Ch2机器学习概述

习题 2-1

分析为什么平方损失函数不适用于分类问题。

答：

平方损失函数 $L(y, f(x,\theta)) = \frac{1}{2}(y - f(x,\theta))^2$

1.从优化角度，使用平方损失函数训练分类模型，不能将损失函数最小化。设输入为X，一共有C类，分类模型输出的是X属于每一类别的概率p_i, $p_i \in [0,1]$。平方损失函数计算的是预测值$f(x,\theta)$与真实标签y之间的距离，如果使用在分类问题中，预测值是概率，$p_i \in [0,1]$，真实标签是具体的类别，难以最小化损失函数来优化模型。

2.从数据分布角度，假设偏差遵循正态分布，使用最大似然估计，MSE 正是用于优化模型的损失函数。而分类问题的输出模型是服从伯努利分布的。

*标签是离散的，不能用连续的函数

习题 2-2

在线性回归中，如果我们给每个样本 ($\boldsymbol{x}(n)$, $y(n)$) 赋予一个权重 $r(n)$，经验风险函数为

$$R(w) = \frac{1}{2}\sum_{n=1}^{N} r^{(n)}(y^n - w^T x^{(n)})^2, \quad (2.91)$$

计算其最优参数$\boldsymbol{w}^*$，并分析权重$r^{(n)}$的作用。

答：

样本数N 特征数量D。

$r = diag(r^{(n)}) \in R^{N\times N}$,$y = [y^{(1)}, \ldots, y^{(N)}]^T \in R^N$,是所有样本的真实标签组成的列向量，$X \in R^{(D+1)\times N}$ ，是由所有样本的输入特征$x^{(1)}, \ldots, x^{(N)}$组成的矩阵：

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ x_D^{(1)} & x_D^{(1)} & \cdots & x_D^{(N)} \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

于是，经验风险函数为

$$\begin{aligned} R(w) &= \frac{1}{2}\sum_{n=1}^{N} r^{(n)}(y^n - w^T x^{(n)})^2 \\ &= \frac{1}{2} r||y - X^T w||^2 \end{aligned}$$

风险函数$R(w)$是关于w的凸函数，其对w的偏导数为

$\frac{\partial R(w)}{\partial w} = \frac{1}{2}\frac{\partial r||y - X^T w||^2}{\partial w} = -Xr(y - X^T w)$

令$\frac{\partial}{\partial w}R(w) = 0$，得到的最优参数$w*$为

$$\begin{aligned} w* &= (XrX^T)^{-1}Xry \\ &= (\sum_{n=1}^{N} x^{(n)} r^{(n)} (x^{(n)})^T)^{-1}(\sum_{n=1}^{N} x^{(n)} r^{(n)} y^{(n)}) \end{aligned}$$

权重$r^{(n)}$是对不同的样本进行加权，调整各个样本对于结果的影响程度。如果有部分样本比较重要，需要关注的，可以将$r^{(n)}$设置得比较大，若部分样本属于噪声或者异常值等情况，希望减轻其对模型对影响，则设置得小些。

习题 2-3

证明在线性回归中，如果样本数量N小于特征数量D + 1，则XX^T的秩最大为N。

答：

样本数N 特征数量D。

$X \in R^{(D+1)\times N}$ ，是由所有样本的输入特征$x^{(1)}, \ldots, x^{(N)}$组成的矩阵：

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ x_D^{(1)} & x_D^{(1)} & \cdots & x_D^{(N)} \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

$R(X) \le min(N, D+1)$=>$R(X) \le N$,且$R(X) = R(X^T)$

因为$R(XX^T) \le min((R(X), R(X^T))$，所以$R(XX^T) \le N$

习题 2-4

在线性回归中，验证岭回归的解为结构风险最小化准则下的最小二乘法估计，见公式(2.44)

$$R(w) = \frac{1}{2}||y - X^T w||^2 + \frac{1}{2}\lambda||w||^2$$

答：

岭回归的解:$w^* = (XX^T + \lambda I)^{-1}Xy$

结构风险最小化准则下的最小二乘法估计的目标函数:$R(w) = \frac{1}{2}||y - X^T w||^2 + \frac{1}{2}\lambda||w||^2$

$\frac{\partial R(w)}{\partial w} = \frac{1}{2}\frac{\partial||y-X^T w||^2}{\partial w} + \frac{1}{2}\lambda\frac{\partial||w||^2}{\partial w}$

$= -X(y - X^T w) + \lambda w$

令$\frac{\partial R(w)}{\partial w} = 0$,则

$Xy = (XX^T + \lambda I)w$

$w^* = (XX^T + \lambda I)^{-1}Xy$，与岭回归的解相同，得证。

习题 2-5

在线性回归中，若假设标签y~$N(w^T x, \beta)$并用最大似然估计来优化参数，验证最优参数为公式(2.52)的解.

答:

y服从均值为$w^T x$，方差为β的高斯分布:

$p(y|X; w, \beta) == \frac{1}{\sqrt{2\pi\beta}}exp(-\frac{(y-w^T x)^2}{2\beta})$

参数w在训练集上对似然函数为

$p(y|X; w, \beta) = \prod_{n=1}^{N}(p(y^{(n)}|x^{(n)}; w, \beta) = \prod_{n=1}^{N} N(y^{(n)}; w^T x^{(n)}, \beta),$

其中，样本特征向量集为{$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$},对应对标签集为{$y^{(1)}, y^{(2)}, \ldots, y^{(n)}$}。令$X = [x^{(1)}, x^{(2)}, \ldots, x^{(n)}]$，$y = [y^{(1)}, y^{(2)}, \ldots, y^{(n)}]$。

则对数似然函数为

$logp(y|X; w, \sigma) = \sum_{n=1}^{N} logN(y^{(n)}; w^T x^{(n)}, \beta) = \sum_{n=1}^{N} log\frac{1}{\sqrt{2\pi\beta}} - \sum_{n=1}^{N}\frac{(y^{(n)}-w^T x^{(n)})^2}{2\beta}$

对w偏导

$\frac{\partial logp(y|X;w,\sigma)}{\partial w} = \frac{\partial}{\partial w}(\frac{1}{2\beta}||y - X^T w||^2) = \frac{1}{\beta}X(y - X^T w)$

令$\frac{\partial logp(y|X;w,\sigma)}{\partial w} = 0$，得 最优参数$w^{ML} = (XX^T)^{-1}Xy$

习题 2-6

假设有 N 个样本$x^{(1)}, x^{(2)}, \cdots, x^{(N)}$服从正态分布$\mathcal{N}(\mu, \sigma^2)$，其中 μ 未知.

1）使用最大似然估计来求解最优参数μ^{ML}；

2）若参数μ为随机变量，并服从正态分布$\mathcal{N}(\mu_0, \sigma_0^2)$，使用最大后验估计来求解最优参数$\mu^{MAP}$.

答:

1）参数w在训练集上对似然函数为

$p(x; \mu, \sigma^2) = \prod_{n=1}^{N}(x^{(n)}; \mu, \sigma^2),$

对数似然函数为

$log(x; \mu, \sigma^2) = \sum_{n=1}^{N} logp(x^{(n)}; \mu, \sigma^2) = \sum_{n=1}^{N} log\frac{1}{\sqrt{2\pi}\sigma} - \sum_{n=1}^{N}\frac{(x^{(n)}-\mu)^2}{2\sigma^2}$

令$\frac{\partial log(x;\mu,\sigma^2)}{\partial \mu} = 0$，$\frac{1}{\sigma^2}\sum_{n=1}^{N}(x^{(n)} - \mu) = 0,$

解得$\mu^{ML} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$

2）参数μ的后验分布为

$p(\mu|x; \mu_0, \alpha_0^2) \propto p(x|\mu; \sigma^2)p(\mu; \mu_0, \sigma_0^2)$

令似然函数$p(x|\mu; \sigma^2)$为高斯密度函数，则后验分布的对数为

$logp(\mu|x; \mu_0, \alpha_0^2) \propto logp(x|\mu; \sigma^2) + logp(\mu; \mu_0, \sigma_0^2)$

$\propto -\frac{1}{\sigma^2}(x^{(n)} - \mu)^2 - \frac{1}{\sigma_0^2}(\mu - \mu_0)^2$

令$\partial logp(\mu|x; \mu_0, \alpha_0^2)/\partial\mu = 0$，得到

$\mu^{MAP} = (\frac{1}{\sigma^2}\sum_{n=1}^{N} x^{(n)} + \frac{\mu_0}{\sigma_0^2})/(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2})$

习题 2-7

在习题2-6中，证明当$N \to \infty$时，最大后验估计趋向于最大似然估计.

答:

$\mu^{MAP} = (\frac{1}{\sigma^2}\sum_{n=1}^{N} x^{(n)} + \frac{\mu_0}{\sigma_0^2})/(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2})$

$= \frac{\sigma_0^2\sum_{n=1}^{N} x^{(n)} + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}$

$\lim_{N\to\infty}\mu^{MAP} = \lim_{N\to\infty}\frac{\sigma_0^2\sum_{n=1}^{N}\frac{x^{(n)}}{N} + \frac{\sigma^2\mu_0}{N}}{\sigma_0^2 + \frac{\sigma^2}{N}}$

当$N \to \infty$时，其他数都为常数，所以

$\lim_{N\to\infty} \mu^{MAP} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)} = \mu^{ML}$

习题 2-8

验证公式(2.61).

公式2.61 $f^*(x) = \mathbb{E}_{y\sim p_r(y|x)}[y]$

答：

以回归问题为例，假设样本的真实条件分布为$p_r(y|x)$，采用平方损失函数，$f^*(x)$为使用平方损失作为优化目标的最优模型，

模型$f(x)$的期望错误为

$$R(f) = \mathbb{E}_{(x,y)\sim p_r(x,y)}[(y - f(x))^2]$$

$$= \mathbb{E}_{(x,y)\sim p_r(x,y)}[(y - \mathbb{E}_{y\sim p_r(y|x)}[y] + \mathbb{E}_{y\sim p_r(y|x)}[y] - f(x))^2]$$

$$= \mathbb{E}_{(x,y)\sim p_r(x,y)}[(y - \mathbb{E}_{y\sim p_r(y|x)}[y])^2 + (\mathbb{E}_{y\sim p_r(y|x)}[y] - f(x))^2 + 2(y - \mathbb{E}_{y\sim p_r(y|x)}[y])(\mathbb{E}_{y\sim p_r(y|x)}[y] - f(x))]$$

其中，只有要学习的模型$f(x)$是变量，要使得$R(f)$最小， $f^*(x) = \mathbb{E}_{y\sim p_r(y|x)}[y]$ 。

习题 2-9

试分析什么因素会导致模型出现图2.6所示的高偏差和高方差情况。

答：

偏差是指一个模型在不同训练集上的平均性能和最优模型的差异。

方差是指一个模型在不同训练集上的差异，可以用来衡量一个模型是否容易过拟合。

方差一般会随着训练样本的增加而减少。而随着模型复杂度的增加，模型的拟合能力变强，偏差减少而方差增大，从而导致过拟合。

图2.6b为高偏差低方差的情况，表示模型的泛化能力很好，但拟合能力不足。图2.6c为低偏差高方差的情况，表示模型的拟合能力很好，但泛化能力比较差，当训练数据比较少时会导致过拟合。图2.6d为高偏差高方差的情况，表示模型选择及模型训练都不成功，此时训练出的模型效果最差。

高方差：数据集不够大

高偏差：选择的模型不合适，可能太过简单

习题2-10

验证公式(2.66).

答：

公式2.66如下

$$\mathbb{E}_D[(f_D(x) - f^*(x)^2)] = (\mathbb{E}_D[f_D(x)] - f^*(x))^2 + \mathbb{E}_D[(f_D(x) - \mathbb{E}_D[f_D(x)])^2]$$

证明如下：

$$\mathbb{E}_D[(f_D(x) - f^*(x)^2)] = \mathbb{E}_D[(f_D(x) - \mathbb{E}_D[f_D(x)] + \mathbb{E}_D[f_D(x)] - f^*(x))^2]$$

$$= \mathbb{E}_D[(f_D(x) - \mathbb{E}_D[f_D(x)])^2] + \mathbb{E}_D[(\mathbb{E}_D[f_D(x)] - f^*(x))^2]$$

$$+ \mathbb{E}_D[2(f_D(x) - \mathbb{E}_D[f_D(x)])(\mathbb{E}_D[f_D(x)] - f^*(x))]$$

因为E[E(X)]=E(X)，将$\mathbb{E}_D$移入括号内，

$\mathbb{E}_D[2(f_D(x) - \mathbb{E}_D[f_D(x)]) = 2(\mathbb{E}_D[f_D(x)] - \mathbb{E}_D[f_D(x)]) = 0$

所以$\mathbb{E}_D[2(f_D(x) - \mathbb{E}_D[f_D(x)])(\mathbb{E}_D[f_D(x)] - f^*(x))] = 0$

$$\mathbb{E}_D[(f_D(x) - f^*(x)^2)] = \mathbb{E}_D[(f_D(x) - \mathbb{E}_D[f_D(x)])^2] + \mathbb{E}_D[(\mathbb{E}_D[f_D(x)] - f^*(x))^2]$$

$$= \mathbb{E}_D[(f_D(x) - \mathbb{E}_D[f_D(x)])^2] + \mathbb{E}_D[(\mathbb{E}_D[f_D(x)] - f^*(x))^2]$$

$$= \mathbb{E}_D[(f_D(x) - \mathbb{E}_D[f_D(x)])^2] + (\mathbb{E}_D[f_D(x)] - f^*(x))^2$$

习题 2-11

分别用一元、二元和三元特征的词袋模型表示文本“我打了张三”和 “张三打了我”，并分析不同模型的优缺点。

答：

（1）一元特征表示有三个词："我"，"打了"，"张三"

对应的词袋模型表示为：

“我打了张三”:$[111]^T$；"张三打了我"：$[111]^T$

（2）二元特征表示有八个词："$我"，"我打了"，“打了张三”，"张三#"，"$张三"，"张三打了"，"打了我"，"我#"

对应词袋模型表示为：

“我打了张三”：$[11110000]^T$;"张三打了我"：$[00001111]^T$

（3）三元特征表示有六个词："$我打了"，"$张三打了"，"我打了张三"，"张三打了我"，"打了张三#"，"打了我#"

对应词袋模型表示为：

“我打了张三”:$[101010]^T$；"张三打了我":$[010101]^T$

一元特征的词袋模型：无法表示语序特征，单词相同，词向量就相同

二元特征的词袋模型：可表示单词相邻顺序，

三元特征的词袋模型：可表示单词前后两个相邻位置信息

若词袋容量太大则可以直接表示句子，失去词袋的意义。

习题2-12

对于一个三分类问题，数据集的真实标签和模型的预测标签如下：

真实标签 1 1 2 2 2 3 3 3 3

预测标签 1 2 2 2 3 3 3 1 2

分别计算模型的精确率、召回率、F1值以及它们的宏平均和微平均.

答：

混淆矩阵如下

真实类别\预测类别	3	2	1
3	2	1	1
2	1	2	0
1	0	1	1

各类别查准率：$P_3=\frac{2}{3}=66.7\%$ $P_2=\frac{2}{4}=50\%$ $P_1=\frac{1}{2}=50\%$

各类别查全率: $R_3=\frac{2}{4}=50\%$ $R_2=\frac{2}{3}=66.7\%$ $R_1=\frac{1}{2}=50\%$

各类别的F1值为：$F_3^1=\frac{2\times P_3\times R_3}{P_3+R_3}=\frac{2\times\frac{2}{3}\times\frac{1}{2}}{\frac{2}{3}+\frac{1}{2}}=\frac{4}{7}$

$$F_2^1=\frac{2\times P_2\times R_2}{P_2+R_2}=\frac{2\times\frac{1}{2}\times\frac{2}{3}}{\frac{1}{2}+\frac{2}{3}}=\frac{4}{7}$$

$$F_1^1=\frac{2\times P_1\times R_1}{P_1+R_1}=\frac{2\times\frac{1}{2}\times\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}}=\frac{1}{2}$$

宏平均：

$P_{macro}=\frac{1}{3}(\frac{1}{2}+\frac{1}{2}+\frac{2}{3})=\frac{5}{9}$

$R_{macro}=\frac{1}{3}(\frac{1}{2}+\frac{1}{2}+\frac{2}{3})=\frac{5}{9}$

$F_1^{macro}=\frac{2\times\frac{5}{9}\times\frac{5}{9}}{\frac{5}{9}+\frac{5}{9}}=\frac{5}{9}$

微平均：

$P_{macro}=\frac{\overline{TP}}{\overline{TP}+\overline{FP}}$

$P_{micro}=\frac{\overline{TP}}{\overline{TP}+\overline{FN}}$

$F_1^{micro}=\frac{2\times P_{macro}\times P_{micro}}{P_{macro}+P_{micro}}$

微平均中每个样本的查准率和查全率相同

每个样本的平均查准率=每个样本的平均召回率=$\frac{1+0+1+1+0+1+1+0+0}{9}=\frac{5}{9}$

所以$F_1^{micro}=\frac{2\times\frac{5}{9}\times\frac{5}{9}}{\frac{5}{9}+\frac{5}{9}}=\frac{5}{9}$