Introduction to Machine Learning
Homework 1

191220103　孙逸扬　1196168404@qq.com

2021 年 10 月 7 日

1　[30pts] Basic concepts

1.1　Probabiliy

Suppose Bob has been tested with a terrible disease. The event T and D represent a person has been tested positive for this disease and actually has this disease, respectively. According to statistics, we know:

$$\begin{cases} \Pr(\mathrm{T}|\mathrm{D}) & = 0.98 \\ \Pr(\mathrm{T}|\neg\mathrm{D}) & = 0.10 \\ \Pr(\mathrm{D}) & = 0.01 \end{cases} \tag{1.1}$$

He wants you to help him calculate the probability that he actually has the disease?

解：按题意要计算 $Pr(D|T)$，根据贝叶斯公式

$$Pr(D|T) = \frac{Pr(T|D)Pr(D)}{Pr(T|D) + Pr(T|\neg D)}$$

可得 $Pr(D|T) = 0.98 \times 0.01/(0.98 + 0.10) \approx 0.0091$，所以 Bob 在检测为阳性的前提下得病的概率约为 0.0091。

1.2　Maximum likelihood estimation

We have an uneven coin, and the probability of tossing it heads up at random is p. Suppose you toss this coin 10 times, 8 of which are heads up. Please estimate p based on the existing information using MLE.

解：设 X 表示抛掷 10 次硬币正面朝上的次数，则根据题意

$$\begin{aligned} Pr(X=8) &= \mathrm{C}_{10}^{8} p^8 (1-p)^2 \\ &= 45 p^8 (1-p)^2 \end{aligned}$$

令 $f(p) = Pr(X=8)$ 且 $f'(p) = 0$，化简可得

$$4p^7(1-p)^2 = p^8(1-p) \Rightarrow 4(1-p) = p$$

所以 $f(p)$ 的极大值点为 $p = 4/5$，即 p 值为 0.8。

1.3 Performance meause

We have a set of samples with binary classes (denoted as 0 and 1) and two classifers C_1 and C_2. For each sample, the classifier gives a score to measure the confidence that the classifier believes that the sample belongs to class 1. Below are the predicted results of two classifiers (C_1 and C_2) for 8 samples, their ground truth labels (y), and the scores for both classifiers (y_{C_1} and y_{C_2}).

y	1	0	1	1	1	0	0	1
y_{C_1}	0.62	0.39	0.18	0.72	0.45	0.01	0.32	0.93
y_{C_2}	0.34	0.12	0.82	0.89	0.17	0.75	0.36	0.97

(1) Calculate the area under the ROC curve (AUROC) for both classifiers C_1 and C_2.

解：将 C_1 的所有预测值从高到低排列，依次划定正例阈值 t（$y_{C_1} \geq t$ 则判定为 1），并计算 FPR 和 TPR——

y	1	1	1	1	0	0	1	0
y_{C_1}	0.93	0.72	0.62	0.45	0.39	0.32	0.18	0.01
TP	1	2	3	4	4	4	5	5
TN	3	3	3	3	2	1	1	0
FP	0	0	0	0	1	2	2	3
FN	4	3	2	1	1	1	0	0
$FPR\ (x_i) = \frac{FP}{FP+TN}$	0	0	0	0	1/3	2/3	2/3	1
$TPR\ (y_i) = \frac{TP}{TP+FN}$	1/5	2/5	3/5	4/5	4/5	4/5	1	1

根据教材的 AUROC 估算公式

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1})$$

可得分类器 C_1 的 ROC 曲线下的面积为 $0 + \frac{1}{3}\times\frac{4}{5} + \frac{1}{3}\times\frac{4}{5} + \frac{1}{3} = 13/15$；同理可得分类器 C_2 的各项数据如下——

y	1	1	1	0	0	1	1	0
y_{C_2}	0.97	0.89	0.82	0.75	0.36	0.34	0.17	0.12
TP	1	2	3	3	3	4	5	5
TN	3	3	3	2	1	1	1	0
FP	0	0	0	1	2	2	2	3
FN	4	3	2	2	2	1	0	0
$FPR\ (x_i) = \frac{FP}{FP+TN}$	0	0	0	1/3	2/3	2/3	2/3	1
$TPR\ (y_i) = \frac{TP}{TP+FN}$	1/5	2/5	3/5	3/5	3/5	4/5	1	1

所以 C_2 的 AUROC$= \frac{1}{3}\times\frac{3}{5} + \frac{1}{3}\times\frac{3}{5} + \frac{1}{3} = 11/15$。

综上，C_1 的 AUROC$= 13/15$，C_2 的 AUROC$= 11/15$。

(2) For the classifier C_1, we select a decision threshold $th_1 = 0.40$ which means that C_1 classifies a sample as class 1, if its score $y_{C_1} > th_1$, otherwise it classifies this sample as class 0. Calculate the confusion matrix and the F_1 score. Do the same thing for the classifier C_2 using a threshold value $th_2 = 0.90$.

解：根据 $th_1 = 0.40$ 可得 C_1 的判定结果如下——

y	1	0	1	1	1	0	0	1
y_{C_1}	0.62	0.39	0.18	0.72	0.45	0.01	0.32	0.93
$label_{C_1}$	1	0	0	1	1	0	0	1

$TP = 4$，$TN = 3$，$FP = 0$，$FN = 1$，所以 $P = TP/(TP + FP) = 1$，$R = TP/(TP + FN) = 4/5$，从而 $F_1 = 2 \times P \times R/(P + R) = 8/9$。

再根据 $th_2 = 0.90$，可得 C_2 的判定结果如下——

y	1	0	1	1	1	0	0	1
y_{C_2}	0.34	0.12	0.82	0.89	0.17	0.75	0.36	0.97
$label_{C_2}$	0	0	0	0	0	0	0	1

$TP = 1,\ TN = 3,\ FP = 0,\ FN = 4$，所以 $P = 1,\ R = 1/5$，从而 $F_1 = 1/3$。
综上，C_1 的混淆矩阵为

真实标记	预测标记	
	0	1
0	3	0
1	1	4

C_1 的 F_1 score $= 8/9$；C_2 的混淆矩阵为

真实标记	预测标记	
	0	1
0	3	0
1	4	1

C_2 的 F_1 score $= 1/3$。

2 [30pts] Linear model

Suppose you are given a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. We want to use a regularized linear regression model to fit this data set, that is, to solve the following minimization problem:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 \tag{2.1}$$

where, $\mathbf{y} \in \mathbb{R}^m, \mathbf{X} \in R^{n\times d}$. Assume that $\mathbf{X}$ is column full-rank matrix.

1. Please give the closed-form solution for Eq.(2.1). You need to give your solution in detail.

   解：目标函数对 w 求偏导，令偏导数为 0 解出 w^*：

   $$\begin{aligned}\frac{\partial}{\partial w}\left(\frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_2^2\right) &= \frac{\partial}{\partial w}\left(\frac{1}{2}(y - Xw)^T(y - Xw) + \lambda w^T w\right)\\ &= \frac{1}{2}\cdot 2\cdot(-X^T)(y - Xw) + 2\lambda w\\ &= X^T(Xw - y) + 2\lambda w\\ &= 0\end{aligned}$$

   所以 $(X^TX + 2\lambda E)w = X^Ty$，即 $w^* = (X^TX + 2\lambda E)^{-1}X^Ty$。

2. The data set D is shown in the Table 1, where each sample has 3 dimensions (F_1, F_2, F_3). Please calculate the optimal solution for $\mathbf{w}$ when $\lambda = 1$.

   解：因为 $\lambda = 1$ 且由表 1 可知，

   $$y = (290,\ 1054,\ 944,\ 964,\ 246,\ 948,\ 488,\ 167,\ 370,\ 598)^T$$

   $$X^T = \begin{bmatrix} 2 & 9 & 8 & 8 & 2 & 8 & 4 & 1 & 3 & 5\\ 9 & 3 & 3 & 8 & 1 & 4 & 3 & 8 & 3 & 3\\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

   再由公式 $w^* = (X^TX + 2\lambda E)^{-1}X^Ty$ 计算可得

   $$w^* = (112.93,\ 6.19,\ 11.98)^T$$

F_1	2	9	8	8	2	8	4	1	3	5
F_2	9	3	3	8	1	4	3	8	3	3
F_3	1	1	1	1	1	1	1	1	1	1
y	290	1054	944	964	246	948	488	167	370	598

表 1: Training set for linear regression

3 [40pts] Logistic Regression

In a binary classification problem, each instance $\mathbf{x}_i$ in a data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ has a label $y_i \in \{0, 1\}$. We have already known that the logistic regression model Eq.(3.1) is a powerful tool to handle this kind of problem.

$$y = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}} \tag{3.1}$$

To simplify this problem, we assume that $\boldsymbol{\beta} = (\mathbf{w}; b)$, $\hat{\boldsymbol{x}}_{\boldsymbol{i}} = (\mathbf{x}_i; 1)$. Because its negative log-likelihood function Eq.(3.2) is convex, we can optimize it efficiently with Gradient Descent method, Newton method, and so on.

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(-y_i \boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_{\boldsymbol{i}} + \ln\left(1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_{\boldsymbol{i}}}\right)\right) \tag{3.2}$$

1. Prove the Eq.(3.2) is convex.

证明：根据凸函数的定义，只需证

$$\ell\left(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2}\right) \leq t\ell\left(\boldsymbol{\beta_1}\right) + (1-t)\ell\left(\boldsymbol{\beta_2}\right), \quad t \in [0,\ 1]$$

令 $f(\boldsymbol{\beta}) = \sum_{i=1}^n (-y_i \boldsymbol{\beta}^T \hat{\boldsymbol{x}}_{\boldsymbol{i}})$ 且 $g(\boldsymbol{\beta}) = \sum_{i=1}^n \ln\left(1 + e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_{\boldsymbol{i}}}\right)$，则 $\ell(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})$。一方面，

$$\begin{aligned} f\left(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2}\right) &= \sum_{i=1}^{n} \left(-y_i (t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2})^T \hat{\boldsymbol{x}}_{\boldsymbol{i}}\right) \\ &= t\sum_{i=1}^{n} (-y_i \boldsymbol{\beta_1}^T \hat{\boldsymbol{x}}_{\boldsymbol{i}}) + (1-t)\sum_{i=1}^{n} (-y_i \boldsymbol{\beta_2}^T \hat{\boldsymbol{x}}_{\boldsymbol{i}}) \quad = tf(\boldsymbol{\beta_1}) + (1-t)f(\boldsymbol{\beta_2}) \end{aligned}$$

另一方面，我们可以证明对 $x > 0,\ t \in [0,\ 1]$ 有如下关系成立——

$$(1+x)^t \geq 1 + x^t \tag{3.3}$$

当 $t = 0$ 或 $t = 1$ 时式(3.3)显然成立；当 $t \in (0,\ 1)$ 时，令 $h(x) = (1+x)^t - x^t - 1$，则

$$\begin{aligned} h'(x) &= t(1+x)^{t-1} - tx^{t-1} \\ &= t\left((1+x)^{t-1} - x^{t-1}\right) \\ &> 0 \end{aligned}$$

从而 $h(x)$ 在 $(0,\ +\infty)$ 单调增，于是 $h(x) > h(0) = 0$，式(3.3)成立。由此可知——

$$\begin{aligned}
tg(\boldsymbol{\beta_1}) + (1-t)g(\boldsymbol{\beta_2}) &= t\sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{\beta_1}^T \hat{\boldsymbol{x}}_i}\right) + (1-t)\sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{\beta_2}^T \hat{\boldsymbol{x}}_i}\right) \\
&= \sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{\beta_1}^T \hat{\boldsymbol{x}}_i}\right)^t + \sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{\beta_2}^T \hat{\boldsymbol{x}}_i}\right)^{(1-t)} \\
&\geq \sum_{i=1}^{n} \ln\left(1 + e^{t\boldsymbol{\beta_1}^T \hat{\boldsymbol{x}}_i}\right) + \sum_{i=1}^{n} \ln\left(1 + e^{(1-t)\boldsymbol{\beta_2}^T \hat{\boldsymbol{x}}_i}\right) \\
&= \sum_{i=1}^{n} \ln\left(\left(1 + e^{t\boldsymbol{\beta_1}^T \hat{\boldsymbol{x}}_i}\right) \cdot \left(1 + e^{(1-t)\boldsymbol{\beta_2}^T \hat{\boldsymbol{x}}_i}\right)\right) \\
&\geq \sum_{i=1}^{n} \ln\left(1 + e^{t\boldsymbol{\beta_1}^T \hat{\boldsymbol{x}}_i + (1-t)\boldsymbol{\beta_2}^T \hat{\boldsymbol{x}}_i}\right) \\
&= \sum_{i=1}^{n} \ln\left(1 + e^{(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2})^T \hat{\boldsymbol{x}}_i}\right) \\
&= g(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2})
\end{aligned}$$

综上，我们有

$$\begin{aligned}
\ell(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2}) &= f\left(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2}\right) + g\left(t\boldsymbol{\beta_1} + (1-t)\boldsymbol{\beta_2}\right) \\
&\leq tf(\boldsymbol{\beta_1}) + (1-t)f(\boldsymbol{\beta_2}) + tg(\boldsymbol{\beta_1}) + (1-t)g(\boldsymbol{\beta_2}) \\
&= t\ell\left(\boldsymbol{\beta_1}\right) + (1-t)\ell\left(\boldsymbol{\beta_2}\right)
\end{aligned}$$

即 $\ell\left(\boldsymbol{\beta}\right)$ 是凸函数，证毕。

2. Suppose we are facing a multi-class classification problem instead of a binary classification problem, where $y_i \in \{1, 2, \ldots, K\}$. Please expand the logistic regression model Eq.(3.1) to a multi-class version and give the log-likelihood function of this multi-class logistic regression model.

解：设样本 x_i 属于第 j 个类的概率为 p_j，则式(3.1)应当被扩展为

$$p_j = \frac{e^{\mathbf{w}_j^T x_i + b_j}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T x_i + b_k}}$$

该式就是 softmax，其最大似然为

$$\text{MLE} = \prod_{i=1}^{n} \frac{e^{\mathbf{w}_i^T x_i + b_i}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T x_i + b_k}}$$

从而其负对数似然为

$$
\begin{aligned}
L &= -\ln(\text{MLE}) = -\ln\left(\prod_{i=1}^{n}\frac{e^{\mathbf{w}_i^T x_i+b_i}}{\sum_{k=1}^{K}e^{\mathbf{w}_k^T x_i+b_k}}\right) \\
&= -\sum_{i=1}^{n}\ln\left(\frac{e^{\mathbf{w}_i^T x_i+b_i}}{\sum_{k=1}^{K}e^{\mathbf{w}_k^T x_i+b_k}}\right) \\
&= \sum_{i=1}^{n}\left(-\left(\mathbf{w}_i^T x_i+b_i\right)+\ln\left(\sum_{k=1}^{K}e^{\mathbf{w}_k^T x_i+b_k}\right)\right)
\end{aligned}
$$

3. Use out-of-the-box machine learning tools (e.g., scikit-learn, ...) to build your logistic regression model and comprehensively evaluate your results on Yeast[1] data set. You are recommended to try different techniques (e.g., OvO, OvR, multi-class logistic regression) for solving this multi-class problem. Briefly showing your analysis, experimental results, and conclusions.

解：本题使用 sklearn 分别对 OvO, OvR 以及多分类逻辑回归三种方法进行测试。所有方法都使用 2/3 的数据进行训练（以及验证）、1/3 的数据进行评估（数据是统一的，但是每次运行会随机划分）。原理就不再赘述，相关代码可以在附带的文件中找到[2]。

（1）OvO

在实验中，我观察到适当加大正则化系数可以提升准确率。这是因为二分类任务相对简单，很容易过拟合（可以输出每个二分类器在它自己任务上获得的准确率）。过拟合之后的二分类器彼此之间会互相冲突，从而导致最终结果趋于随机。或者更准确的说，每个二分类器都认为样本只有两个类，但这是不对的。一个过拟合的二分类器，会认为整个样本空间某一部分就是 0，另一部分就是 1，这样的判断会对最终结果产生影响，因为最后的分类是投票决定的。我们希望的是，每个二分类器在它有把握的时候以接近 1 的概率输出 1 或 0；在它没有把握的时候，以 0.5 的概率输出 0 或 1，这样不会影响其它分类器的投票结果。如果我们不加大正则项，将会观察到单个二分类器的准确率很高（接近 1.0），但在所有类上的整体正确率只有 0.1（接近随机，一共 10 个类）。

通过在验证集上调参（训练集进一步划分，用 1/5 做验证），最终确定正则化系数为 10（sklearn 中的 C 参数为 0.1）。在测试集上获得的准确率为 0.3 左右，这虽然不高，但是好于随机猜测（0.1）。

[1] http://archive.ics.uci.edu/ml/datasets/Yeast

[2] ./Prob01.ipynb

(2) OvR

根据 (1) 中的分析，可以预见到 OvR 的效果要比 OvO 好。因为 OvR 只判断某个样本属不属于该类；如果不属于，给其它所有类都投一票，这样不影响其它分类器的分类结果。不过实验中，我观察到 OvR 使用稍小一些的正则化系数表现更好。这可能是因为 OvR 是不平衡的，负类样本明显多于正类，如果正则化项太大，就会将分类器泛化到接近随机；换句话说，OvR 需要精准识别正类，因此它的正则化项不能太大。

通过在训练集上进行 3 折交叉验证，最终确定正则化系数为 0.05（sklearn 中的 C 参数为 20）。在测试集上取得 0.55 到 0.60 的准确率，这比 OvO 好得多。

(3) multiclass logistic regression

多分类逻辑回归最好先对数据进行归一化，归一化后的数据梯度比较稳定，有利于收敛（sklearn 的逻辑回归默认的优化方法是二阶的）。其它没有太多需要注意的，多分类回归对正则化项不敏感，但也不要过分加大正则化项，防止欠拟合。

使用 sklearn 默认的正则化系数（1.0），在测试集上也是取得 0.55 到 0.60 的准确率，和 OvR 差不多。

(4) 结论

最终结论是建议对多分类任务采用多分类逻辑回归，也即 softmax。假设一共有 C 个类，则 OvO 需要训练 $C(C-1)/2$ 个二分类器，这将消耗大量的时间和空间，并且 OvO 的准确率（泛化性能）没有其他两种方法好。OvR 的准确率虽然和多分类逻辑回归差不多，但是 OvR 要训练 C 个分类器，在时间和空间上都没有多分类的逻辑回归高效。