期末考试
CS282 机器学习

2022 年 6 月 1 日

基于自己的理解，回答以下问题：

1.（15 分）什么是期望风险极小化，经验风险极小化，结构化风险极小化？期望风险和经验风险之间的关系是什么？

2.（10 分）什么是过拟合？如何能够检测到过拟合？如何来修正过拟合效应？

3.（10 分）正则化参数如何选取？现在有两种策略（已知正则化参数大约在 10^{-6} 到 1 之间）：（1）通过细密验证的选取，如每隔 10^{-6} 为一个间隔，选出表现最好的正则化参数。（2）通过粗略的选取，如从 $10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1$ 中间选取出表现最好的。你认为哪个策略好？原因是什么？

4.（6 分）比较 SVM、回归类（logistic regression）之间的优势（相较而言）。

5.（$3+3+4=10$ 分）如果现在考虑一个贷款用户的审批模型。变量 $x_1, ..., x_n$ 里面存在大量的（有序、无序）类别变量（例如：住宅年代、住宅邮编、学位学历）。也就是对于每一个 x_i，经过 dummy coding 可以形成扩充后的变量 $x_{ij}, j = 1, ..., i_k$。现在考虑以下场景，给出相应的正则项：

(1) 想知道哪个变量对于模型重要，而剔除掉不重要的变量。问应采取哪种正则项。

(2) 想知道每个变量里的类别哪些重要，哪些类别不重要，从而剔除不重要的类别。问应采取哪种正则项。

(3) 现在假定每个变量（年龄、收入、住房年限、负债等）都是由连续变量进行区间分划得到，例如住房的年限为连续变量，但采样得到的一般为类别 $[1920, 1930), [1930, 1940), ..., [2010, 2020)$。现在想知道每个变量内部这样的区间分划是否能够进一步整合（例如将 1920–1970 之间的区间统一整合为 [1920, 1970) 以降低未来问卷采样工作量）。问应采取哪种正则项。

6.（8 分）给映射 $K(x_1, x_2) = x_1^T x_2$，$\forall x_1, x_2 \in \mathbb{R}^m$，验证 K 是 Kernel。

7.（10 分）阐述对于 k-fold validation，样本容量为 $N = 1000$ 的情况。对于 $k = 2$ 和 $k = N$ 的两种选取方式，他们存在的优缺点是什么。

8.（6 分）说明 svm 为什么验证误差（01 泛化误差，线性可分的情况）只受到支撑向量（support vectors）的影响。

9.（$4 + 5 = 9$ 分）关于 VC 维度，完成以下两个证明：

（1） 考虑一个有限元素的假设函数集合 $\mathcal{H} = \{h_1, \ldots, h_M\}$。证明 $d_{vc}(\mathcal{H}) \leq \log_2 M$。

（2） 说明 d 维空间中，SVM 最优分离平面的 VC 维度 $d_{vc}(\text{SVM}) \leq d + 1$。

10.（$4 + 4 = 8$ 分）（1）利用 VC 不等式，讨论过高的模型复杂度为何会带来泛化性能变差。（2）利用方差偏差分解公式，讨论过高的模型复杂度为何会带来泛化性能变差。（考虑带噪音的情形，即 $y(x) = f(x) + \epsilon$，其中 ϵ 是期望为 0，方差为 σ^2 的噪音。可直接给出分解公式进行讨论，不需要证明之。）

11.（$4 + 4 = 8$ 分）对于 ℓ_2 正则形式的 logistic regression 问题（在目标中添加 ℓ_2 正则项，正则参数为正数），如果调用梯度下降法，回答以下三个问题：

(1) 算法如果用不同的初始点启动，是否会收敛到不同的最优解？为什么？

(2) 为什么在机器学习尤其是大规模数据的情况下，不建议用线搜索的方式（如 Backtracking-Armijo 步长）选取学习率。