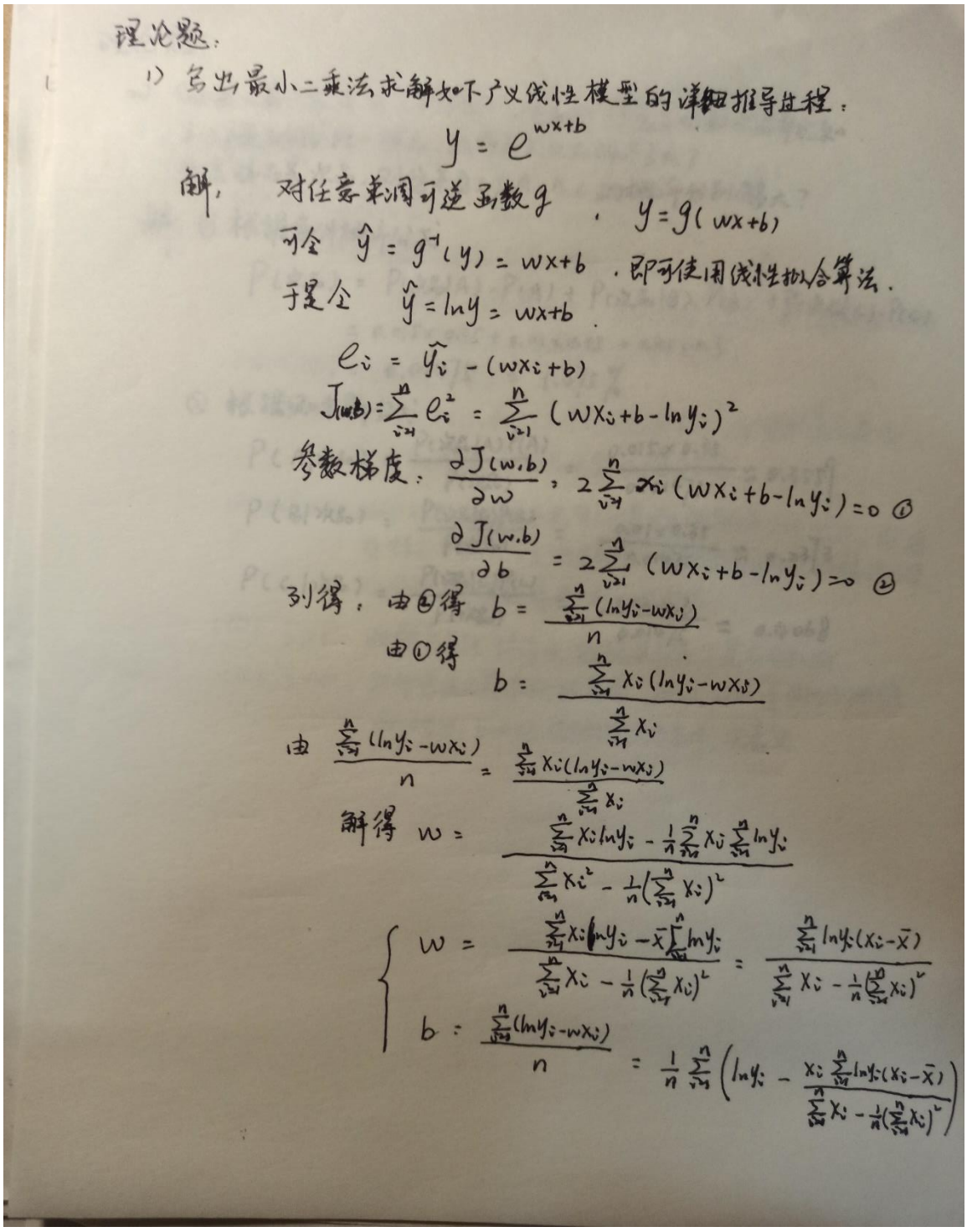
Homework1's Answer

第一次作业

一、理论题

1、



理论题：

1）写出最小二乘法求解如下广义线性模型的详细推导过程：

$$y = e^{wx+b}$$

解：对任意单调可逆函数 g，$y = g(wx+b)$

可令 $\hat{y} = g^{-1}(y) = wx+b$，即可使用线性拟合算法。

现令 $\hat{y} = \ln y = wx+b$

$$e_i = \hat{y}_i - (wx_i+b)$$

$$J(w,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (wx_i+b-\ln y_i)^2$$

参数梯度：$\frac{\partial J(w,b)}{\partial w} = 2\sum_{i=1}^{n} x_i(wx_i+b-\ln y_i) = 0$ ①

$$\frac{\partial J(w,b)}{\partial b} = 2\sum_{i=1}^{n} (wx_i+b-\ln y_i) = 0$$ ②

到得：由②得 $b = \frac{\sum_{i=1}^{n}(\ln y_i - wx_i)}{n}$

由①得 $b = \frac{\sum_{i=1}^{n} x_i(\ln y_i - wx_i)}{\sum_{i=1}^{n} x_i}$

由 $\frac{\sum_{i=1}^{n}(\ln y_i - wx_i)}{n} = \frac{\sum_{i=1}^{n} x_i(\ln y_i - wx_i)}{\sum_{i=1}^{n} x_i}$

解得 $w = \frac{\sum_{i=1}^{n} x_i \ln y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} \ln y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$

$$\begin{cases} w = \frac{\sum_{i=1}^{n} x_i \ln y_i - \bar{x}\sum_{i=1}^{n} \ln y_i}{\sum_{i=1}^{n} x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} = \frac{\sum_{i=1}^{n} \ln y_i (x_i - \bar{x})}{\sum_{i=1}^{n} x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \\ b = \frac{\sum_{i=1}^{n}(\ln y_i - wx_i)}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(\ln y_i - \frac{x_i \sum_{i=1}^{n} \ln y_i (x_i - \bar{x})}{\sum_{i=1}^{n} x_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}\right) \end{cases}$$

2、

理论题

2) 假设三家厂家A,B,C共同生产一种台灯，产品占比和次品率已知。
某次随机抽检一样品，该样品是次品概率多大？
若该样品是次品，则它来自工厂A,B,C的概率分别有多大？

解：① 根据条件概率公式

$$P(\text{次品}) = P(\text{次品}|A)\cdot P(A) + P(\text{次品}|B)\cdot P(B) + P(\text{次品}|C)\cdot P(C)$$
$$= 0.015\times 0.35 + 0.01\times 0.35 + 0.02\times 0.3$$
$$= 0.01475 = 1.475\%$$

② 根据贝叶斯公式

$$P(A|\text{次品}) = \frac{P(\text{次品}|A)P(A)}{P(\text{次品})} = \frac{0.015\times 0.35}{0.01475} \approx 0.3559$$
$$P(B|\text{次品}) = \frac{P(\text{次品}|B)P(B)}{P(\text{次品})} = \frac{0.01\times 0.35}{0.01475} \approx 0.2373$$
$$P(C|\text{次品}) = \frac{P(\text{次品}|C)P(C)}{P(\text{次品})} = \frac{0.02\times 0.3}{0.01475} \approx 0.4068$$

3、

理论题

3) 说明支持向量机中松弛变量与不同取值范围对应的含义。

解：原本我们对样本点的要求是 $y_i(w^Tx_i+b)\geq 1$

引入松弛变量后变为 $y_i(w^Tx_i+b)\geq 1-\xi_i$

优化问题的目标函数也变为 $\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$

（其中C为惩罚因子，当惩罚因子越大时，说明我们越不愿意放弃离群点（ξ_i对应点）。）

下面讨论ξ不同取值的含义：

(i) $\xi=0$：此时与未引入松弛变量的情况一致，即要求分类面最近的样本点，距离也要大于等于$\frac{1}{\|w\|}$。

(ii) $0<\xi<1$：此时允许有些样本点到分类面的距离小于$\frac{1}{\|w\|}$，即这些样本不满足分类内条件，但在放弃这些点的精确分类的同时，我们获得更大的几何间隔。

(iii) $\xi\geq 1$：此时会导致样本点的错误分类，是不可取的。

(iv) $\xi<0$：此时已经不算作松弛变量，而且可以适当比例放缩 $w\to 2w$，$b\to 2b$ 得到分类内条件，无意义。

二、实践题一

（所有的代码都集成到了 Final.py 这个文件中，运行这个文件即可获得所有实践题的结果）

1、实现 LDA 线性分类器并在西瓜 3.0 数据集上用前 80%训练、后 20%测试时的精度

解：

预处理：西瓜数据集 3.0 中同时有离散特征和连续特征，为了便于数据的处理，我使用 one-hot 的方法，将其特征向量扩展为了 19 维，其中离散特征用 1 表示该特征存在，用 0 表示该特征不存在。

训练：通过计算 Sw 和正类、负类分别的均值，运用公式可以计算出权值向量 w。

验证：通过权值向量 w 将测试样本投影到分类直线，通过比较投影点到两类样本哪类更近可以进行分类。在西瓜 3.0 数据集上用前 80%训练、后 20%测试时的精度如下：

```
The answer for Question1:
The accuracy for LDA(80% to 20% validation): 0.7
```

2、实现 Naïve Bayes 分类器并在西瓜 3.0 数据集上测试 K=5 重交叉验证精度

解：

训练： 首先计算先验概率，在分别计算离散特征的类分布概率和连续特征的类分布概率：其中，离散特征的类分布概率我是通过字典的方式存储的，连续特征的类分布概率则用高斯分布曲线来拟合。

验证：通过贝叶斯公式分别计算某个测试样本对于正类和负类的概率，哪

个概率大则分到哪一类。在西瓜 3.0 数据集上用 K=5 重交叉验证时的精度如下：

```
The answer for Question2:
The accuracy for NBC(cross validation): 0.6666666666666666
```

3、比较 SVM 使用不同（至少四种）核函数时，西瓜 3.0 数据集上用前 80%训练、后 20%测试的精度（可使用任意 SVM 算法实现软件包）

解：

预处理：西瓜数据集 3.0 中同时有离散特征和连续特征，为了便于数据的处理，我使用 one-hot 的方法，将其特征向量扩展为了 19 维，其中离散特征用 1 表示该特征存在，用 0 表示该特征不存在。

训练、预测：通过调用 sklearn 中的 SVM 包，分别运用 Linear 核、Polynomial 核、Gaussian 核、Sigmoid 核进行训练和预测。

验证：将预测标签与真实标签对比。在西瓜 3.0 数据集上用前 80%训练、后 20%测试时的精度如下：

```
The answer for Question3:
The accuracy of LinearMethod(80% to 20% validation): 0.6333333333333333
The accuracy of PolynomialMethod(80% to 20% validation): 0.7999999999999999
The accuracy of GaussianradialbasisMethod(80% to 20% validation): 0.7666666666666666
The accuracy of SigmoidMethod(80% to 20% validation): 0.6666666666666666
```

三、实践题二

实现对数几率回归并在西瓜 3.0 和 Iris 数据集上与线性分类器、Naïve Bayes 分类器和 SVM 做性能比较（5 折交叉验证）。

解：

1)对数几率回归的实现：

预处理：西瓜数据集 3.0 中同时有离散特征和连续特征，为了便于数据的处理，我使用 one-hot 的方法，将其特征向量扩展为了 19 维，其中离散特征用 1 表示该特征存在，用 0 表示该特征不存在。

训练：采用梯度下降法，控制迭代次数的上限，并设置当迭代前后 sigmoid 函数值差量小于一个极小量定值时退出迭代。迭代完成后，得到模型的权值向量 w 和偏置向量 b。

验证：通过权值向量 w 和偏置向量 b 以及 sigmoid 函数预测测试样本属于正类的概率，通过概率将测试样本分类。

2)分别在西瓜数据集 3.0 和 Iris 数据集上对比四种分类算法。

因为 Iris 数据集的分类属于三分类问题，所以在 LDA、LOG 和 SVM 中我运用了一对多(OVR)的策略。

在两个数据集上分别采用四种分类算法，运用 k=5 重交叉验证得到的精度如下：

```
The performance of NBC in Watermelon 3.0:
The accuracy for NBC(cross validation): 0.6

The performance of LDA in Watermelon 3.0:
The accuracy for LDA(cross validation): 0.6

The performance of LOG in Watermelon 3.0:
The accuracy for LOG(cross validation): 0.6666666666666667

The performance of SVM in Watermelon 3.0:
The accuracy of LinearMethod(cross validation): 0.5333333333333333
The accuracy of PolynomialMethod(cross validation): 0.8
The accuracy of GaussianradialbasisMethod(cross validation): 0.6666666666666666
The accuracy of SigmoidMethod(cross validation): 0.7333333333333333
```

```
The performance of NBC in Iris:
The accuracy for NBC(cross validation): 0.9517241379310345

The performance of LDA in Iris:
The accuracy for LDA(cross validation): 0.8137931034482758

The performance of LOG in Iris:
The accuracy for LOG(cross validation): 0.8344827586206897

The performance of SVM in Iris:
The accuracy of LinearMethod(cross validation): 0.7241379310344828
The accuracy of PolynomialMethod(cross validation): 0.9655172413793103
The accuracy of GaussianradialbasisMethod(cross validation): 0.9517241379310345
The accuracy of SigmoidMethod(cross validation): 0.3379310344827587
```

①可以看出在西瓜数据集 3.0 上各种分类算法精度普遍都低于在 Iris 数据集上的精度，是因为 Iris 数据集的样本数量远远高于西瓜数据集的样本数量。

②在西瓜数据集 3.0 上，四种分类算法的精度基本上比较接近。

③在 Iris 数据集上，Naïve Bayes 分类器表现明显好于 LDA 和对数几率回归分类器。而在 SVM 分类器中，多项式核和高斯核的 SVM 分类器明显好于线性核和 sigmoid 核。