机器学习作业 PRML-HW25F03

一、

1. 机器学习：

① 给计算机不用~~来看定程~~编程就能自动学习的领域。

② 使用经验Experience，表现在任务Task上，用评价指标评估 Performance Measure

③ 人工智能的分支，赋予计算机不通过显式编程即从数据中学习的能力。

2. 监督学习：

使用已知标签的数据进行训练。

非监督学习：训练数据没有标签。

3.

线性回归模型： $h_w(x) = \underline{\underline{w_0}} + \underline{\underline{w_1}} x$

目标：确定参数w.

↓ 方法

求损失函数 $J(w_0,w_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_w(x^{(i)}) - y^{(i)}\right)^2$ 的最小值 $\arg\min\limits_{w_0,w_1} J(w_0,w_1)$

确定参数后进行预测

4. 损失函数 $J(w_0,w_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_w(x^{(i)}) - y^{(i)}\right)^2$

① 使用m来求均方误差，得到平均平方误差，代表模型对每个样本的~~误~~模型误差。

② 再乘 $\frac{1}{2}$ 是为了简化后续的求~~导过程~~导的梯度计算过程。

5. 逻辑回归为什么能用于分类：通过 Sigmoid 函数将线性回归的连续结果映射到 (0, 1) 区间，将该区间值解释为"样本属于正类的概率"，再通过设定阈值将概率转化为离散的分类结果。

步骤：

① 构建线性预测函数

② 通过 sigmod 函数转换为概率

③ 设定阈值确定分类结果

6. 二分类器 $\xrightarrow{\text{组合构建}}$ 多分类器

↓

策略

是苹 / 蕉

苹 蕉 橙

① 一对一：(One-vs-One, OvO)

每类训练一个二分类器，共需 $C(N,2) = \frac{N \times (N-1)}{2 \times 1} = \frac{N(N-1)}{2}$

问题：分类器数量随类别数呈平方增长

复杂度：$O(N^2)$

② 一对多 (One-vs-Rest, OvR)

苹、香蕉+橙 ① → 是/不是苹

香 苹+橙 ②

橙 苹+香 ③

问题：类别数不平衡时，分类器偏向负类，预测准确率低

复杂度：$O(N)$

③ 多对多 (Many-vs-Many, MvM)

问题：编码规则设计复杂，且解释性差。

复杂度：$O(M)$ ← 编码长度

	第一位 ①	第二位 ②
苹果	1 (正)	1 (正)
香	1 (正)	0 (负)
橙	0 (负)	1 (正)

① → 0 / 1　　② → 0 / 1

其他：① 层次化分类　　e.g. 是否带皮吃 ……

② 集成式多类转换

7. 判别函数：用于分类任务的核心函数，通过计算样本的特征输入，输出一个用于判断样本所属类别的值。
优点：计算效率高，可解释性强，鲁棒性较好
缺点：仅适用于线性可分数据、特征表达能力弱、对类别不平衡的敏

8、线性回归模型：

	误差来源	减小策略
①	模型偏差(Bias)	提升模型复杂度、更换更灵活的模型
②	模型方差(Variance)	正则化、数据增强、简化模型
③	随机噪声(Noise)	数据清洗、特征工程

9. 过拟合：模型过度贴合训练数据的特征，导致其在训练集上误差极小，但在未见过的测试数据或新数据上泛化能力差、预测误差显著增高的现象。

10. Univariate Linear Regression 单变量线性回归
Supervised Learning 监督学习 Cost Function 损失函数
Gradient decent 梯度下降
Multivariate Linear Regression 多元线性回归
Features ~~偏差~~ 特征 Bias 偏差 Regularization 正则化
Overfitting 过拟合 Maximum likeihood 最大似然
least squares 最小二乘 Decision surface 决策面
Perceptron 感知机 Error Function 误差函数
Logistic Regression 逻辑回归

二、线性回归 $y=w^Tx+b$

1. Cost function $J(w,b)=\frac{1}{2m}\sum_{i=1}^{m}(h(x^{(i)})-y^{(i)})^2$

$$=\frac{1}{2m}\sum_{i=1}^{m}(w^Tx^{(i)}+b-y^{(i)})^2$$

2. $w^{(k+1)}=w^{(k)}-\frac{\partial J(w,b)}{\partial w}\cdot\alpha$

$$=w^{(k)}-\frac{1}{m}\sum_{i=1}^{m}(w^Tx^{(i)}+b-y^{(i)})\cdot x^{(i)}\cdot\alpha$$

$$b^{(k+1)}=b^{(k)}-\frac{\partial J(w,b)}{\partial b}\cdot\alpha$$

$$=b^{(k)}-\frac{1}{m}\sum_{i=1}^{m}(w^Tx^{(i)}+b-y^{(i)})\cdot\alpha$$

3. $J(w,b)=\frac{1}{2m}\left[\sum_{i=1}^{m}(\underbrace{h(x^{(i)})}_{w^Tx^{(i)}+b}-y^{(i)})^2+\lambda\sum_{j=1}^{n}w_j^2\right]$

$$w^{(k+1)}=w^{(k)}-\frac{\partial J(w,b)}{\partial w}\cdot\alpha$$

$$=w^{(k)}-\frac{1}{m}\left[\sum_{i=1}^{m}(\overset{w^Tx^{(i)}+b}{\cancel{h(x^{(i)})}}-y^{(i)})x^{(i)}+\lambda\sum_{j=1}^{n}w_j\right]\cdot\alpha$$

$$b^{(k+1)}=b^{(k)}-\frac{\partial J(w,b)}{\partial b}\cdot\alpha$$

$$=b^{(k)}-\frac{1}{m}\left[\sum_{i=1}^{m}(w^Tx^{(i)}+b-y^{(i)})\right]\cdot\alpha$$

三、感知器

1. $y(x)=f(\underbrace{w^T}_{\text{权重向量}}\underbrace{\varphi(x)}_{\text{特征向量}})$

$$f(a)=\begin{cases}+1, & a\geq 0\\ -1, & a<0\end{cases}$$

等价误分类点：$y_i(w^T x_i + b) \leq 0$

2、更新：$w^{(t+1)} = w^{(t)} - \eta \nabla E_p(w) = w^{(t)} + \eta y_n x_n$

3、感知器收敛定理

若训练数据集是线性可分的，则感知器学习算法经过有限次迭代后一定收敛，能找到一个可行的权重向量 w 和偏置 b，使得对于所有训练样本 (x_i, y_i)，都有 $y_i(w^T x_i + b) > 0$

四、逻辑回归

1、

$$J(w) = \frac{1}{m}\sum_{i=1}^{m} cost(h_w(x^{(i)}), t^{(i)})$$

$$cost(h_w(x), t) = \begin{cases} -\log h_w(x), & if\ t=1 \\ -\log(1-h_w(x)), & if\ t=0 \end{cases}$$

$$\Rightarrow cost(h_w(x), t) = -t\log h_w(x) - (1-t)\log(1-h_w(x))$$

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\left[t\log h_w(x) + (1-t)\log(1-h_w(x))\right]$$

2

$$w_j = w_j - \alpha\frac{\partial}{\partial w_j}J(w)$$

$$\left.\begin{aligned} h_w(x) &= g(w^T x) \\ g(z) &= \frac{1}{1+e^{-z}} \end{aligned}\right\} z = w^T x$$

$$\frac{\partial J}{\partial w} = -\frac{1}{m}\sum_{i=1}^{m}\left[t^{(i)}\frac{1}{h_w(x^{(i)})}\frac{\partial h_w(x^{(i)})}{\partial w} + (1-t^{(i)})\frac{1}{1-h_w(x^{(i)})}\cdot\frac{-\partial h_w(x^{(i)})}{\partial w}\right]$$

$$g'(z) = g(z)(1-g(z))$$

$$\therefore \frac{\partial h_w(x)}{\partial w} = \cancel{h_w(x)(1-h_w(x))} \quad \frac{\partial g(w^T x)}{\partial w}$$

$$= x\cdot g'(w^T x)$$

$$= x\cdot g(w^T x)[1-g(w^T x)]$$

$$= x\cdot h_w(x)\cdot[1-h_w(x)]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[t^{(i)}x^{(i)}[1-h_w(x^{(i)})] + (\cancel{1-t}\ t^{(i)})\cdot x^{(i)}h_w(x^{(i)})\right]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[t^{(i)}x^{(i)} - x^{(i)}h_w(x^{(i)})\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m}(h_w(x^{(i)}) - t^{(i)})x^{(i)}$$

3. $J(w) = -\frac{1}{m}\sum_{i=1}^{m}\left[t^{(i)}\log h_w(x^{(i)}) + (1-t^{(i)})\log(1-h_w(x^{(i)})) + \lambda\sum_{j=1}^{n} w_j\right]$

4. 通过观测数据 $D = \{(x_i, y_i)\}_{i=1}^{m} \xrightarrow{\text{推断}} P(w|D)$

$$P(w|D) = \frac{P(wD)}{P(D)} = \frac{P(w)P(D|w)}{P(D)}$$

$$P(w|D) \propto \underbrace{P(w)}_{\text{先验项}}\underbrace{P(D|w)}_{\text{似然项}}$$

寻找使 $P(w|D)$ 最大的 w.
(最大化后验概率)

先验项：反映对参数 w 的先验知识

似然项：$\prod_{i=1}^{m} P(y_i|x_i; w)$

五、

1. $f(w_i) = \frac{e^{w_i} - e^{-w_i}}{e^{w_i} + e^{-w_i}}$

$$\frac{\partial f}{\partial w_i} = \frac{(e^{w_i}+e^{-w_i})^2 - (e^{w_i}-e^{-w_i})^2}{(e^{w_i}+e^{-w_i})^2} = 1 - \left(\frac{e^{w_i}-e^{-w_i}}{e^{w_i}+e^{-w_i}}\right)^2$$

2. $f(x) = \frac{1}{2}x^T x = \frac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2)$

$$\nabla_x f(x) = \frac{\partial f}{\partial x} = x$$

$x = [\,]$

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

$$\nabla_x f(x) = \frac{\partial f}{\partial x} = Ax + b^T$$