《机器学习》期中考试试卷

1.请给出一般机器学习框架，基于该框架分别说明监督学习的训练和预测过程，并详细给出学过的分类算法以及列出至少**2**个实际应用例子。

一般机器学习框架：

1）数据收集

寻找具有代表性的公开数据集，若没有满足需求的数据集则可以自己采集数据。

2）数据预处理

- 真实的数据可能会有缺失、噪声和异常值等问题；所以需要对数据进行清理，处理缺失值，异常点等
- 如有需要，可将多个数据源中的数据集成到一个一致的数据存储中
- 根据需求可以对数据进行一些变换，例如标准化，维度变换等操作

3）模型选择和训练

- 根据任务目标、数据特征等选择具体模型，选择合适的方法对模型进行训练
- 训练过程中可以不断调整模型和训练过程

4）模型评估与优化

在测试集上计算一些评价模型的指标来评估模型；如果测试效果不好（例如过拟合等现象），可以分析问题所在，对模型进行优化并重新训练

监督学习的训练过程：

- 监督学习使用有真实标签的数据来训练模型
- 将训练数据输入模型，选择合适的损失函数计算模型输出的预测值和真实标签之间的误差
- 优化器根据这个误差来调整模型的参数，不断迭代至模型收敛或验证集准确率不再提高

监督学习的预测过程：

- 将预测数据输入模型，比较模型输出的预测值和真实标签
- 计算误差准确率等指标，对模型进行评估，效果不好时根据问题调整模型，重新进行训练

已学过的分类算法：

KNN，Logistic 回归，SVM，MLPNN等...

实际应用例子：

KNN进行鸢尾花分类；MLPNN进行手写数字识别等...

2.简述支持向量机的分类原理（包括映射函数和目标函数），说明支持向量机如何处理线性不可分以及非线性问题，并详细说明支持向量机的优缺点。

分类原理：

SVM基本思想是求解能够正确划分训练数据集并且几何间隔最大的超平面。即：使所有训练样本到该超平面的最小距离最大。对于线性可分的数据来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的超平面是唯一的。

映射函数：

在超平面的两边的样本为两个类，即映射函数为：

$$y = sgn(wx + b)$$

目标函数：

目标是求解几何间隔最大的超平面，可以表示为以下的约束最优化问题：

$$max \quad \gamma$$

$$s.t. \quad y_i(\frac{wx_i + b}{||w||}) \geq \gamma$$

通过推导，上述的约束问题等价于：

$$min \quad \frac{1}{2}||w||^2$$

$$s.t. \quad y_i(wx_i + b) \geq 1$$

用拉格朗日乘子法求其对偶问题，得到新的拉格朗日目标函数：

$$L = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i(y_i(wx_i + b) - 1)$$

为了得到对偶问题的具体形式，令 L 对 w 和 b 的偏导为0，得到：

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

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

将上述等式代入原式消去 w 和 b ，得到：

$$L = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

利用SMO算法求解，支持向量可以求解 b

线性不可分问题：

加入松弛变量和惩罚因子

非线性问题：

SVM引入核函数来解决非线性问题

- 核函数计算实例在非线性变换后的内积，对于非线性问题，可以通过非线性变换将它转化为某个特征空间中的线性问题，在高维特征空间中学习线性支持向量机
- 由于在线性支持向量机学习的对偶问题里，目标函数和分类决策函数都只涉及实例和实例之间的内积，所以不需要显式地指定非线性变换，而是用核函数替换当中的内积

SVM的优点：

- 核函数可以将数据映射到高维，解决非线性问题
- SVM是一种小样本学习方法
- 决策函数只取决于少数支持向量
- 简化了问题，同时剔除了大量样本使得模型具有更好的鲁棒性

SVM的缺点：

- 大规模数据训练困难，SVM求解二次规划涉及n阶矩阵的计算（n为样本数量），当n数目很大时计算复杂度高
- SVM是二分类器，不适用于多分类问题，需用多个二分类器才能进行多分类

3.请结合历史股价数据，给出至少两个可能的预测器，并详细阐述其工作原理。

股价数据是连续数值，可以采用回归模型进行预测，例如线性回归和多项式回归等

1）线性回归：

线性回归使用线性函数作为映射函数，输入特征，输出为连续的预测值，即：

$$y = wx + b$$

一般使用平方和误差作为目标函数，即：

$$loss = \frac{1}{2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

再利用最小二乘法或梯度下降法进行求解得到最终的模型

对于历史股价数据，训练一个线性回归模型，输入可以是有关股价市场的特征，输出预测股价。

为了防止过拟合，提高模型的泛化能力，可以引入正则化项，即在损失函数中加入参数的L1范数(Lasso回归)或L2范数(岭回归)

2）多项式回归：

相较于线性函数，多项式回归的映射函数不再是简单的线性函数，而是多项式函数，即：

$$y = w_0 + w_1 x + w_2 x^2 + \ldots + w_n x^n$$

多项式函数的表达能力比线性函数更强，也可以引入正则化项防止过拟合。目标函数和优化方法与线性回归类似

4.请基于手写体图像数据设计一个分类系统，包括数据采集，特征设计和模型设计等详细描述。

数据采集：

获取公开的MNIST数据集，例如可以从python的Keras模块或者利用pytorch来加载数据集

特征设计：

- 对数据进行清理，处理缺失值，异常点等
- 原始的图像数据为28×28的矩阵，将其转换为1×784的向量，以便后续处理
- 把数据缩放到[0,1]，可利用min-max 归一化实现。min-max 归一化在图像处理上非常常用，因为大部分的像素值范围是 [0, 255]
- 对数据进行PCA降维，减少计算量。在每个数字的图像中，图像边界部分的像素几乎都是白色的，把这些像素删除并不会对识别有什么影响

模型设计：

- 选择SVM模型进行分类任务
- 将训练数据输入模型，选择合适的损失函数计算模型输出的预测值和真实标签之间的误差
- 根据这个误差来调整模型的参数，不断迭代至模型收敛或验证集准确率不再提高
- 可比较使用不同核的SVM模型的性能。核函数可选用线性函数，poly函数，rbf函数或者sigmoid函数
- 将预测数据输入模型，比较模型输出的预测值和真实标签
- 计算误差准确率等指标，对模型进行评估，效果不好时根据问题调整模型，重新进行训练

5.请给出感知器，最小二乘法和Logistic Regression(LR)模型的映射函数以及对应的目标函数，并详细说明设计该目标函数的物理意义。

1）感知器：

映射函数为：

$$y = f(wx + b)$$

其中 f 为激活函数。分类任务目标函数为二分类交叉熵：

$$BCEloss = \frac{1}{n}\sum_{i=1}^{n}(-y_i log(\hat{y}_i) - (1 - y_i) log(1 - \hat{y}_i))$$

物理意义为：衡量预测值的概率分布与真实值的概率分布之间的差异

回归任务的目标函数选用平方和误差：

$$loss = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

物理意义为：衡量预测值和真实值之间的差异

2）最小二乘法：

映射函数为：

$$y = wx + b$$

目标函数为：

$$loss = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

物理意义为：衡量预测值与真实值之间的误差

3）LR：

映射函数为：

$$y = sigmoid(wx + b)$$

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

目标函数同非线性感知器，为二分类交叉熵(BCE)

物理意义同样为：衡量预测概率分布和实际概率分布之间的误差