第二周

注意：回顾题一般为课本上的题目，需要用自己的话进行总结打卡。补充题则为课本上没有另行补充的题目，会提供一些参考答案。

Day1-Day2

1. **回顾：请参考 svm 系列视频以及课本 3.1 节在白纸上完成硬间隔 svm 公式的推导。其中推导中可略过 SMO 算法。**

参考答案：参考讲解视频以及讲义进行推导。

2. **补充：为什么要将求解 SVM 的原始问题转换为其对偶问题？**

参考答案：首先明确一点，不转化为对偶问题也能求解。所以该问题可以从转化为对偶问题后带来的优点进行阐述。

引入对偶问题所带来的优势是:

（1） 对偶问题有时候更易求解，w、b 是维度相关的，而 lambda 是维度无关的，与样本数量相关，所以对高维数据且样本数量一定，适用于对偶问题。

（2） 对偶问题产生内积，方便核函数的引入，进而推广到非线性分类问题。

3. **补充：为什么 SVM 要引入核函数？**

参考答案：首先明确一点，核函数并非 SVM 独有，它是一种解决问题的方法。

当样本在原始空间线性不可分时，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。而引入这样的映射后，所要求解的对偶问题的求解中，无需求解真正的映射函数，而只需要知道其核函数。核函数的定义：$K(x,y)=<\phi(x),\phi(y)>$，即在特征空间的内积等于它们在原始样本空间中通过核函数 K 计算的结果。一方面数据变成了高维空间中线性可分的数据，另一方面不需要求解具体的映射函数，只需要给定具体的核函数即可，这样使得求解的难度大大降低。

4. **补充：为什么 SVM 对缺失数据敏感？哪些模型对缺失数据不那么敏感？（这里说的缺失数据是指缺失某些特征数据，向量数据不完整）**

参考答案：

SVM 没有处理缺失值的策略。而 SVM 希望样本在特征空间中线性可分，所以特征空间的好坏对 SVM 的性能很重要。缺失特征数据将影响训练结果的好坏。

对于缺失数据敏感和不敏感的面试题可以参考下面这么答:

树模型一般对于缺失值的敏感度较低，大部分时候可以在数据有缺失时使用。

一般涉及到距离度量时，如计算两个点之间的距离，缺失数据就变得比较重要。因为涉及到“距离”这个概念，那么缺失值处理不当就会导致效果很差，如 K 近邻算法(KNN)和支持向量机(SVM)。

线性模型的损失函数往往涉及到距离的计算，计算预测值和真实值之间的差别，这容易导致对缺失值敏感。

神经网络的鲁棒性强，对于缺失数据不是非常敏感，但一般没有那么多数据可供使用。

贝叶斯模型对于缺失数据也比较稳定，数据量很小的时候首推贝叶斯模型。
总结来看，对于有缺失值的数据在经过缺失值处理后：
数据量很小，用朴素贝叶斯
数据量适中或者较大，用树模型，优先 xgboost
数据量较大，也可以用神经网络
避免使用距离度量相关的模型，如 KNN 和 SVM

5. 补充：谈谈你是怎么使用 SVM 中的核函数的。
参考答案：
一般选择线性核和高斯核，也就是线性核与 RBF 核。 线性核：主要用于线性可分的情形，参数少，速度快，对于一般数据，分类效果已经很理想了。 RBF 核：主要用于线性不可分的情形，参数多，分类结果非常依赖于参数。有很多人是通过训练数据的交叉验证来寻找合适的参数，不过这个过程比较耗时。 如果 Feature 的数量很大，跟样本数量差不多，这时候选用线性核的 SVM。 如果 Feature 的数量比较小，样本数量一般，不算大也不算小，选用高斯核的 SVM。

6. 补充: SVM 的优缺点：
参考答案：
优点：
(1) 由于 SVM 是一个凸优化问题，所以求得的解一定是全局最优而不是局部最优。
(2) 不仅适用于线性问题还适用于非线性问题(用核技巧)。
(3) 拥有高维样本空间的数据也能用 SVM，这是因为数据集的复杂度只取决于支持向量而不是数据集的维度，这在某种意义上避免了“维数灾难”。
(4) 理论基础比较完善(例如神经网络就更像一个黑盒子)。
缺点：
(5) 二次规划问题求解将涉及 m 阶矩阵的计算(m 为样本的个数)，因此 SVM 不适用于超大数据集。(SMO 算法可以缓解这个问题)
(6) 只适用于二分类问题。(SVM 的推广 SVR 也适用于回归问题；可以通过多个 SVM 的组合来解决多分类问题)