持同分布

[2] PCA 降维，保证正交

[3] kernel trick 升维：如果数据在当前空间中不是线性可分的，则需做 transform，将数据变换到更高的维度空间中，kernel function 的本质是计算内积。

[4] 概率图模型来表征变量之间的联系

[5] deep learning: dispose the features

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?

补充：Humans are limited that can not easily perceive or deal with high-dimensional problem.

[1] 之所以能对高维数据进行降维，是因为数据的原始表示常常包含大量冗余：

[2] 有些变量的变化比测量引入的噪声还要小，因此可以看作是无关的

[3] 有些变量和其他的变量有很强的相关性(例如是其他变量的线性组合或是其他函数依赖关系)，可以找到一组新的不相关的变量。

[4] 从几何的观点来看，降维可以看成是挖掘嵌入在高维数据中的低维线性或非线性流形。这种嵌入保留了原始数据的几何特性，即在高维空间中靠近的点在嵌入空间中也相互靠近

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

[1] 当没有足够多的数据支持高复杂度模型时，模型易受噪声样本影响而过拟合：方差（variance）用来测量预测结果对于任何给定的测试样本会出现多大的变化。方差过高表明模型无法将其预测结果泛化到更多的数据。对训练集高度敏感也称为过拟合（overfitting）。

补充11题：tend to capture some important features; overfitting is worse than underfitting; ease the pain from feature selection and more easy to learn; loss possible to get a local sol

[2] 符合奥卡姆剃刀原理。

12) What is a loss function? Give three examples (Square, Log, Hinge) and describe their shapes and behaviors;

（1）what is the loss function? （three examples）

Loss function: quantifies the amount by which the prediction deviates from the actual values. (In classification problem, it represents the price paid for inaccuracy of prediction.)

损失函数（loss function）是用来估量模型的预测值f(x)与真实值Y的不一致程度，它是一个非负实值函数,通常使用L(Y, f(x))来表示，损失函数越小，模型的鲁棒性就越好。

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i; \theta) + \lambda \, \Phi(\theta)$$

Hinge loss（用于SVM）

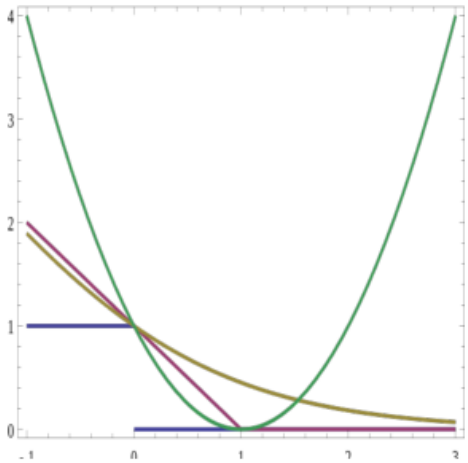
$$\ell_{hinge}(z) = \max(0, 1 - z)$$

Least square（用于Linear regression）

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y - xW^T\}^2$$

Logistic loss（用于Logistic regression）

$$\ell_{log}(z) = \log(1 + \exp(-z))$$



Green is the square loss function. Purple is the hinge loss function. Yellow is the logistic loss function.

Summary:

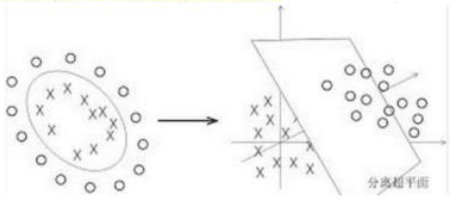
- logistic loss 和 hinge loss的收敛速度是相似的，但是hinge（用于SVM），在支持向量边界外分类正确的点不会有penalty，也就是说，这些点的loss值为0，Hinge loss的“零区域”，使得支持向量机的解具有稀疏性；logistic损失函数中，所有点都有penalty，只不过对于置信度较高的点来说cost value会非常小。这种性质使得logistic的解依赖于更多的训练样本，预测开销更大；但logistic 回归的主要优势是在给出预测标记的同时也给出了概率
- logistics loss 和 least square 都是smooth的，但是least square 过分地对那些outlier敏感了，因此它比hinge、logistics都要收敛的慢
- 三个函数都是凸的且连续，但是注意，hinge是不光滑的，追求相对正确（与margin有关）

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;

用hinge loss 的SVM：

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \max\left(0, 1 - y_i(w^T x_i + b)\right)$$

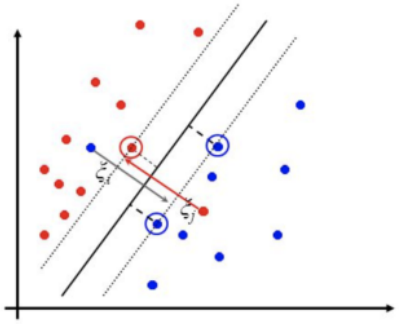
Risk 1：Linear Non-separable（线性不可分）



- **The kernel trick**
  - 原始样本空间内也许不存在可以正确划分样本的超平面，可以将样本从原始空间映射到一个更高维到特征空间（feature mapping），使得样本在这个特征空间内线性可分。如果原始空间是有限维，那么一定存在一个高维特征空间使得样本可分。
  - 为什么需要kernnel function? $<\Phi(x),\Phi(z)>$ 是样本x、z映射到特征空间后到内积，由于维数高，计算是困难的，所以设想 $K(x,z)=<\Phi(x),\Phi(z)>$，把问题转化为求在原始样本空间中求 K(X,Z) 的结果
  - 对称函数所对应的核矩阵半正定，那么这个函数就为kernel function

Risk 2：regularization in Non-separable case

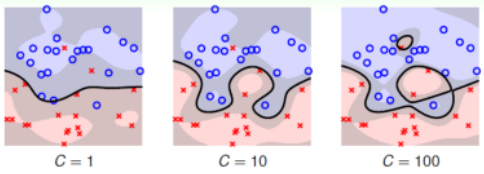
- **soft margin（软间隔）**



- 即便找到了某个核函数使得样本在特征平面线性可分，很难判定这个结果不是由过拟合造成的，所以软间隔的作用就是允许某些样本不满足约束，优化目标为

$$\min_{b,\mathbf{w},\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\cdot\sum_{n=1}^{N}\xi_n$$

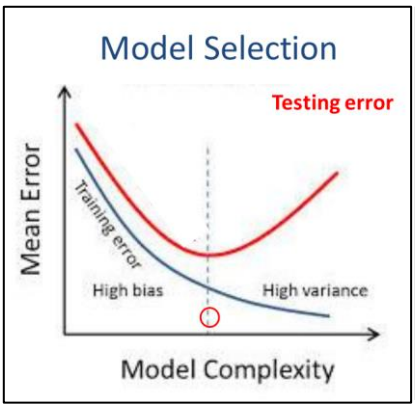
  - 每个样本的松弛变量用以表征样本不满足约束的程度
  - 用C表示更重视margin还是重视犯错误的多少



$C = 1$ $C = 10$ $C = 100$

- 软间隔支持向量机的最终模型仅仅与支持向量有关

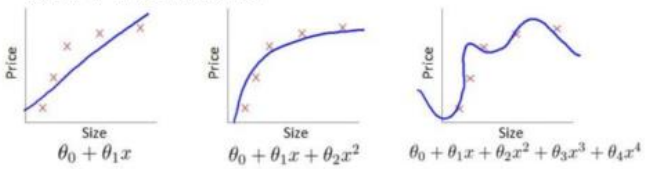
14) Describe biases (under-fitting) and variance (over-fitting) issue in learning, and how can we select and validate an appropriate model?



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

噪声：刻画了学习问题本身的难度



High bias (underfit) | "Just right" | High variance (overfit)

模型选择： 需要使用交叉验证来帮助选择模型

- We are given training data D and test data D_{test}, and we would like to fit this data with a model $p_i(x;\theta)$ from the family $\mathcal{F}$ (e.g, an LR), which is indexed by i and parameterized by θ.
- K-fold cross-validation (CV)
  - Set aside αN samples of D (where $N = |D|$). This is known as the held-out data and will be used to evaluate different values of i.
  - For each candidate model i, fit the optimal hypothesis $p_i(x;\theta^*)$ to the remaining $(1-\alpha)N$ samples in D (i.e., hold i fixed and find the best θ).
  - Evaluate each model $p_i(x|\theta^{*i})$ on the held-out data using some pre-specified risk function.
  - Repeat the above **K times**, choosing a **different** held-out data set each time, and the scores are averaged for each model $p_i(.)$ over all held-out data set. This gives an estimate of the risk curve of models over different i.
  - For the model with the lowest risk, say $p_{i^*}(.)$, we use all of D to find the parameter values for $p_{i^*}(x;\theta^*)$.

- How to decide the values for K and α
  - Commonly used $K = 10$ and $\alpha = 0.1$.
  - when data sets are small relative to the number of models that are being evaluated, we need to decrease α and increase K
  - K needs to be large for the variance to be small enough, but this makes it time-consuming.
- Bias-variance trade-off
  - Small α usually lead to low bias. In principle, *LOOCV* provides an almost unbiased estimate of the generalization ability of a classifier, especially when the number of the available training samples is severely limited; but it can also have high variance.
  - Large α can reduce variance, but will lead to under-use of data, and causing high-bias.
- One important point is that the test data D_{test} is never used in CV, because doing so would result in overly (indeed dishonest) optimistic accuracy rates during the testing phase.

2.2 K-fold Cross Validation

另外一种折中的办法叫做K折交叉验证，和LOOCV的不同在于，我们每次的测试集将不再只包含一个数据，而是多个，具体数目将根据K的选取决定。比如，如果K=5，那么我们利用五折交叉验证的步骤就是：

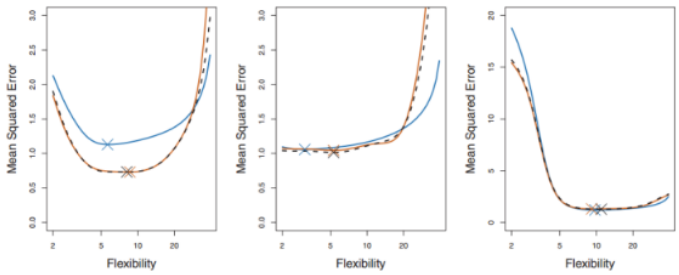
1.将所有数据集分成5份

2.不重复地每次取其中一份做测试集，用其他四份做训练集训练模型，之后计算该模型在测试集上的 MSE_i

3.将5次的 MSE_i 取平均得到最后的MSE

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$

不难理解，其实LOOCV是一种特殊的K-fold Cross Validation（K=N）。再来看一组图：



每一幅图种蓝色表示的真实的test MSE，而黑色虚线和橙线则分贝表示的是LOOCV方法和10-fold CV方法得到的test MSE。我们可以看到事实上LOOCV和10-fold CV对test MSE的估计是很相似的，但是相比LOOCV，10-fold CV的计算成本却小了很多，耗时更少。

15) How to control model complexity in the regression of a linear model? Are there supposed to be a unique low-dimensional model for a given high dimensional problem?

为了避免出现过拟合的情况，一般加入对模型复杂度的限定，即正则项（Regularizer）。一般常用的正则项有L1和L2正则项

- linear regression
  - $$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$
- logistic regression
  - $$J(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}\times log(h_\theta(x^{(i)})) + (1-y^{(i)})\times log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

上述的带正则项的优化过程就是以结构风险最小化为原则，在寻求经验风险最小的同时还要结合当前模型的复杂度，复杂度越高，惩罚程度也就越大。

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?

对于linear regression：

1）将原问题表现为下面这种形式（ε为线性拟合时所产生的误差，这一误差项我们可以看作是数据集内在的未捕捉到的特征或者看成随机的噪声）

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

2）*if the underlying distribution is Gaussian（误差项满足高斯分布）*

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

3）将上述两项组合起来得到

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

即 $y^{(i)} \mid x^{(i)};\theta \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$

每一组样本(x,y)之间是毫无关联，因此从这一概率密度函数p中抽取得到的结果就相当于每次抽取的结果的乘积，称为参数θ相对于样本集X的似然函数(Likelihood Function)

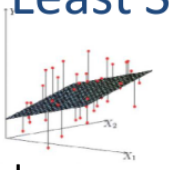
$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

4）对数似然函数（对数函数单调递增，不会对原函数有影响）

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\ &= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2 \end{aligned}$$

可以看到，求对数似然函数的最大值即为求第二项最小值，而它正好是最小二乘法

To Prove the Gaussian Nature of Least Square



- Assume data are i.i.d. (independently identically distributed)
  - Likelihood of L(θ) = the probability of y given x parameterized by θ

$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y} \mid X;\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta)$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

- What is Maximum Likelihood Estimation (MLE)?
  - Chose parameters θ to maximize the function $L(\theta) = p(\vec{y} \mid X;\theta)$, so to make the training data set as probable as possible.

The Equivalence of MLE and LS

instead maximize the **log likelihood** $\ell(\theta)$:

$$\begin{aligned}\ell(\theta) &= \log L(\theta) \\ &= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\ &= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2\end{aligned}$$

Hence, maximizing $\ell(\theta)$ gives the same answer as minimizing

$$\frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2,$$

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?

17题补充：convexity means the global optimum is unique and we can use gradient-based method easily to find it

1） **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lost function的角度建模，一般就是求使得loss fucntion最小的参数。

loss function 是关于参数的凸函数，由于在凸函数中，任何极小值都是最小值，因此当它关于w和b的导数为0的时候，得到w和b的最优解，因此，用梯度下降法可以找到最优解的基础。

2）**likelihood-based**

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，需要有分布假设。

两者都是把估计问题变为了优化问题，用梯度下降（上升）可以找到最优解，因为每一个step，都会令模型更好，不会浪费，最终会找到全局最优的点。若不是凸优化问题，则会使得优化问题找到的解是局部最优解。

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?

SMO算法主要用于解决支持向量机目标函数的最优化问题。考虑数据集$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$的二分类问题，其中$\mathbf{x}_i$是输入向量，$y_i \in \{-1, 1\}$是向量的类别标签，只允许取两个值。一个软边缘支持向量机的目标函数最优化等价于求解以下二次规划问题的最大值：

$$W = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j K(x_i, x_j)\alpha_i \alpha_j,$$

满足：

$$0 \le \alpha_i \le C, \quad \text{for } i = 1, 2, \ldots, n,$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0,$$

其中，C是SVM的参数，而$K(\mathbf{x}_i, \mathbf{x}_j)$是核函数。这两个参数都需要使用者制定。

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta).$$

1.是不是一个凸的梯度？如果出现局部解，怎么从局部解跳出来？
解决方法：同时搜索，某一些搜索可以找到更好的局部解
2.alpha的选择
3.下降的速度：一般用一阶导，二阶导下降更快
4.同时要有很多梯度，梯度搜索并行协同的问题，如何优化加快？通过distribute同步，增加搜索效率
5.梯度的顺序：先去搜索哪个参数
（随机梯度下降：不是按照模型本身的顺序，而是随机挑选数据和梯度进行下降，可以比较快的收敛）

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(最大似然估

计)-based generative methods from a Bayesian perspective?

Any type of classification can be seen as a *probabilistic generative model* by modeling the class-conditional densities `p(x|C_k)` (i.e. given the class `C_k`, what's the probability of `x` belonging to that class), and the class priors `p(C_k)` (i.e. what's the probability of class `C_k`), so that we can apply Bayes' theorem to obtain the posterior probabilities `p(C_k|x)` (i.e. given x, what's the probability that it belongs to class `C_k`). It is called *generative* because, as Bishop says in his book, you could use the model to generate synthetic data by drawing values of `x` from the marginal distribution `p(x)`.

This all just means that every time you want to classify something into a specific class (e.g. size of a tumor being malignant of benign), there will be a probability of that being right or wrong.

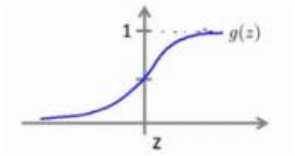
Logistic regression uses a sigmoid function (or logistic function) in order to classify the data. Since this type of function ranges from 0 to 1, you can easily use it to think of it as probability distributions. Ultimately, you're looking for `p(C_k|x)` (in the example, `x` could be the size of the tumor, and C_0 the class that represents benign and C_1 malignant), and in the case of logistic regression, this is modeled by:

```
p(C_k|x) = sigma( w^t x )
```

where `sigma` is the sigmoid function, `w^t` is the transposed set of weights `w`, and `x` is your feature vector.

1) Discriminative learning algorithm （判别学习算法）：直接去预测后验 P(y|x)，即直接预测判别函数的算法

2）概率学解释：逻辑回归是**判别学习算法**，从输入直接可以映射到{0,1}



sigmoid函数：

在逻辑回归中，令

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

逻辑回归中，假设事件符合**伯努利分布**，因此有

$$P(y=1 \mid x;\theta) = h_\theta(x)$$
$$P(y=0 \mid x;\theta) = 1 - h_\theta(x)$$
$$\longrightarrow \quad p(y \mid x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

因为y非0即1

（理解：当h>=0.5 时，若预测y=1，预测正确的概率为h，预测错误的概率为1-h，同理...）

3）逻辑回归与极大似然法的关系

由上述得到似然方程&对数似然方程

$$L(\theta) = p(\vec{y} \mid X;\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) = \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)}))$$

通过证明可以得到对数似然方程是凹函数，在凹函数中，任何极大值也是最大值

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（指数分布族）？ 见手写笔记

[1] 后验概率公式为：$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}$，逻辑回归公式为：略

[2] 指数分布族的定义：$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$。一组确定的 T，a 和 b 定义了这样一个以 η 为参数的分布族。对于不同的 η，我们可以得到指数分布族中不同的分布。

[3] 逻辑回归认为 P(Y|X)服从伯努利二项分布，伯努利二项分布属于指数分布族的证明如下：

3、二项分布属于指数分布族的证明

对于二项分布（伯努利分布），每一个取不同均值的参数Φ，就会唯一确定一个y属于{0,1}之间的分布。所以可以表示为

$$p(y=1;\phi)=\phi;\ p(y=0;\phi)=1-\phi.$$

故二项分布的分布函数只以Φ作为参数，统一这样表示二项分布：

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi+(1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y+\log(1-\phi)\right) \end{aligned}$$

这样，自然参数为：$\eta=\log(\phi/(1-\phi))$，翻转一下，有：$\phi=1/(1+e^{-\eta})$

为了进一步将二项分布向指数分布族靠拢，我们可以进行如下表示：

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

这显示了二项分布可以被写成是指数分布族的形式，所以二项分布属于指数分布族。

进一步地，我们用指数分布族的性质去验证一下，有：

$$\dot{a}(\eta)=\frac{d}{d\eta}(\log(1+e^{\eta}))=\frac{e^{\eta}}{1+e^{\eta}}=\phi,$$

$$\ddot{a}(\eta)=\frac{d^2a(\eta)}{d\eta^2}=\frac{d\dot{a}(\eta)}{d\eta}=\frac{d}{d\eta}\left(\frac{e^{\eta}}{1+e^{\eta}}\right)=\frac{e^{\eta}}{(1+e^{\eta})^2}=\phi(1-\phi)$$

刚好是二项分布的期望与方差，故满足性质。

22) Can you provide a probabilistic comparison for linear and logistic regression? 见手写笔记

23) Why the log of odd would be something related to entropy and effective information? 见手写笔记

24答案：(computation)computationally better since the model complexity is much lower, which limits to 0-1; (conceptually)convert the problem to a binary problem, even can further divide the divided part;(robustness)逻辑回归更鲁棒tolerant to noise, sometimes value isnot important, the threshold is more informative than a num.

Shannon Entropy (1948)

$$H=-\sum_{i=1}^{n} p_i \log p_i$$

- H(p) is monotonically decreasing in p
- An increase in the probability of an event decreases the information from an observed event, and vice versa.
- H(p) ≥ 0 – information is a non-negative quantity.
- H(1) = 0 – events that always occur do not communicate information.
- H(p1 p2) = I(p1) + I(p2) – information due to independent events is additive.

24) Why often we want to convert a linear to a logistics regression, conceptually and computationally?

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

Discriminative models

A discriminative model is a model of the conditional probability of the target Y, given an observation X, symbolically, $P(Y|X)$. Classifiers computed without using a probability model are also referred to loosely as "discriminative."

- Probabilistic: $p_\theta(y|x)$
- Deterministic: $y = g_\theta(x)$ Sometimes, deterministic models can be represented as $y = \text{argmax}_y w^\top \phi(x,y)$. In this case, $p_\theta(y|x) \propto e^{w^\top \phi(x,y)}$.

In deep learning, a discriminative neural network with the softmax output layer is

$$p_\theta(y = k|X) = p_k = \frac{1}{Z(\theta)} \exp(f_\theta^{(k)}(X)),$$

where $f_\theta^{(k)}(X)$ is a K-dimensional network output with respect to the input X before the softmax layer. $f_\theta^{(k)}(X)$ indicates the k-th output dimension.
for $k = 1, \ldots, K$, and

$$Z(\theta) = \sum_k \exp(f_\theta^{(k)}(X)).$$

Generative models

In statistical classification, two main approaches are called the generative approach and the discriminative approach. These compute classifiers by different approaches, differing in the degree of statistical modelling.
Given an observable variable X and a target variable Y, a generative model is a statistical model of the joint probability distribution on $X \times Y$, $p_\theta(X,Y)$;

Do the conditional inference.

$$p_\theta(Y|X) = \frac{p_\theta(X,Y)}{p_\theta(X)} = \frac{p_\theta(X,Y)}{\sum_{Y'} p_\theta(X,Y')}$$

In deep learning, the generative neural network g_θ is usually given as

$$h \sim p(h) \sim \text{N}(0,I),\ X = g_\theta(h) + \varepsilon,$$

where h is the input of the generative neural network, which is a low-dimensional vector. X is the network output, *e.g.* an generated image.

$$p_\theta(X) = \int p(h) p_\theta(X|h) dh.$$

$$X - g_\theta(h) = \varepsilon \sim \text{N}(0, \sigma^2 \mathbf{I}) \quad \Rightarrow \quad X \sim \text{N}(g_\theta(h), \sigma^2 \mathbf{I})$$

$$\log p_\theta(X|h) = -\frac{\|X - g_\theta(h)\|^2}{2\det(\sigma^2 \mathbf{I})} + \text{constant},$$

where $\Sigma = \sigma^2 \mathbf{I}$ is the covariance matrix of ε.

$$\log p(h) = \frac{1}{2}\|h\|^2 + \text{constant}.$$

26) What multinomial(多项式) is the most important assumption for something Naïve but still very effective? For instance, for classifying different documents? 见手写笔记

多项式分布（Multinomial Distribution）是二项式分布的推广。 二项分布的典型例子是扔硬币，硬币正面朝上概率为 p，重复扔 n 次硬币，k 次为正面的概率即为一个二项分布概率。把二项分布公式推广至多种状态，就得到了多项分布。

在单次试验中，假设一共有k种可能情况，记这k种可能发生的概率为$\mu = [\mu_1, \ldots, \mu_k]$，并且$\sum_{i=1}^k \mu_i = 1$，记$\mathbf{x} = [x_1, \ldots, x_k]$，其中$x_i \in \{0,1\}$，并且$\sum_{i=1}^k x_i = 1$，即$x_i$中只有一个为1，其他均为0，也就是每次试验只有一种可能发生，x_i取1的概率为μ_i，那么，$\mathbf{x}$的概率为

$$P(\mathbf{x}|\mu) = \prod_{i=1}^k \mu_i^{x_i}$$

将试验进行N次，记第i种可能发生的次数为m_i，$\sum_{i=1}^k m_i = N$，那么多项分布表示m_i的联合概率分布

$$P(m_1, \ldots, m_k|N, \mu) = Multi(m_1, \ldots, m_k|N, \mu) = \frac{N!}{m_1! \cdots m_k!} \prod_{i=1}^d \mu_i^{m_i}$$

参考资料：https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note07-2up.pdf

4 The multinomial document model

In the multinomial document model, the document feature vectors capture the frequency of words, not just their presence or absence. Let $\mathbf{x}_i$ be the multinomial model feature vector for the i th document D_i. The t th element of $\mathbf{x}_i$, written x_{it}, is the count of the number of times word w_t occurs in document D_i. Let $n_i = \sum_t x_{it}$ be the total number of words in document D_i.

Let $P(w_t|C)$ again be the probability of word w_t occurring in class C, this time estimated using the word frequency information from the document feature vectors. We again make the naive Bayes assumption, that the probability of each word occurring in the document is independent of the occurrences of the other words. We can then write the document likelihood $P(D_i|C)$ as a multinomial distribution (equation 6), where the number of draws corresponds to the length of the document, and the proportion of drawing item t is the probability of word type t occurring in a document of class C, $P(w_t|C)$.

$$P(D_i|C) \sim P(\mathbf{x}_i|C) = \frac{n_i!}{\prod_{t=1}^{|V|} x_{it}!} \prod_{t=1}^{|V|} P(w_t|C)^{x_{it}}$$
$$\propto \prod_{t=1}^{|V|} P(w_t|C)^{x_{it}}. \quad (7)$$

We often often won't need the normalisation term $(n_i!/\prod_t x_{it}!)$, because it does not depend on the class, C. The numerator of the right hand side of this expression can be interpreted as the product of word likelihoods for each word in the document, with repeated words taking part for each repetition.

As for the ~~Bernoulli~~ model, the parameters of the likelihood are the probabilities of each word given the document class $P(w_t|C)$, and the model parameters also include the prior probabilities $P(C)$. To estimate these parameters from a training set of documents labelled with class $C = k$, let z_{ik} be an indicator variable which equals 1 when D_i has class $C = k$, and equals 0 otherwise. If N is again the total number of documents, then we have:

$$\hat{P}(w_t \mid C = k) = \frac{\sum_{i=1}^{N} x_{it} z_{ik}}{\sum_{s=1}^{|V|} \sum_{i=1}^{N} x_{is} z_{ik}}, \quad (8)$$

an estimate of the probability $P(w_t \mid C = k)$ as the relative frequency of w_t in documents of class $C = k$ with respect to the total number of words in documents of that class.

The prior probability of class $C = k$ is estimated as before (equation 4).

$$\hat{P}(C = k) = \frac{N_k}{N}. \quad (4)$$

Thus given a training set of documents (each labelled with a class) and a set of K classes, we can estimate a multinomial text classification model as follows:

1. Define the vocabulary V; the number of words in the vocabulary defines the dimension of the feature vectors.
2. Count the following in the training set:
   - N the total number of documents,
   - N_k the number of documents labelled with class $C = k$, for each class $k = 1, \ldots, K$,
   - x_{it} the frequency of word w_t in document D_i, computed for every word w_t in V.
3. Estimate the likelihoods $P(w_t \mid C = k)$ using (8).
4. Estimate the priors $P(C = k)$ using (4).

To classify an unlabelled document D_j, we estimate the posterior probability for each class combining (1) and (7):

$$P(C|D_j) = P(C|\mathbf{x}_j)$$
$$\propto P(\mathbf{x}_j|C)\,P(C)$$
$$\propto P(C) \prod_{t=1}^{|V|} P(w_t|C)^{x_{it}}. \quad (9)$$

27) What would be the most effective way to obtain a really universal prior? And what would be the most intriguing implications(启示/暗示) for human intelligence?

[1] 本体(ontology)，知识图谱(knowledge graph)：从抽象层面看，本体最抽象，其次是知识库(knowledge base)，最后才是知识图谱。举个例子，如果我们要做图书领域的知识库或者知识图谱，首先要对图书进行分类，这个分类就是本体，比如说，图书分为计算机类和电子类，计算机类有分为网络、人工智能；有了这个分类后，我们就可以把图书都分到每个类别，比如说《Zero to One》是一本进口原版书，然后这本书有各种属性－属性值，比如说书的作者是 Peter Thiel，这些数

据就构成了一个图书知识图谱（前面讲的分类可以认为不是这个知识图谱的一部分），而这里分类和知识图谱一起可以看成是一个图书知识库。也就是说，本体是强调概念关系，知识图谱强调实体关系和实体属性值，知识库则是所有知识的集合。但是知识库不局限于分类和图谱，知识库可以包括规则，包括过程性知识等。而本体也可以定义得很抽象，任何概念的内涵和外延可以定义本体。

priors become defined, refined and confined

[2] 人类智慧的基础是先验（PPT 原话），在进化中不断 ~~define and re-define~~ the structure of our mind；人类智慧和机器智能的根本区别：人类智慧是遗传得到的，机器智能是穷举习得的。大佬note: human knowledge form intriguing structures

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

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

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

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

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?

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?

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

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

[1] *人们提出的所有计算模型都能够用图灵机模型模拟*：根据图灵的研究，直观上讲，所谓计算就是计算者（人或机器）对一条两端可无限延长的纸带上的 0 和 1 执行操作，一步一步地改变纸带上的 0 或 1 的值，经过有限步骤最终得到一个满足预先要求的符号串的变换。他将此计算过程符号化为一个五元组，机器从给定的纸带上某起始点出发，其动作完全由初始状态及五元组来决定。图灵机是一种为可解问题设计的一种计算装置，它不是一台具体的现代意义上的计算机，但它却是一种操作十分简单且运算能力很强的计算装置，用其来计算所有可能想象到的可计算函数。图灵机模型的实际意义在于：图灵证明，只有图灵机能解决的计算问题，实际计算机才能解决；如果图灵机不能解决的计算问题，则实际计算机也无法解决。即可计算性=图灵可计算性。因此，图灵机的能力概括了数字计算机的计算能力，对计算机的一般结构、可实现性和局限性产生了深远的影响。图灵机等计算模型均是用来解决“能行计算”问题的，理论上的能行性隐含着计算模型的正确性，而实际实现中的能行性还包含时间与空间的有效性。

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

[1] *停机问题在图灵机上是不可判定问题。*

[2] 计算机（包括图灵机等类似计算模型）只是二阶逻辑的物理实现，二阶逻辑上不可判定的问题计算机当然也做不出来。

[3] 图灵机是串行机，所以做不到真正的并行

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

[1] *希尔伯特第十问题又称“判定丢番图方程的可解性”对其具体的描述为“给定一个系数均为有理整数，包含任意个未知数的丢番图方程：设计一个过程，通过有限次的计算，能够判定该方程在有理数整数上是否可解。”*

[2] 希尔伯特第十问题留下了两个悬念。第一个悬念是科学的“算法”定义。在那个年代，有限的、机械的证明步骤在数学上还没有严格的定义，人们只能凭着感觉去定义这样一种模糊的表达方式。今天我们知道，这样一个问题，本质就是“算法”的概念。第二个悬念则是问题的答案。如果问题的答案是否定的，那将意味着可能存在着大量数学问题人们永远无法知道其答案是否存在，自然也就无法去找到解决它的办法。人们面对这样的问题只能束手无策。因此一个可以计算的机器可能从诞生之初就有其无法逾越的极限。

[3] 到了 30 年代，图灵和丘奇分别从不同的抽象角度提出了有效机械算法的概念。其中图灵提出的图灵机模型倾向于硬件性，且模型直观形象，很快得到了人们的普遍接受。通过图灵机模型，人们第一次理解了“算法”这一基本的深刻概念。也正因为图灵奠定的理论基础，人们才有可能发明改变现代文明的工具：计算机。然而，图灵的立足点不仅于此。为了解决计算机是否存在着理论上的极限这个悬念，图灵对计算机这一概念有了更深的思考。这就是著名的图灵机的停机问题。

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

[1] *希尔伯特第十三问题：Can a given high-dimensional problem be described by a composition with a finite number of bivariate functions? 证明 f^{7}+xf^{3}+yf^{2}+zf+1=0 这个方程式的七个解，若表成系数为 x, y, z 的函数，则此函数无法简化成两个变数的函数。*

[2] 阿诺德对连续函数（仅有连续函数的前提下才能得到答案）的情形解决了希尔伯特第十三问题，得到了每一个三元连续

函数都能表示为二元连续函数的叠加的结论。Kolmogorov–Arnold Representation Theorem (1956) – Every multivariate continuous function can be represented as a superposition of continuous functions of two variables

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. （开放题）

[1] 可计算理论是让你认清图灵机的界限在哪里，也即人类无法通过机械计算得出结论的界限在哪里，而不是让你认清人类的极限在哪里。人类的极限属于哲学才能回答的范畴。

[2] 如果我们认同人类的大脑也是个计算模型，那么按照哥德尔不完备性定理来说，一定也存在这我们人类无法判断是真是伪的东西，这也就是说，其实怀疑论是对的。我们人类并无感知和了解世界的能力。多大的一个遗憾啊。

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

[1] 毫无疑问，人类进化的成功在很大程度上是因为人类大脑掌握了贝叶斯数学原则。在贝叶斯概率中，推理是基于由经验而来的概率。这就如同是人类学会了如何应对这个或然的世界，并做出适合的行为。

[2] 如图：

为了明确样本集中的样本个数，采用以下记号：

$$D^n=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\},\ D^{n-1}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{n-1}\},\ \ldots,\ D^1=\{\mathbf{x}_1\}$$

当n>1时，有

$$p(D^n\mid\boldsymbol{\theta})=\prod_{k=1}^{n}p(\mathbf{x}_k\mid\boldsymbol{\theta})=p(\mathbf{x}_n\mid\boldsymbol{\theta})\prod_{k=1}^{n-1}p(\mathbf{x}_k\mid\boldsymbol{\theta})=p(\mathbf{x}_n\mid\boldsymbol{\theta})p(D^{n-1}\mid\boldsymbol{\theta})$$

$$\begin{aligned}p(\boldsymbol{\theta}\mid D^n)&=\frac{p(D^n\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D^n\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}\\&=\frac{p(\mathbf{x}_n\mid\boldsymbol{\theta})p(D^{n-1}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}_n\mid\boldsymbol{\theta})p(D^{n-1}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}\\&=\frac{p(\mathbf{x}_n\mid\boldsymbol{\theta})p(\boldsymbol{\theta}\mid D^{n-1})}{\int p(\mathbf{x}_n\mid\boldsymbol{\theta})p(\boldsymbol{\theta}\mid D^{n-1})d\boldsymbol{\theta}}\end{aligned}$$

注意，当尚未有观测样本时，令 $p(\theta|D^0)=p(\theta)$ 。反复运用上述公式，能够产生一系列的概率密度函数： $p(\theta),p(\theta|x_1),p(\theta|x_1,x_2)$ 等等。这一过程被称为参数估计的**递归的贝叶斯估计**。这一过程属于增量学习或在线学习算法，其特点是学习的过程随着观察数据的的不断获得而不断进行。如果这一概密度函数的序列最终能够收敛到一个中心在参数真实值附近的狄拉克函数，那么就实现了贝叶斯学习过程。

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

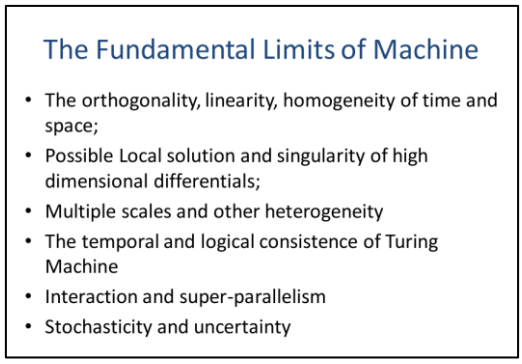
44) Explain the core idea of machine learning to decompose a complex problem into an integration of individual binary problems, its mathematical and computational frame works.

[1] 首先，阿诺德表示奠定了理论基础。Kolmogorov–Arnold Representation Theorem (1956) – Every multivariate continuous function can be represented as a superposition of continuous functions of two variables，在机器学习中这个二元边界问题可以用 sigmoid function 这种概率模型来表示。

[2] from computing point of view: most feasible approach

[3] from complexity point of view: 0-1 is the most primitive(原始的) and representative thing that a computer can do.

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



46) Why differentials of composite non-linear problems(复合非线性函数) can be very complex and even singular?

[1] 从数学形式上看，深度神经网络就是一个多层复合函数。可以证明（Kolmogorov–Arnold 表示定理），任意一个多元函数可以表示成若干个单变量函数的复合。这为使用深度神经网络用来逼近任意高维函数提供了理论基础。由于整个网络是一个复合函数，因此，训练的过程就是不断地应用“复合函数求导”（链式法则）及“梯度下降法”。

[2] complex: back propagation，链式法则，local solution，a number of differentials couple to each other，梯度消失问题

[3] singular: 复合函数的导数非常复杂时，可能会是奇异的。系统可能会进入混沌状态（神经网络的初值敏感问题）

[4] 混沌：对一个非线性系统的某个吸引子，如果从具有相近的初始条件开始的运动轨道随着时间越来越分开，则称这种吸引子为奇异吸引子，并称此时系统处于混沌状态。奇异吸引子的主要特点是运动对初始条件的敏感依存性。

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（英语：halting problem）是逻辑数学中可计算性理论的一个问题。通俗地说，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。该问题等价于如下的判定问题：是否存在一个程序 P，对于任意输入的程序 w，能够判断 w 会在有限时间内结束或者死循环。艾伦·图灵在 1936 年用对角论证法证明了，不存在解决停机问题的通用算法。这个证明的关键在于对计算机和程序的数学定义，这被称为图灵机。停机问题在图灵机上是不可判定问题。这是最早提出的决定性问题之一。

[2] question 1: temporal order。图灵机是串行机。In the Turing Machine, as in any procedural account of events. temporal order enters as an external parameter.

[3] question 2: In Turing machines, the iteration of logics takes order, then the result can be valued. No true parallelism.

48) What is mathematical heterogeneity and multiple scale?

[1] mathematical heterogeneity：异质性（heterogeneity）其实也就是我们经常所谓的差异、差别。它可以是个体层面上，也可以是群体层面上。前者属于个体异质性，后者属于总体异质性。异质性无处不在，这也是社会科学研究的真正本质。很多的统计方法都假定总体是同质的。所以实际研究中，经常看到数据被当作仿佛是从一个单一总体中得到而加以分析，尽管往往样本中所有的个体可能并不具有相同的一套参数值。实际上，研究者们也经常意识到一个总体可能异质的，是由多个不同的子总体混合而成的，比如男性和女性、城镇居民与农村居民。为此，尝试在模型设定和选择上尽可能地考虑能放宽同质性总体假定，以便得到更合理的认识或对更复杂的理论假说做出实证检验。

[2] multiple scale：图像金字塔很重要，将不同 scale 的图像送入网络提取出不同 scale 的特征做融合，对于整个网络性能的提升很大，但是由于图像金字塔的多尺度输入，造成计算且保存了大量的梯度在内存，从而导致对硬件的要求很高，而且测试时，增加了计算时间。CNN 网络的每一个参数，几乎都会影响到产生的特征图。 这隐含着固定架构的网络，往往会学习到特定尺度的特征，也可能具有了一定的尺度不变性。 同时，这些参数往往与当前的任务紧密相关，不能轻易修改，于是使得模型很难学到多尺度的信息。

[3] YB：RNN for example, for different time point T, each of their local temporal loss cannot directly integrate into the global loss because each have different scales and cannot be compared in a universal system. 而 CNN 等模型，we take it for granted that the inputs and outputs are normalized and scaled. 所以我们需要用 deeper networks to learn the integration of all the heterogenous features，这其中最重要的是非线性的激活函数去 categorize the outputs.

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

[1] 复合函数从理论上说是非凸问题，但是有些时候是凸函数（神经网络中固定(fix)其他网络层的参数只更新当前层的参数, fix the composite function and only focus on the certain linear combination）。Deep learning is proposed to relax those propagations allowing any kinds of combinations, thus the whole learning will obtain local solutions.

[2] 两种方法：

i. Take advantage of something linear, RELU, potentially probabilistic??????

ii. Partition your networks：regional combination, assuming the regional network is convex. (Capsule Network)

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

Probabilistic graphical model:

1) Compare the graphical representation with feature vector-based and kernel-based representations;
2) Explain why sometime a marginal distribution has to be computed in a graphical model;

1. 有答案还是要总结一下：
2. 减少变量，降低复杂度

3. 它是执行其他任务的先决条件
4. 省掉部分隐变量的影响

3) Why class labels might be the key factor to determine if presumptively different data distributions can be indeed discriminated? (标签噪声)
   1. 在监督学习中，训练数据所对应的标签质量对于学习效果至关重要。如果学习时使用的标签数据都是错误的，那么不可能训练出有效的预测模型。标签噪声可能会增加模型复杂度，降低数据集的利用率，降低分类精度等。
   2. *降低分类精度：*标签噪声对于分类器的预测性能的影响是难以忽略的，甚至训练样本集中小部分标签噪声的引入也会对分类器的决策产生较大大的偏差和失真。常用机器学习算法，如感知器、K 近邻算法、决策树算法和支持向量机等，分类性能都会受到标签噪声的影响。K 近邻算法中，特别是当 K 值取 1 时，一个样本标签的错误标记可能会使得周围邻近的一些样本的错误分类。相对来说，由于支持向量机的良好性能取决于精确的决策面的构建，如果有一个距离分类决策面较远的样本的类别标记错误，也会使得决策面发生位移和扭曲，分类器的预测性能更容易受其影响。
   3. *增加模型复杂度：*标签噪声的存在也会导致训练样本数目的增加和学习模型复杂度的变大。当训练数据集中有标签噪声存在时，决策树的尺寸会相应地增加，其结点数目相对较多，使得训练得到的模型异常复杂。
   4. *集成学习*也受到标签噪声的影响：通过实验发现，Bagging 和 Boosting 两种常用的集成方法在处理有标签噪声的数据集时，展现了不同的效果。数据集中随机加入少量的噪声样本，Bagging 方法中基分类器的差异性会随之增加，最终构建的学习模型具有更好的分类性能。相对来说，Boosting 方法更容易受到标签噪声的负面影响，特别是 AdaBoost 算法，随着迭代次数的增加，由于算法会更多地关注于错分类的样本，必然会使得噪声样本的权值越来越大，进而增加了模型复杂度，降低了算法性能。

   参考：http://gb.oversea.cnki.net/KCMS/detail/detail.aspx?filename=1016211364.nh&dbcode=CMFD&dbname=CMFDTEMP

4) Why knowledge-based ontology (representation) be a possible solution for many prior-based inference problems?
   1. 定义：知识图谱就是把所有不同种类的信息（Heterogeneous Information）连接在一起而得到的一个关系网络。知识图谱提供了从“关系”的角度去分析问题的能力。它可以模拟人类的逻辑推理能力，利用知识作为优化目标的约束，指导深度学习模型的学习；通常是将知识图谱中知识表达为优化目标的后验正则项
   2. *from the logic point of view,* knowledge-based ontology is another data mining approach, trying to understand data in terms of some universal structure; It helps to describe our understanding in terms of some *hierarchy* knowledge (either generative or specific), therefore we are able to classify and cluster。
   3. *from the human understanding point of view*, standardized ontology can generate labels and classification trees since the knowledge-based ontology provides a prior structure that the “document” is supposed to be complied with (遵守) and restricted for the distribution.
5) Why a graphical model with latent variables can be a much harder problem?
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?
7) Why does EM not guarantee a global solution? What is a simple proof for that?
8) Why is K-mean only an approximate and local solution for clustering?
9) How to interpret the HMM-based inference problem from a Bayesian perspective, using the forward/backward algorithm?
10) Show how to estimate a given hidden state for a given series of observations using the alpha and beta factors;
11) How a faster inference process would be constructed, given a converging network?
    1. It is necessary that each individual propagation provides good graphic model structure, good prior, good data;
    2. predominant, estimated, deterministic, from expectation point of view
    3. 在实际应用中，常通过模型量化/剪枝实现网络加速
        i. 量化即通过减少表示每个权重所需的比特数来压缩原始网络。一般模型的内部的计算都采用了浮点数计算，浮点数的计算会消耗比较大的计算资源，如果在不影响模型准确率的情况下，模型内部可以采用其他简单数值类型进行计算的话，计算速度会提高很多，消耗的计算资源会大大减小。
        ii. 神经网络的参数众多，但其中有些参数对最终的输出结果贡献不大而显得冗余，剪枝顾名思义，就是要将这些冗余的参数剪掉。首先，需要根据对最终输出结果的贡献大小来对模型的神经元们排序，然后，舍去那些贡献度低的神经元来，使得模型运行速度更快、模型文件更小。

12) How can an important node (where inference can be significantly and sensitively affected) be detected using an alpha and a beta process? alpha*beta, where the value changes dramatically is 关键节点
    1. Solution: alpha times beta, result is supposed to be sigmoidal (0~1).
    2. when alpha is the determinant one, 只需要看 alpha 就可以判断 where inference can be deterministic
13) Why often an alpha process (forward) is more important than beta (backward)?
    1. prior is important from Bayesian point of view → constrain, eliminate, provide smaller and smaller scenarios

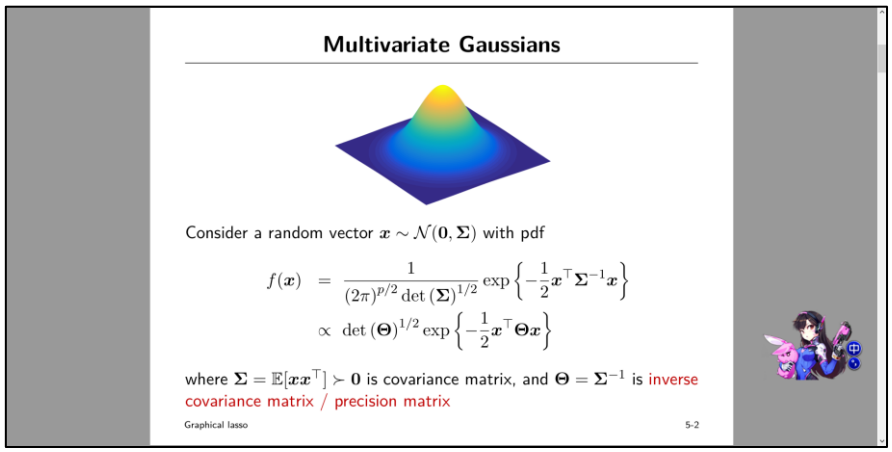
alpha process updates the prior which is important for humans.

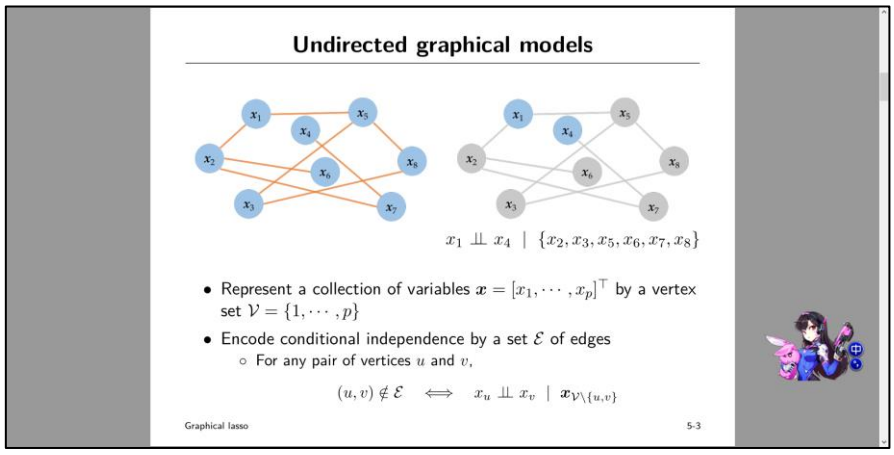
14) What are the key differences between an alpha and a beta process from human and machine intelligence point of views?
    1. machine: no assumption what the model should be, so machine pick the model with highest likelihood → exhausting problem
    2. human: prior → already limits the searching space; alpha becomes predominant more quickly; don't have to see all data, enumerate all selections, understand the whole sequence

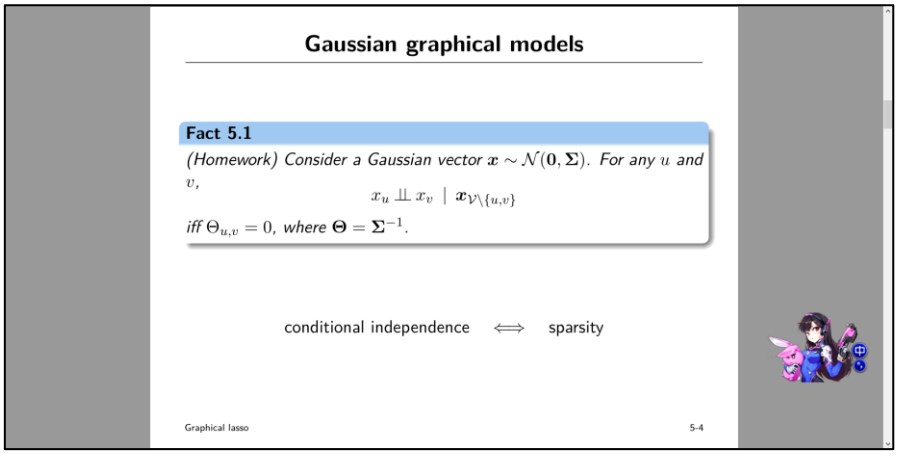
machine: have to exhaust all possible results, can only use data to max likelihood

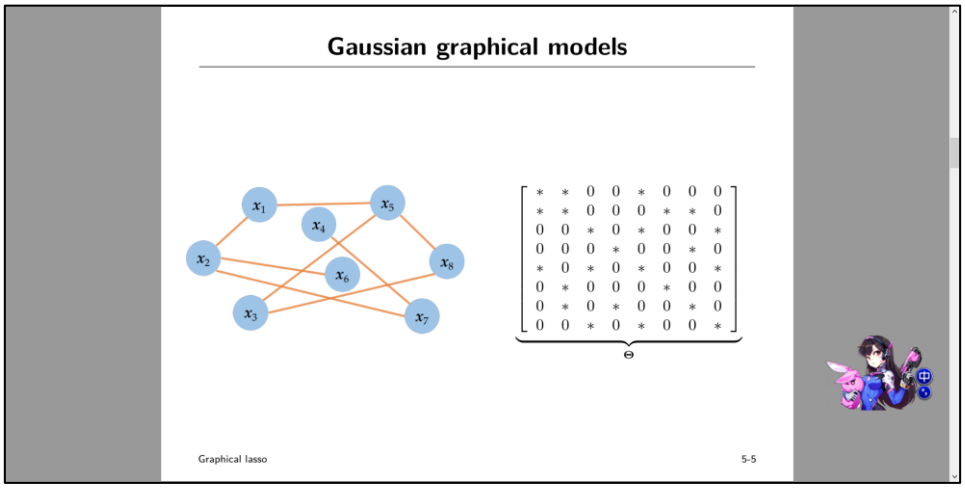
human: take advantages of priors, do not need whoe sequence

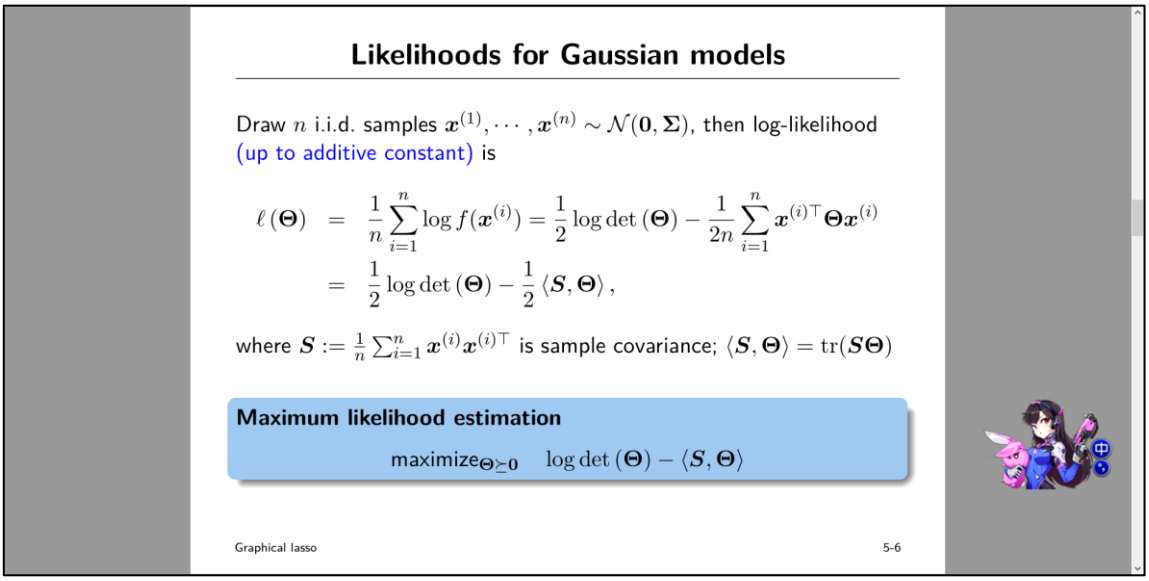
15) How data would contribute to the resolution of an inference process from a structural point of view?
    1. YB: every single update has two parts?
16) For a Gaussian graphical model, what is the implication of sparsity for such a graphical model? How is such sparsity achieved computationally?
17) Explain the objective function of Gaussian graphical model? Thus the meaning of MLE?











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

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?

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

类比L0/L1/L2范数

1. nuclear norm（核范数）定义：矩阵奇异值的求和

$$\|A\|_* = \operatorname{trace}\left(\sqrt{A^*A}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i(A).$$

where $\sigma_i(A)$ are the singular values of A.

1. 既然秩可以度量相关性，而矩阵的相关性实际上有带有了矩阵的结构信息。如果矩阵之间各行的相关性很强，那么就表示这个矩阵实际可以投影到更低维的线性子空间，它就是低秩的。低秩表征着一种冗余程度。秩越低表示数据冗余性越大。rank(W)的凸近似就是核范数。所以核范数常用来约束低秩。

2. spectral norm（谱范数，矩阵的 2 范数）定义：A*A 矩阵的最大特征值开平方根（A*共轭转置）

$$\|A\|_2 = \sqrt{\lambda_{\max}\left(A^*A\right)} = \sigma_{\max}(A)$$ [1] where A^* denotes the conjugate transpose of A.

3. Frobenius norm 定义：

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\operatorname{trace}\left(A^{\mathsf{T}}A\right)} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)},$$

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）

$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$ --------(2)

其中 $U_{m\times m}$ 和 $V_{n\times n}$ 均为正交矩阵， $\Sigma_{m\times n}$ 为对角矩阵

奇异值分解要解决的问题是将 $A_{m\times n}$ 矩阵分解为对角矩阵 $\Sigma_{m\times n}$ ， $\Sigma_{m\times n}$ 中对角元素 σ_i 称为矩阵 $A_{m\times n}$ 的奇异值

2、问题求解方法

$$A^TA = (U\Sigma V^T)^TU\Sigma V^T = V^T\Sigma^TU^TU\Sigma V = V^T\Sigma^T\Sigma V = V^T\Sigma^2V$$

所以 V 是 A^TA 特征值分解的特征向量按行组成的正交矩阵， Σ^2 是 A^TA 特征值组成的对角矩阵，也可以看出 $A_{m\times n}$ 的奇异值 σ_i 是是 A^TA 特征值 λ_i 的平方根。

$\sigma_i = \sqrt{\lambda_i}$ --------(3)

假如 A^TA 的特征向量为 v_i ，则根据式2和式3， U 中对应的 u_i 则可以由下式求出：

$u_i = \frac{Av_i}{\sigma_i}$ --------(4)

也即**奇异值分解的关键在于对 A^TA 进行特征值分解。**

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

1. PCA求解关键在于求解协方差矩阵 $C = \frac{1}{m}XX^T$ 的特征值分解
2. SVD关键在于 A^TA 的特征值分解。

很明显二者所解决的问题非常相似，都是对一个实对称矩阵进行特征值分解，

如果取：

$$A = \frac{X^T}{\sqrt{m}}$$

则有：

$$A^TA = \left(\frac{X^T}{\sqrt{m}}\right)^T\frac{X^T}{\sqrt{m}} = \frac{1}{m}XX^T$$

SVD与PCA等价，所以PCA问题可以转化为SVD问题求解，那转化为SVD问题有什么好处?

有三点：

1. 一般 X 的维度很高， A^TA 的计算量很大
2. 方阵的特征值分解计算效率不高
3. SVD除了特征值分解这种求解方式外，还有更高效且更准确的迭代求解法，避免了 A^TA 的计算

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)

Robust Principal Component Analysis (PCA)

- Would like to split matrix $M \in \mathbb{R}^{n_1 \times n_2}$ into

$$M = L_0 + S_0,$$

where L_0 is low rank and S_0 is sparse.

- Solve the *Principal Component Pursuit (PCP)* problem

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} & L + S = M \end{array}$$

with variables L, $S \in \mathbb{R}^{n_1 \times n_2}$ and problem data $M \in \mathbb{R}^{n_1 \times n_2}$.

- $\|L\|_* = \sum_{i=1}^{r} \sigma_i(L)$ is the nuclear norm
- $\|S\|_1 = \sum_{i,j} S_{ij}$ is the ℓ_1-norm of the matrix S thought of as a vector

Relationship to Classical PCA

- seek the best rank-k estimate of L_0 by solving

$$\begin{array}{ll} \text{minimize} & \|M - L\|_2 \\ \text{subject to} & \mathbf{rank}(L) \leq k. \end{array}$$

- efficiently solved via SVD
- assumes $M - L$ is small noise (in 2-norm)

Advantages of Robust PCA

- robust to *grossly* corrupted observations of M: single large corruption in an entry of M renders arbitrarily bad estimates of L_0.
- sparsity pattern of S is unknown ahead of time
- naturally extends to matrix completion

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; Oliver的答案已整理到手写笔记上

首先，矩阵可以认为是一种线性变换，而且这种线性变换的作用效果与基的选择有关。

以Ax = b为例，x是m维向量，b是n维向量，m,n可以相等也可以不相等，表示矩阵可以将一个向量线性变换到另一个向量，这样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。
A= $\mu\Sigma\sigma^T$ ， μ和σ是两组正交单位向量， Σ是对角阵，表示奇异值，它表示我们找到了μ和σ这样两组基，A矩阵的作用是将一个向量从σ这组正交基向量的空间**旋转**到μ这组正交基向量空间，并对每个方向进行了一定的**缩放**，缩放因子就是各个奇异值。如果σ维度比μ大，则表示还进行了**投影**。可以说奇异值分解将一个矩阵原本混合在一起的三种作用效果，分解出来了。

而**特征值分解**其实是对旋转缩放两种效应的归并。（有投影效应的矩阵不是方阵，没有特征值）特征值，特征向量由Ax= λx得到，它表示如果一个向量v处于A的特征向量方向，那么Av对v的线性变换作用只是一个**缩放**。也就是说，求特征向量和特征值的过程，我们找到了这样一组基，在这组基下，矩阵的作用效果仅仅是存粹的**缩放**。对于实对称矩阵，特征向量正交，我们可以将特征向量式子写成 $A = x\lambda x^T$ ，这样就和奇异值分解类似了，就是A矩阵将一个向量从x这组基的空间旋转到x这组基的空间，并在每个方向进行了缩放，由于前后都是x，就是没有旋转或者理解为旋转了0度。

总结一下，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。我感觉特征值分解其实是一种找特殊角度，让旋转效果不显露出来，所以并不是所有矩阵都能找到这样巧妙的角度。仅有缩放效果，表示、计算的时候都更方便，这样的基很多时候不再正交了，又限制了一些应用。

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? 见手写笔记

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;

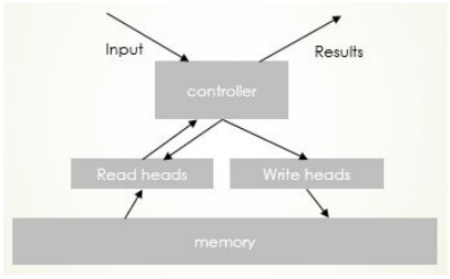
L2：cancel to relative weight and average inputs
L1：keep pruning connections between the layers
sparse networks are computationally tolerant.

- 问题二：sigmoid re-normalize ~~re-categorize~~ all outputs into something ~~zeros and ones~~ (probabilistic); by the way, tanh is better.

recalibrate
get probabilistic outputs(Logistic Loss)

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; 见手写笔记
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? 已与Oliver答案整合，见手写笔记

- 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：



组成是一个表示memory的矩阵，读写头，及RNN作为控制器，上图是Grave提出的架构。另外提一下，NTM的具体细节Grave的论文中只有读写头和memory的交互，至于RNN controller的部分是没有提及的。controller部分的内容是Grave的另一篇论文提及的：

> Generating Sequences With Recurrent Neural Network
> arxiv.org/abs/1308.0850

- RNN can handle sequential data while CNN cannot. YB keyword: RNN do logical process
  可参考 https://datascience.stackexchange.com/questions/11619/rnn-vs-cnn-at-a-high-level

Difference between CNN and RNN are as follows :

CNN:

1. CNN take a fixed size input and generate fixed-size outputs.
2. CNN is a type of feed-forward artificial neural network - are variations of multilayer perceptrons which are designed to use minimal amounts of preprocessing.
3. CNNs use connectivity pattern between its neurons is inspired by the organization of the animal visual cortex, whose individual neurons are arranged in such a way that they respond to overlapping regions tiling the visual field.
4. CNNs are ideal for images and videos processing.

RNN:

1. RNN can handle arbitrary input/output lengths.
2. RNN, unlike feedforward neural networks, can use their internal memory to process arbitrary sequences of inputs.
3. Recurrent neural networks use time-series information (i.e. what I spoke last will impact what I will speak next.)
4. RNNs are ideal for text and speech analysis.

14) What is the central additional difficulty of RNN compared to CNN? 见手写笔记

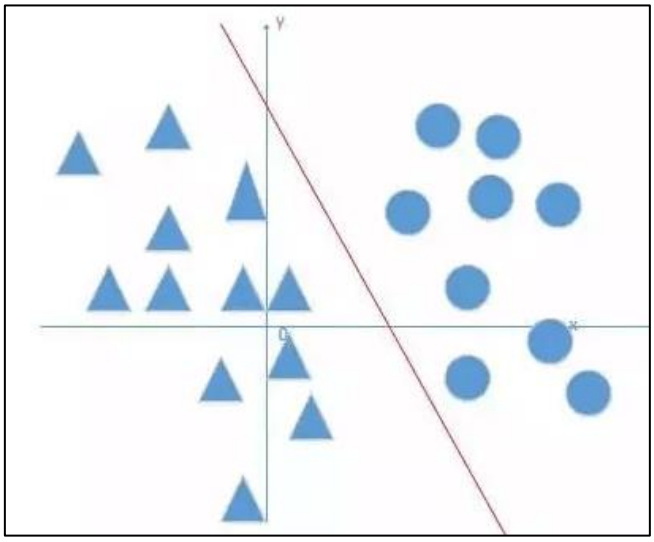
1、RNN/seq2seq的模型，此类模型中的变量是按照时间戳一步步向前（反向传播时，为向后）进行的，如Encoder为例，$h_t = f(h_{t-1}, x_t)$,其中 h 代表隐状态， x 代表输入。这也就决定了循环网络中的计算是按照时间戳（输入、输出位置）来分解计算的，加之 输入往往是可变长序列，且我们想捕获的就是这种时序上的依赖，这种固有的序列化本质阻碍了并行化处理RNN网络的加速。

2、在CNN网络中比较方便利用批处理技术，即 batch by batch的处理数据，而且在per batch内部可以有效地利用向量化技术来加速。RNN网络中就很难做到这一点儿。当然也有一些计算分解技术，但是实现起来就不是那么直观。

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? Oliver keywords: global & local loss, LSTM相关内容见手写笔记

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;
   a) 图灵完备性：https://en.wikipedia.org/wiki/Turing_completeness

b) 数据挖掘的极限：https://content.wisestep.com/data-mining-purpose-characteristics-benefits-limitations

Oliver:
Singular issue;
Multi-scale
Local solutions

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?