PRML3

Zhu Yongshan

March 2025

1 理论题

1.1 试述什么是机器学习。

机器学习是计算机通过学习知识，不断学习、修正，以利用知识的过程。

1.2 试述监督学习的概念。其与非监督学习有何不同?

监督学习的数据是数据的特征-数据的标签的数据对，而非监督学习的数据一般没有数据的标签。

1.3 描述线性回归模型的数学形式。线性回归模型是如何进行预测的?

一元线性回归模型的数学形式可以表示为

$$y = \beta_0 + \beta_1 x \tag{1}$$

其中y是因变量，x是自变量，β_0是偏置，β_1是斜率

拓展到多元线性回归的情况

$$y = \theta_1 X + \theta_0 \tag{2}$$

其中θ_1是变量系数，θ_0是偏置

线性回归模型的预测先是利用训练数据通过最小二乘法等方法估计回归系数，再把新样本的自变量值带入模型计算得到因变量的预测值

1.4 定义线性回归的损失函数。为什么我们选择这种损失函数

线性回归常用的损失函数的是均方误差（MSE）损失函数，对于m个样本的数据集$(\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\})$，其均方误差损失函数为

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{3}$$

其中$(h_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$是线性回归模型的预测函数，$\theta = (\theta_0, \theta_1, \cdots, \theta_n)$是模型的参数，$x^{(i)}$是第$i$个样本的特征向量，$y^{(i)}$是第$i$个样本的真实值。

选择原因：可微，计算量小，可以有效避免较大的偏差出现

1.5 逻辑回归常用于分类任务。解释为什么逻辑回归能够用于分类，并描述它是如何做到的。

能用于分类的原因：通过逻辑函数（如 Sigmoid 函数）将线性回归输出映射到 [0, 1] 的概率值区间，概率值可衡量样本属于某类的可能性。

分类过程：先构建线性组合 $z = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$，再将 z 输入 Sigmoid 函数得到样本属于正类的概率 $\hat{p} = \sigma(z)$，最后根据阈值（通常 0.5）决策分类，$\hat{p} \geq 0.5$ 为正类，$\hat{p} < 0.5$ 为负类。

1.6 有哪些策略可以讲二类分类起组合构造为多类分类器？这些策略存在什么问题？各策略的计算复杂度是多少？除课堂讲解过的组合策略，是否存在其他组合策略？

策略：一对多（OvR）、一对一（OvO）、误差纠正输出码（ECOC）。

问题：一对多存在类别不平衡问题；一对一训练和预测的计算量都较大；误差纠正输出码编码设计和解码过程复杂。

计算复杂度：一对多训练 $O(Kn)$、预测 $O(K)$；一对一训练 $O(\frac{K(K-1)}{2}n)$、预测 $O(\frac{K(K-1)}{2})$；误差纠正输出码训练 $O(Ln)$、预测 $O(L(K+1))$。

其他策略：基于图的方法

1.7 什么是鉴别函数？用于分类的线性鉴别函数有什么优点和缺点？

1.8 鉴别函数定义

在模式识别和分类问题中，鉴别函数是一种用于对样本进行分类的函数。对于给定的样本特征向量，鉴别函数计算出一个值，根据这个值来判断样本属于哪个类别。例如，在两类分类问题中，可以通过比较两个类别的鉴别函数值大小来确定样本的类别归属；在多类分类问题中，可能会有多个鉴别函数，样本被分配到鉴别函数值最大的那个类别。

1.8.1 优缺点

优点：计算简单、可解释性强、训练速度快。

缺点：受线性可分假设限制、表达能力有限、对特征相关性敏感。

1.9 线性回归模型的误差来源有哪些？在实际构建模型分别应当采用什么策略来减小相应的误差？

1.10 误差来源

模型假设误差：线性回归模型假设因变量和自变量之间存在线性关系，可能不符合实际情况。

数据噪声误差：数据本身可能会因为各种原因存在噪声。

特征缺失误差：如果缺少了对因变量存在印象的变量，模型便无法充分捕捉数据的信息。

过拟合和欠拟合误差

1.11 策略

模型假设误差：使用非线性回归模型或对自变量变换。
数据噪声误差：数据平滑或增加数据量。
特征缺失误差：全面收集特征或进行特征工程。
过拟合和欠拟合误差：过拟合采用正则化和交叉验证；欠拟合增加模型复杂度。

1.12 什么是过拟合?

在模型参数量较大，训练集数据较少的情况下，模型训练过程不考虑数据真实分布情况下，最小化损失函数，导致在测试集上效果远差于训练集称为过拟合。

1.13 总结线性模型相关的英文术语

- Linear Regression（线性回归）
- Simple Linear Regression（简单线性回归）
- Multiple Linear Regression（多元线性回归）
- Regression Coefficient（回归系数）
- Intercept（截距）
- Independent Variable（自变量）
- Dependent Variable（因变量）
- Mean Squared Error (MSE)（均方误差）
- Gradient Descent（梯度下降）
- Least Squares Method（最小二乘法）
- Overfitting（过拟合）
- Underfitting（欠拟合）
- Regularization（正则化）

2 线性回归

令线性回归模型为

$$y = w^T x + b$$

2.1 使用损失函数为均方误差，推导出关于权重w的损失函数。

给定 m 个样本 $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$，预测值 $\hat{y}^{(i)} = w^T x^{(i)} + b$，真实值为 $y^{(i)}$。则均方误差损失函数 $L(w, b)$ 为:

$$\mathcal{L}(w,b) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2 = \frac{1}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b - y^{(i)})^2 \tag{4}$$

2.2 使用梯度下降法，推导出更新权重w和偏差b的公式。

$$w = w - \alpha\frac{2}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b - y^{(i)})x^{(i)} \tag{5}$$

$$b = b - \alpha\frac{2}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b - y^{(i)}) \tag{6}$$

2.3 假设采用l_2正则化项，给出加入正则化后的损失函数，并推导出权重更新公式。

如果加入正则化

$$\mathcal{L}_{reg} = \frac{1}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b - y^{(i)})^2 + \frac{\lambda}{m}w^T w \tag{7}$$

损失函数w求导

$$\Upsilon_w \mathcal{L}_{reg}(w,b) = \frac{2}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b` y^{(i)})x(i) + \frac{\lambda}{m}w \tag{8}$$

则权重w的更新公式为:

$$w = w - \alpha\left(\frac{2}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b - y^{(i)})x^{(i)} + \frac{\lambda}{m}w\right) \tag{9}$$

权重b的更新公式为:

$$b = b - \alpha\frac{2}{m}\sum_{i=1}^{m}(w^T x^{(i)} + b - y^{(i)}) \tag{10}$$

3 感知器

3.1 描述感知器的数学模型，并给出它的激活函数。

数学模型 给定输入向量$x = (x_1, x_2, \cdots, x_n)$，权重向量为$w = (w_1, w_2, \cdots, w_n)$，偏置为$b$。感知器每一层输出为

$$w^T + b = \sum_{i=1}^{n} w_i x_i + b \tag{11}$$

感知器常用的激活函数有

$$ReLU(z) = \begin{cases} x, & z \geq 0 \\ 0, & z < 0 \end{cases} \tag{12}$$

$$Sigmoid(z) = \frac{1}{1 + \exp(-z)} \tag{13}$$

3.2 用公式描述感知器的学习规则。当一个样本被错误分类时，权重是如何更新的?

学习规则：感知器的学习规则基于误差修正学习。设学习率为 η（$0 < \eta \leq 1$），对于一个样本 (x, y_{true})，其中 y_{true} 是真实标签，感知器的预测值为 $y_{pred} = u(w^T x + b)$。权重更新公式为 $w_i(t+1) = w_i(t) + \eta(y_{true} - y_{pred})x_i$，$i = 1, \cdots, n$，$b(t+1) = b(t) + \eta(y_{true} - y_{pred})$，其中 t 表示迭代次数。
错误分类时权重更新：当样本被错误分类时，即 $y_{true} \neq y_{pred}$。若 $y_{true} = 1$ 且 $y_{pred} = 0$（即 $w^T x + b < 0$），则 $w_i(t+1) = w_i(t) + \eta x_i$，$b(t+1) = b(t) + \eta$；若 $y_{true} = 0$ 且 $y_{pred} = 1$（即 $w^T x + b \geq 0$），则 $w_i(t+1) = w_i(t) - \eta x_i$，$b(t+1) = b(t) - \eta$。

3.3 感知器收敛性定理的内容是什么?

如果样本集是线性可分的，那么经过有限次的迭代，感知器的学习算法一定会收敛，即可以找到一个权重向量 w 和偏置 b，使得对于所有的样本，感知器都能正确分类。

4 逻辑回归

令逻辑回归模型为

$$h_w(x) = \frac{1}{1 + \exp(-w^T x)} \tag{14}$$

4.1 推导出交叉熵损失函数。

对于二分类逻辑回归，假设样本的真实标签 $y \in \{0, 1\}$，预测值 $\hat{y} = h_w(x)$。交叉熵损失函数 $L(w)$ 定义为单个样本损失的期望。单个样本的交叉熵损失为：

$$L_{single}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{15}$$

对于 m 个样本的数据集，交叉熵损失函数为：

$$L(w) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_w(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_w(x^{(i)}))] \tag{16}$$

4.2 使用梯度下降法，推导出更新权重w的公式。

首先求 $L(w)$ 关于 w 的梯度。对 $L(w)$ 中的每一项求导，根据复合函数求导法则，已知 $h_w(x) = \frac{1}{1+e^{-w^T x}}$，其导数为 $h_w(x)(1-h_w(x))$。根据梯度下降法 $w := w - \alpha \nabla_w L(w)$，权重 w 的更新公式为：$w := w + \frac{\alpha}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) x^{(i)}$

4.3 当引入l_1正则化项时，写出正则化后的损失函数。

ℓ_1 正则化项为 $\frac{\lambda}{m} \sum_{j=1}^{n} |w_j|$，其中 λ 是正则化参数，n 是权重 w 的维度。正则化后的损失函数$L_{reg}(w)$为：

$$L_{reg}(w) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_w(x^{(i)})) + (1-y^{(i)}) \log(1-h_w(x^{(i)}))] + \frac{\lambda}{m} \sum_{j=1}^{n} |w_j| \tag{17}$$

4.4 试给出线性模型参数优化的贝叶斯解释。

在贝叶斯框架下，线性模型的参数 w 被视为随机变量，具有先验分布 $P(w)$。给定数据集 $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$，根据贝叶斯定理，参数的后验分布为 $P(w|D) = \frac{P(D|w)P(w)}{P(D)}$。最大后验估计（MAP）是找到使后验概率 $P(w|D)$ 最大的参数 w。通常，$P(D)$ 是一个与 w 无关的归一化常数，所以等价于最大化 $P(D|w)P(w)$。$P(D|w)$ 是似然函数，在线性模型中可以基于数据的概率分布（如高斯分布假设下）来定义。$P(w)$ 是先验分布，例如选择高斯先验分布。ℓ_1 和 ℓ_2 正则化可以看作是在最大似然估计的基础上引入了不同的先验分布。ℓ_2 正则化对应高斯先验，ℓ_1 正则化对应拉普拉斯先验。通过最大化后验概率，实现对线性模型参数的优化，这就是线性模型参数优化的贝叶斯解释。

5 计算题

已知$w_i \in R, i = 0, \cdots, d, \mathbf{x} \in R^d$，求下列函数的偏导数$\frac{\partial f}{w_i}$或梯度$\nabla x f(x)$

5.1 $f(w_i) = \frac{\exp(w_i) - \exp(-w_i)}{\exp(w_i) + \exp(-w_i)}$

对w_i求导

令$u = exp(w_i)$, $v = exp(-w_i)$，则$f(w_i) = \frac{u-v}{u+v}$。

下面的平方，上导下不导减上不导下导的

$$\frac{\partial f}{\partial w_i} = \frac{(u+v)(u+v) - (u-v)(u-v)}{(u+v)^2} = \frac{4uv}{(u+v)^2} \tag{18}$$

代入得到

$$\frac{\partial f}{\partial w_i} = \frac{4\exp(w_i)\exp(-w_i)}{(\exp(w_i) + \exp(-w_i))^2} = \frac{4}{(\exp(w_i) + \exp(-w_i))^2} \tag{19}$$

5.2 $f(x) = \frac{1}{2}\mathbf{x}^T\mathbf{x}$

根据矩阵相乘规则,需要$\mathbf{x}$满足$\mathbf{x} = (x_1, x_2, \cdots, x_d)^T$，才能相乘
故而有

$$f(x) = \frac{1}{2}\sum_{i=1}^{d} x_i^2 \tag{20}$$

$f(x)$对x_iq求导，有

$$\frac{\partial f(x)}{\partial x_i} = xi \tag{21}$$

$$\nabla_x f(x) = (x_1, x_2, \cdots, x_d)^T = \mathbf{x} \tag{22}$$

5.3 $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x}$

设$x = (x_1, x_2, \cdots, x_d)^T$，$A = (a_{ij})_{d\times d}$，$b = (b_1, b_2, \cdots, b_d)^T$。
$x^T Ax = \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}x_i x_j$，则$\frac{1}{2}x^T Ax = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}x_i x_j$。
$\frac{\partial(\frac{1}{2}x^T Ax)}{\partial x_k} = \frac{1}{2}\sum_{i=1}^{d}(a_{ik}x_i + a_{ki}x_i)$（对$x_k$求偏导时，$x_i x_j$中只有$i = k$或$j = k$的项有非零导数）。$b^T x = \sum_{i=1}^{d} b_i x_i$，$\frac{\partial(b^T x)}{\partial x_k} = b_k$。所以

$$\nabla_x f(x) = \frac{1}{2}(A + A^T)x + b \tag{23}$$