Final

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?

a) SVM比Logistic Regression更鲁棒的原因：

- SVM使用内积或核函数，保证了可以用一个非距离测度来表征数据，即所有的运算都是在统一的测度下进行的；LR中样本$x_i = (x_i^1, \cdots, x_i^p)$每一纬度可能不处于同样的测度，即$x_i^j, x_i^k$是无法直接被比较的
- SVM使用hinge loss，其数学形式为$h_{hinge}(z) = \max(0, 1-z)$，中间有一块平坦的区域，因此保证了SVM的稀疏性；而LR采用的是logistic loss，数学形式为$h_{log}(z) = \log(1+\exp(-z))$，是光滑的单调递减函数，在不加范数约束的条件下，无法降维和光滑
- SVM对偶问题的KKT条件中$\alpha_i(1-y_i f(x_i))$，最终模型只与少部分支持向量有关，可以看出SVM比较了所有的数据点，最终选出了支持向量；而LR没有这个筛选过程，直接假设了一个概率模型，然后通过最大化似然函数$L(\theta) = \prod_{i=1}^m p(y_i|x_i;\theta)$来求解

b) 有两种方式来控制SVM的边界：

- 通过核函数将在样本空间线性不可分的数据映射到高维的特征空间，使其线性可分，本质上是一个对内积的映射，即$K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$
- 使用软间隔，允许一部分样本点不满足SVM的约束条件。引入松弛变量ξ_i表征该样本不满足约束的程度，引入惩罚因子C。优化目标可以重写为：

  $\min\limits_{\omega,b,\xi_i} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^m \xi_i$

  $\mathbf{s.t.}\ y_i(\omega^T x_i + b) \geq 1 - \xi_i$

  $\xi_i \geq 0, \quad i = 1, 2, \cdots, m$

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?

a) 深度学习的关键是把背景、噪音变得越来越相对，把重要的特征突出出来。卷积神经网络要得到更加整体的特征，而不是强调某一个具体的特征。通过不断地卷积和池化，对局部的特征做平滑处理，使局部的特征变得很平均，因为弱信号的组合关系比某一具体的特征要更加重要。卷积神经网络抹平了局部信号，同时没有损失特异性，虽然降低某种具体信号的强度，但这些很弱的信号组合在一起是非常强的统计学信号

b) 我们认为，真正的特征不是它原始的一些数值和向量（比如图片中的像素不能作为特征），而是这些数值、向量在分解后的重新组合。这些分解产生最基础的二分类问题（概念上），它们的组合（结构上）才是真正能够描述数据的特征。线性回归没有这个分解和组合的过程，所以它利用的特征在概念上和结构上是没有意义的

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?

a) 两种特征选择方法比较：

- 核方法：把非线性问题映射到更高纬度，但是在求解的过程中绕开了求解从低维空间到高维特征空间的ϕ变换，用内积的形式$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$表示数据之间的相似性以及相对的距离关系。但是模型完全依赖于训练数据，存在过拟合的问题
- 神经网络方法：侧重于通过数据的原始表达的组合来提取数据之间的关系，通过同一空间中更复杂的高维非线性函数来表征非线性关系，本质上是参数学习问题。但是会出现局部解。

b) 使用复杂神经网络做非线性分类的难点在于：

- 由于问题非凸，且高阶导数非线性和奇异性，会存在很多局部解
- 基于马尔可夫假设，只利用了同层间的一些简单函数的组合，忽略了高维的不同层之间的非线性关系，表征能力受限

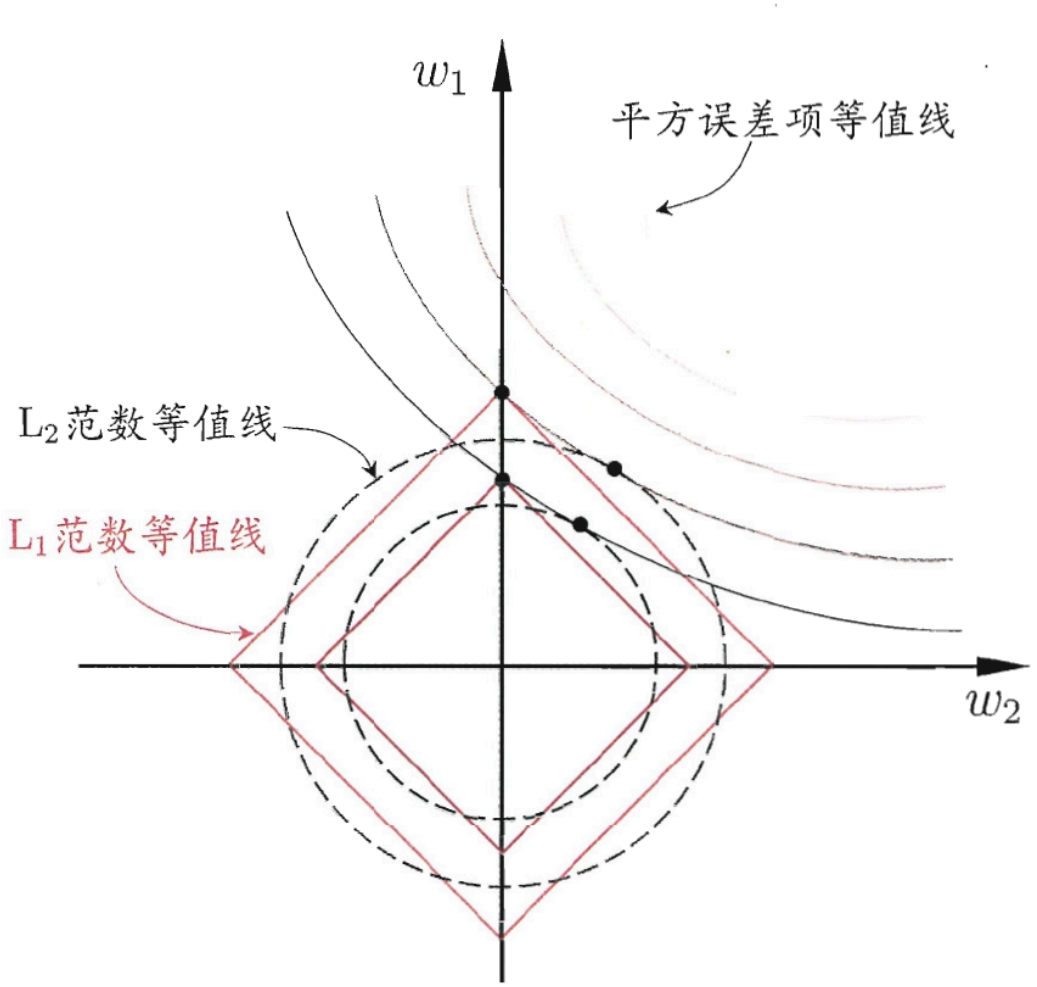
4) For any learning problems, (a) why a gradient-based search is much more favorable than other types of searches? (b) what would be the possible ramifications of having to impose some kinds of sequentiality in both providing data and observing results?

a) 我们碰到的大部分问题都是可微的，局部光滑的；也就是说使用梯度法的前提条件是可以满足的，而其他的搜索方法（剪枝，dfs，hash等）在复杂度上要比梯度高很多。梯度下降也保证了可行解的存在性，并且求得的解是合理的

b) 如果数据和观测值有强加的顺序，那么可能会导致模型不准确，比如模型可能会在开始时的数据中丢弃掉一些重要的特征（为了保证稀疏性），而这些特征对于后来的数据可能更加有用，但此时已无法找回被丢弃的特征。我们可以使用一些特殊的处理方法来缓解使用数据的问题，比如用batch的方法批处理数据，或者用交叉验证的方法组合使用观测数据

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?

a) 在线性回归中，我们使用平方误差，假定样本x仅有两个属性，即求出的w只有两个分量，可以绘制出如下的图像。



L1正则化项与误差项等值线交点更有可能出现在坐标轴上，因而有若干个纬度的值为0，起到了稀疏的效果。而L2正则化项与误差项等值线交点分布较为均匀，可以得到更为光滑的结果。

b) 要求学习是完整的，使用的数据有很好的独立性，数据的使用顺序有很好的随机性。由于L1是按照数据的使用顺序丢弃若干参数的，只有数据的使用顺序足够随机，才有可能L1丢弃的最终结果是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).

a) 监督学习与无监督学习最大的区别在于无监督学习没有label，只有一堆混合在一起的数据，通常需通过期望最大化或其他方法来进行聚类学习。同时，在做期望估计时需要穷举所有的可能性，即$P(Y|X)=\prod\limits_{Z}P(Y|Z,X)$，增大了问题的难度

b) 以EM算法为例，EM算法的优化函数是：
$\theta^{t+1}=\arg\max\limits_{\theta}Q(\theta|\theta^t)=\arg\max\limits_{\theta}\mathbb{E}_{Z|X,\theta^t}LL(\theta|X,Z)$，即优化的是对数似然函数关于$Z$的期望，这个函数非凸，因此不能保证收敛到全局最优解。另一方面，我们可以证明有如下不等式：
$\log p(X|\theta)\geq Q(\theta|\theta^t)+H(\theta|\theta^t)$，即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，无法保证实际目标函数的最大化。

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?

a) HMM的前向算法中利用了贝叶斯的先验知识来进行推断。

现假定有一个HMM模型$\theta = [A, B, \pi]$，其中，A表示在各状态间转换的概率，B表示根据当前状态获得各观测值的概率，π表示初始时刻各状态出现的概率。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$\alpha_i(t)$，记作：

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

我们可以递推地求得前向概率

$$\alpha_j(t) = b_j(o_t) \sum_{i=1}^{N-1} a_{ij}\alpha_i(t-1)$$

当有完整的观测序列O时，可以根据前向概率推测观测序列的概率
$P(O|\theta) = \alpha_N(T+1) = \sum_{i=1}^{N} \alpha_i(T)$

可以看出，我们计算观测序列概率时利用的是先验的概率，先验的先验的概率.....通过递归展开使得先验越来越清晰

b) 图模型的指向性很强，每一步的推导都依赖先验或后验知识，路径非常清晰，因此可能会造成先入为主的问题。在设计的时候希望图模型可以收敛，但是在最开始的时候先验的指向性并不是非常强，而是均匀的边缘分布，在之后的过程中边缘分布向着一个方向逐步收敛。这样使得过程出现较多的变化，有助于泛化。

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?

a)

- 先验的递归和收敛和图模型的结构有关系，比如如果有多个链汇聚到一个点的时候，在聚焦点会有个重要的更新，有一个去伪存真的过程。
- 在图模型中，先验的推断是一个不断迭代的过程，因此迭代的深度也是很重要的，如果深度过大的话可能需要很多的数据去进行监督。所以，模型和问题本身的复杂度应该相对应。

b) 先验的设计通常需要在开始的时候就相对比较清楚，所以关键是这个聚焦点能不能提高一些比较好的信息，以助于先验尽快收敛，把搜索局限在一个较小的范围。因此开始的时候似然归纳有助于确定先验，这样的归纳应该越清晰越早越好，可以尽早确定搜索方向。

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?

a) 主要采用的策略有：

- 将模型简化为原始形式：尽可能对原问题进行分解，将原问题分解为一系列的二分类问题，并将这些问题重新组合。好处是可以绕过模型的穷举，提高了模型的鲁棒性和泛化能力
- 通过分层来提取更多结构表征：对应于网络预训练步骤，通过构建并学习层状结构对数据进行分解，来学习如何表征数据本身具有的结构。经过这样的预处理后，模型会有更大的概率达到良好的效果
- 控制模型的复杂性：可以采用降维和正则化等方法，得到稀疏低秩的低复杂度解

b) 原因：

- 线性模型的复杂度较低，没有考虑数据之间的高阶相互作用，低复杂度模型往往具有更好的泛化能力
- 线性模型构造出的数据的边界也是线性的，这样的边界拓展性较好，鲁棒性更高

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?

a) 在逻辑回归中，令 $f_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$，假设事件符合**伯努利分布**，有 $P(y=1|x;\theta) = f_\theta(x), P(y=0|x;\theta) = 1 - f_\theta(x)$，因此可以推导出

$$p(y|x;\theta) = (f_\theta(x))^y (1 - f_\theta(x))^{1-y}$$

可以得到对数似然函数

$$l(\theta) = \log L(\theta)$$

$$= \log \prod_{i=1}^m p(y_i|x_i;\theta)$$

$$= \sum_{i=1}^m y_i \log f_\theta(x_i) + (1-y_i)\log(1 - f_\theta(x_i))$$

通过上述推导，可以看出逻辑回归与最大似然估计：

- 一致性：逻辑回归中最小化对数损失实际上优化目标是$\min -l(\theta)$，和最大似然估计中最大化对数似然函数的优化边界是一致的
- 差异性：
    - 物理含义不同：逻辑回归是条件概率的监督学习，来最小化损失函数；最大似然估计是联合概率分布的最大化
    - 稳定性不同：逻辑回归主要取决于两个标签之间的数据，受边界值的影响较大；最大似然估计用到了所有的数据，稳定性较好

b) 逻辑回归对于数据分布具有较强的普适性，逻辑回归仅仅要求数据的标签分布服从伯努利分布，并且每一标签下的数据服从指数族分布即可，而大部分混合模型都满足这个条件。