高级机器学习
作业一

MF1833040, 刘森, mf1833040@smail.nju.edu.cn

2018 年 12 月 11 日

1 [25pts] Multi-Class Logistic Regression

教材的章节3.3介绍了对数几率回归解决二分类问题的具体做法。假定现在的任务不再是二分类问题，而是多分类问题，其中$y \in \{1, 2 \ldots, K\}$。请将对数几率回归算法拓展到该多分类问题。

(1) **[15pts]** 给出该对率回归模型的“对数似然”(log-likelihood);

(2) **[5pts]** 计算出该“对数似然”的梯度。

提示1：假设该多分类问题满足如下$K-1$个对数几率，

$$\begin{aligned} \ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^{\mathrm{T}}\mathbf{x} + b_1 \\ \ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^{\mathrm{T}}\mathbf{x} + b_2 \\ &\cdots \\ \ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^{\mathrm{T}}\mathbf{x} + b_{K-1} \end{aligned}$$

提示2：定义指示函数$\mathbb{I}(\cdot)$，

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{若}y\text{等于}j \\ 0 & \text{若}y\text{不等于}j \end{cases}$$

Solution. 此处用于写解答*(中英文均可)*

解：

(1) 设$p(y=K|\mathbf{x})$为t

则对于任意$j \in \{1, 2 \ldots, K-1\}$我们有

$$p(y=j|\mathbf{x}) = te^{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}+b_j}$$

把$\mathbf{w}_j^{\mathrm{T}}\mathbf{x} + b_j$简写为$\beta_j^{\mathrm{T}}\hat{x}$,又因为

$$\sum_{j=1}^{K} p(y=j|\mathbf{x}) = 1$$

所以解得

$$t = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\beta_j^{\mathrm{T}} \hat{x}}}$$

同理可以解出任意的$p(y = j|\mathbf{x})$，$j \in \{1, 2 \ldots, K-1\}$
所以对于第i次的测试事件来说，

$$p(y_i|x_i; w, b) = \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) p_j(\hat{x}_i; \beta_j) + \mathbb{I}(y_i = K) p_K(\hat{x}_i; \beta_K)$$

所以对数似然为

$$l(w, b) = \sum_{i=1}^{m} lnp(y_i|x_i; w, b) = \sum_{i=1}^{m} ln(\sum_{j=1}^{K-1} \mathbb{I}(y_i = j) p_j(\hat{x}_i; \beta_j) + \mathbb{I}(y_i = K) p_K(\hat{x}_i; \beta_K))$$

即

$$l(\beta) = \sum_{i=1}^{m} (\sum_{j=1}^{K} \mathbb{I}(y_i = j) ln(p_j(\hat{x}_i; \beta_j)))$$

(2) 梯度$gradl(\beta_1, \beta_2 \ldots, \beta_{K-1}) = (\frac{\partial l}{\partial \beta_1}, \frac{\partial l}{\partial \beta_2} \ldots, \frac{\partial l}{\partial \beta_{K-1}})$，其中

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{m} \hat{x}_i (\mathbb{I}(y_i = 1) - p_1(\hat{x}_i, \beta_1))$$

$$\frac{\partial l}{\partial \beta_2} = \sum_{i=1}^{m} \hat{x}_i (\mathbb{I}(y_i = 2) - p_2(\hat{x}_i, \beta_2))$$

$$\ldots$$

$$\frac{\partial l}{\partial \beta_{K-1}} = \sum_{i=1}^{m} \hat{x}_i (\mathbb{I}(y_i = K-1) - p_{K-1}(\hat{x}_i, \beta_{K-1}))$$

2 [15pts] Semi-Supervised Learning

我们希望使用半监督学习的方法来对文本文档进行分类。假设我们使用二进制指示符的词袋模型描述各个文档，在这里，我们的词库有10000个单词，因此每个文档由长度为10000的二进制向量表示。

对于以下提出的分类器，说明其是否可以用于改进学习性能并提供简要说明。

1. **[5pts]** 使用EM的朴素贝叶斯；
2. **[5pts]** 使用协同训练的朴素贝叶斯；
3. **[5pts]** 使用特征选择的朴素贝叶斯；

Solution. *此处用于写解答(中英文均可)*

1. *使用EM的朴素贝叶斯:*
   可以改进，因为朴素贝叶斯假设属性之间相互独立这个假设在实际应用中往往是不成立的，在属性个数比较多或者属性之间相关性较大时，分类效果不好。如果采用EM算法进行训练，用已标记数据集作为初始数据集，来初始化一个朴素贝叶斯分类器，接着我们循环EM算法的E步骤和M步骤，当分类器趋于稳定的时候终止循环，最后输出分类器。因为EM算法本身是自收敛的，不需要事先设定类别，也不需要数据间的两两比较合并等操作，还可以使计算结果更稳定准确。

2. *使用协同训练的朴素贝叶斯:*
   可以改进，因为基于协同训练算法的复苏贝叶斯是输入：标记数据集L，未标记数据集U。用$L1$训练视图$X1$上的朴素贝叶斯分类器$f1$，用$L2$训练视图$X2$上的朴素贝叶斯分类器$f2$。用$f1$和$f2$分别对未标记数据U进行分类，把$f1$对U的分类结果中，前k个最置信的数据（正例p个反例n个）及其分类结果加入$L2$。把$f2$对U的分类结果中，前k个最置信的数据及其分类结果加入$L1$，把这2（$p+n$）个数据从U中移除，重复上述过程，直到U为空集。这样可以进行交叉验证，使结果趋于准确

3. *使用特征选择的朴素贝叶斯:*
   可以改进，因为在进行文本文档分类的时候，如果进行了特征选择，可以缩减朴素贝叶斯分类器的计算开销并且提升分类性能。即我们可以通过筛选掉一些不相关的特征来降维，来提升朴素贝叶斯分类效率和准确性。

3 [60pts] Dimensionality Reduction

请实现三种降维方法：PCA，SVD和ISOMAP，并在降维后的空间上用1-NN方法分类。

1. 数据：我们给出了两个数据集，都是二分类的数据。可以从`https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`找到，同时也可以在提交作业的目录文件夹中找名为“two datasets”的压缩文件下载使用。每个数据集都由训练集和测试集组成。

2. 格式：再每个数据集中，每一行表示一个带标记的样例，即每行最后一列表示对应样例的标记，其余列表示对应样例的特征。

具体任务描述如下：

1. **[20pts]** 请实现PCA完成降维（方法可在参考书`http://www.charuaggarwal.net/Data-Mining.htm` 中 Section 2.4.3.1 中找到）

   首先，仅使用训练数据学习投影矩阵；

   其次，用学得投影矩阵将训练数据与测试数据投影到 k-维空间 $(k = 10, 20, 30)$；

   最后，在降维后空间上用1-NN预测降维后k维数据对应的标记 $(k = 10, 20, 30)$，并汇报准确率。注意，测试数据集中的真实标记仅用来计算准确率。

2. **[20pts]** 请实现SVD完成降维（方法在上述参考书 Section 2.4.3.2 中找到）

   首先，仅使用训练数据学习投影矩阵；

   其次，用学得投影矩阵将训练数据与测试数据投影到 k-维空间 $(k = 10, 20, 30)$；

   最后，在降维后空间上用1-NN预测降维后k维数据对应的标记 $(k = 10, 20, 30)$，并汇报准确率。注意，测试数据集中的真实标记仅用来计算准确率。

3. **[20pts]** 请实现ISOMAP完成降维（方法在参考书 Section 3.2.1.7 中找到）

   首先，使用训练数据与测试数据学习投影矩阵。在这一步中，请用4-NN来构建权重图。（请注意此处4仅仅是用来举例的，可以使用其他 k-NN, $k \geq 4$并给出你选择的k。如果发现构建的权重图不连通，请查找可以解决该问题的方法并汇报你使用的方法）

   其次，用学得投影矩阵将训练数据与测试数据投影到 k-维空间 $(k = 10, 20, 30)$。

   最后，在降维后空间上用1-NN预测降维后k维数据对应的标记 $(k = 10, 20, 30)$，并汇报准确率。注意，测试数据集中的真实标记仅用来计算准确率。

可以使用已有的工具、库、函数等直接计算特征值和特征向量，执行矩阵的SVD分解，计算graph上两个节点之间的最短路。PCA/SVD/ISOMAP 和 1-NN 中的其他步骤必须由自己实现。

报告中需要包含三个方法的伪代码和最终的结果。最终结果请以表格形式呈现，表中包含三种方法在两个数据集中，不同 $k = 10, 20, 30$ 下的准确率。

Solution. *此处用于写解答(中英文均可)*

实验测得的准确率如下

Accuracy *Dim* / *Method*	$k=10$	$k=20$	$k=30$	*Dataset*
PCA	0.5825242718446602	0.5631067961165048	0.5631067961165048	*sonar*
PCA	0.7581609195402299	0.7627586206896552	0.7356321839080460	*splice*
SVD	0.5922330097087378	0.5825242718446602	0.5631067961165048	*sonar*
SVD	0.7586206896551724	0.7641379310344828	0.7480459770114942	*splice*
ISOMAP	0.4174757281553398	0.4174757281553398	0.4368932038834951	*sonar*
ISOMAP	0.6809195402298851	0.6901149425287356	0.6919540229885057	*splice*

*(1) PCA*算法伪代码：

输入 ： *train.txt, test.txt*

输出 *：* 准确率

1. 训练数据转换成矩阵并做去中心化处理
2. 计算协方差矩阵
3. 计算协方差矩阵的特征值和特征向量
4. 将特征值从大到小排序
5. 保留特征值较大的M个向量
6. 使用上述M个向量构建投影向量W
7. 分别把训练矩阵和测试矩阵用投影向量W投影到k维空间
8. 在降维后空间上用*1-NN*预测降维后k维数据对应的标记 *(k = 10,20,30)*，并汇报准确率

*(2) SVD*算法伪代码：

输入 ： *train.txt, test.txt*

输出 *：* 准确率

1. 输入训练数据矩阵$D(m,n)$，以及降维目标K
2. 进行*SVD*矩阵分解，得到$U(m,m),S(1,k),VT(n,n)$
3. 对S进行排序，取最大k个数据的索引
4. 分别取出VT中对应位置的列数据，拼接后转置成投影矩阵W

5. 分别把训练矩阵和测试矩阵用投影向量W投影到k维空间

6. 在降维后空间上用*1-NN*预测降维后k维数据对应的标记 *(k = 10,20,30)*，并汇报准确率

*(3) ISOMAP*算法伪代码：

输入 ： *train.txt, test.txt*

输出 *：* 准确率

1. 把训练数据矩阵和测试数据矩阵进行拼接得到矩阵D，目标降维维度k

2. 设定领域点个数，计算邻接距离矩阵，不在邻域之外的距离设置为无穷大

3. 求每对点之间的最小路径，将邻接矩阵转换成为最小路径矩阵

4. 把最小路径矩阵带入*MDS*算法中，具体操作如下

4. 计算出欧式距离$dist_{i\cdot}^2$,$dist_{\cdot j}^2$,$dist_{\cdot\cdot}^2$

5. 根据式$b_{ij} = -\frac{1}{2}(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2)$计算出矩阵$B$

6. 对B做特征值分解得到，取k个最大特征值所构成的对角矩阵$\tilde{\Lambda}$,$\tilde{V}$为对应的特征向量矩阵

7. 输出矩阵 $\tilde{\Lambda}\tilde{V}^{1/2}$，每一行是一个样本的低维坐标

8. 把输出矩阵拆分成训练和测试两块，用*1-NN*预测降维后k维数据对应的标记*(k = 10,20,30)*，并汇报准确率

*(4) ISOMAP*算法中注意点：

1. 在*sonar*中如果邻域点个数取*4*，则图不连通。逐次实验直到k=*6*时，可以图可以连通。

2. 在*splice*中邻域点个数取*4*，图是联通的，可以直接计算。