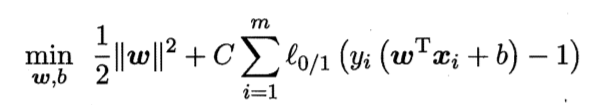
**Previous Exam**

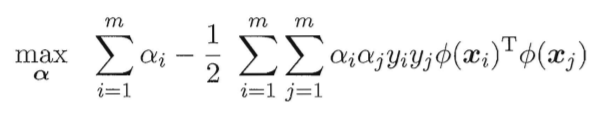
1. **SVM is a linear classifier with a number of possible risks to be incurred, particularly with very high dimensional and overlapping problems. Use a simple and formal mathematics to show and justify (a) how a margin-based liner classifier like SVM can be even more robust than Logistic regression? (b) how to control the overlapping boundary?**
2. 1. LR采用的是欧式测度，而SVM使用点积作为度量标准，由于数据本身是非欧式、dirty的，采取非欧式的测度可能更robust。

2. SVM损失函数中自带二范数，因此得到的结果维数也更低，模型复杂度更低，更robust。

3.SVM中只有在边界附近的数据（即支撑向量）才对边界有影响，不依赖于数据的所有分布。

1. 1.通过控制penalty前的系数控制边界的宽度：

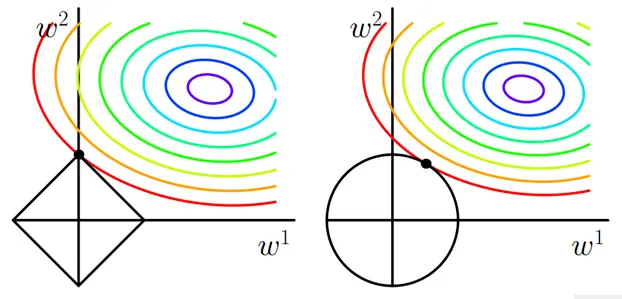
C越大则错误分类的惩罚越多，margin越宽，越深入数据中。

2.通过kernel，将数据非线性的映射到高维空间，期望数据在高维空间是可分的：

1. **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?**
2. **与传统的从背景中提取特征的方法不同，CNN学习到的是所关注特征与背景之间的关系，是对于pattern的学习，因此可能触及到更本质的问题。**
3. **Regression只是单纯的从各种特征中提取到对结果贡献大的那些，而这些特征本身可能是许多更本质的影响因素的组合，regression无法学习到这些背后的组合与模式，进行的特征提取是流于表面，缺乏解释性的。**
4. **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?**
5. **Kernel完全是面向数据的，不学习其背后的模式与本质，仅仅是希望能将数据投影后分开，减小overlapping，而且有overfitting的风险；NN是针对特征的组合与分解，学习feature出现的pattern。**
6. **这题没讲，自己想的**

* **如何设计神经网络的结构使得模型复杂度和数据的复杂度能够match**
* **局部解和overfitting**
* **网络过深导致的梯度消失**

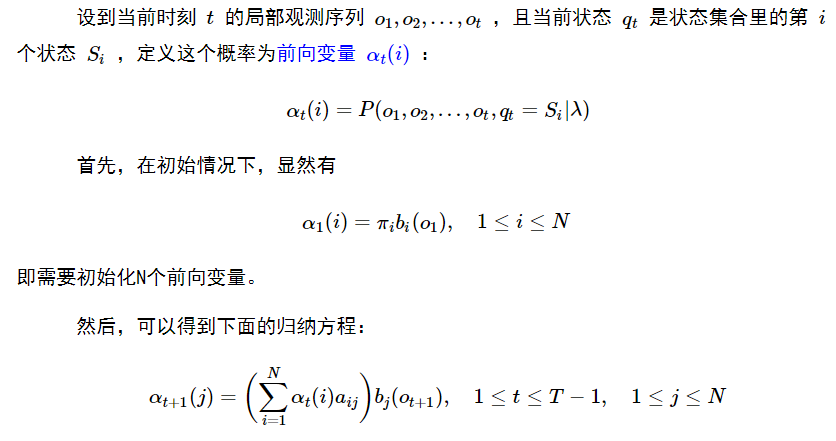
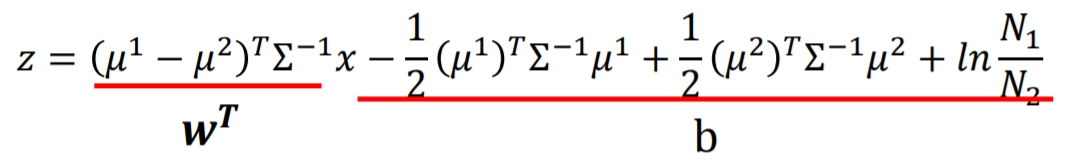
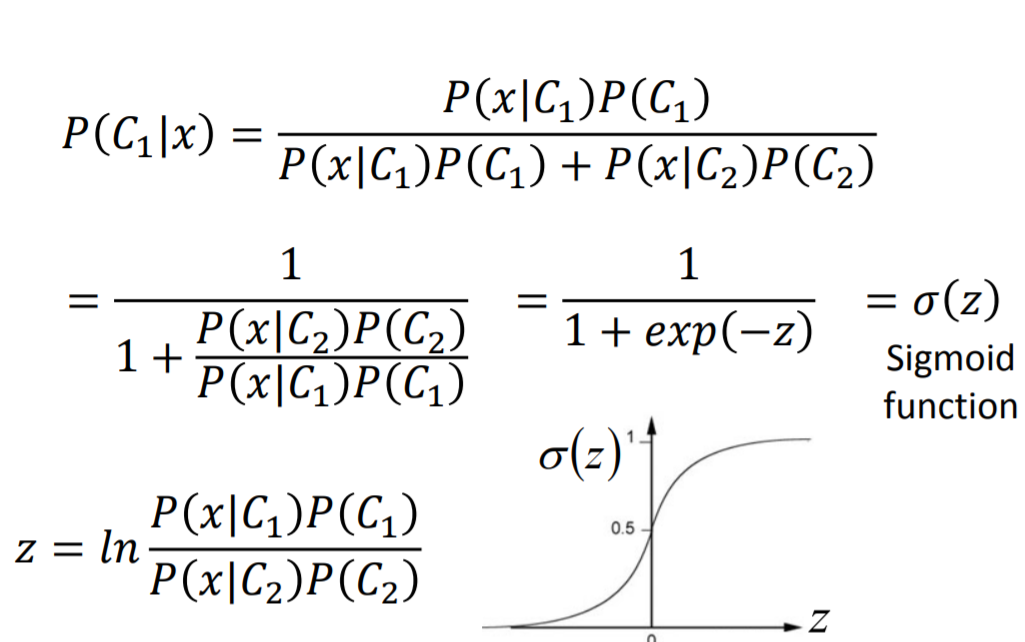
1. **For any learning problems, (a) why a gradient-based search is much more favorable than other types of searches? (b) what would be the possible ramifications of having to impose some kinds of sequential in both providing data and observing results?**
2. **Gradient提供了一条到solution的最短路径，当问题是凸的gradient可以确保到达全局最优解，而且其中的每一步都不会浪费。**
3. **由于现有的机器不可能实现真正的并发，模型学习的进度是与数据提供的顺序有很大关系的，而学习的过程又是不断丢弃参数使得模型简化稀疏的过程，有可能因为数据分布的缘故在一开始就丢弃了一些实际上很重要的参数。**
4. **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?**



如图圆圈表示原loss的等值线，黑色边界表示范数的约束范围，采用L1的情况交点更有可能在轴上，此时,对应的feature在最终模型中不起作用，稀疏化。

1. 0范数的最优化问题是一个NP hard问题，理论上有证明，L1范数是L0范数的最优凸近似。（我觉得）当数据本身比较冗余，彼此之间依赖关系比较多的时候L1会是比较好的近似。
2. **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).**
3. 无监督学习没有学习的目标label，因此要引入隐变量，要计算边缘概率：

无监督学习可以通过EM算法求解，不断对目标函数的lower bound进行近似，因此无法确定可以得到全局解。

1. 写出上式，先要求和再连乘，难以保证问题是凸的，想要得到global solution只能进行穷举。
2. **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?**
3. 是可以递归的，也就是说现在得到的后验概率可以作为下一刻的先验并不断传播。
4. 通过设计使得收敛得不至于过早，再观察到了足够的数据后再收敛，以此来增加泛化能力。
5. **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?**
6. 深度越深考虑到的情况就越多，观察的范围也越广，对先验、递归包括数据的支持要求也更高，才能开始收敛。缺点是当数据不足以支撑模型的时候先验比较混乱，始终没有出现占据主要地位的（同7）。
7. 不会、听不懂、编不出来
8. **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?**
9. 使用凸的模型；稀疏、低秩约束；降维；正交归一；kernel投影；DL将问题分解重构；
10. 线性模型是凸的，不会有局部解；模型涉及到的参数少，复杂度较低。
11. **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?**
12. LR对分布不做假设，仅仅是寻找boundary，属于判别性模型；mixture model先对分布进行假设，有了先验假设后看数据与先验的吻合程度，属于生成性模型。从后验概率的角度来看两者在形式上是统一的：

代入高斯分布可以得到：

不同的是LR的w和b仅仅与边界附近的数据有关，而mixture model的参数是由所有数据共同的分布计算得到的，因此mixture model更为稳定。

1. 可以证明只要是指数集族得到的混合模型都可以的到sigmoid的形式，上面就是利用高斯分布的例子；从直观上讲0-1决策是更多复杂模型的基础和本质。