MS328 Assignment5

周李韬 518030910407

2020 年 5 月 12 日

找一个有意思的真实数据或者模拟一个高维数据，(基于 glment) 探索 LASSO 方法的算法、或参数选择或特征选择等。

建议选择一个角度深入探索 LASSO 方法。

1 实验准备

本实验中我们将生成高维数据，并使用 LASSO 方法进行特征提取，主要探索 LASSO 方法中参数的选择对结果的影响。

首先我们调用 sklearn 中 make_regression 方法生成 400 个 500 维的样本，其中有 10 个有用特征，并且将方差为 5 的高斯噪音作用于样本。

[1]:
```
import numpy as np
from sklearn.datasets import make_regression
reg_data, reg_target = make_regression(n_samples=400, n_features=500,␣
 ↪n_informative=10, noise=5)
print (reg_data.shape)
print (reg_target.shape)
```

```
(400, 500)
(400,)
```

2 LASSO 方法

针对 $Y = (Y_1, \ldots, Y_n)^\top$， $X = (X_1, \ldots, X_n)$，LASSO 问题为

$$\arg\min_{\beta \in \mathcal{R}^p} \frac{1}{2n} \left\| Y - X^\top \beta \right\|_2^2 + \alpha |\beta|_1$$

其中 l1-norm 的惩罚项保证了结果能够较大可能落在坐标轴上，从而达到降维提取特征的目的。针对这一问题，Python 提供了相关的算法模块，并且可以调节 Lasso 优化目标中的惩罚系数 α。本试验将直接调用该模块。

[2]:
```
from sklearn.linear_model import Lasso
lasso = Lasso()
lasso.fit(reg_data, reg_target)
```

[2]:
```
Lasso(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=1000,
      normalize=False, positive=False, precompute=False, random_state=None,
      selection='cyclic', tol=0.0001, warm_start=False)
```

默认情况下 Lasso 问题的系数为 1。计算得到如下非零特征数，可见 Lasso 算法达到了有效的降维效果。

[3]:
```
import numpy as np
np.sum(lasso.coef_ != 0)
```

[3]:
```
10
```

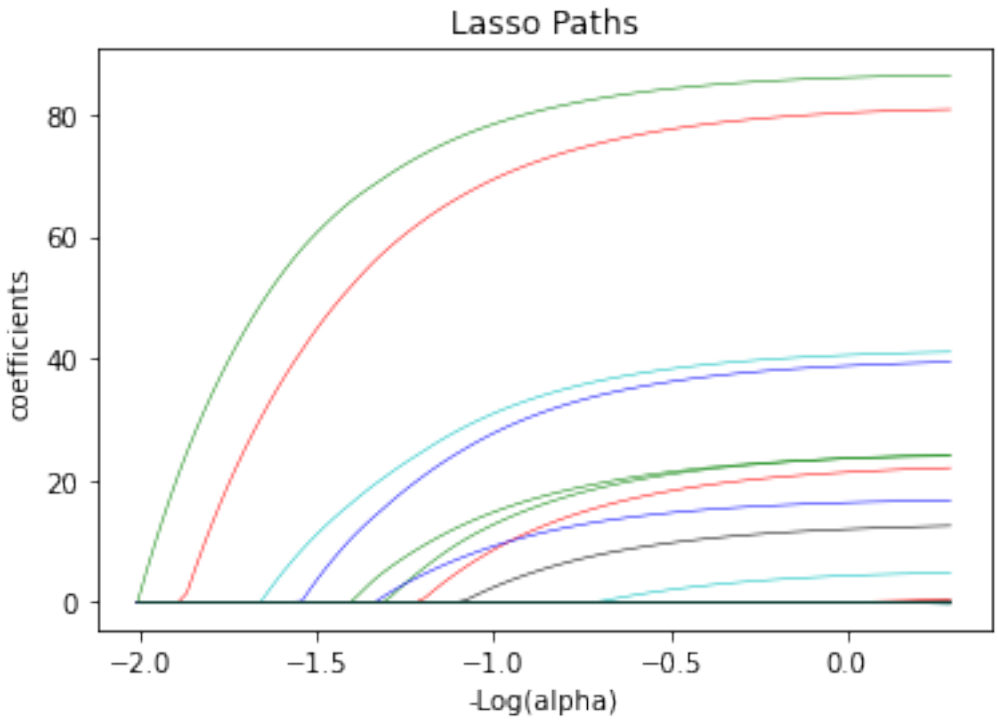
3 实验分析

我们可以调用 Python 中的 lasso_path 方法计算参数大小对特征系数的影响。该方法可以展示在当前数据集下，随着 alpha 的改变，各特征系数大小的变化情况。

[4]:
```
%matplotlib inline
from itertools import cycle
import numpy as np
import matplotlib.pyplot as plt

from sklearn.linear_model import lasso_path
from sklearn import datasets

# Compute paths

eps = 5e-3  # the smaller it is the longer is the path
```

```
alphas_lasso, coefs_lasso, _ = lasso_path(reg_data, reg_target, eps,␣
 ↪fit_intercept=False)

# Display results
plt.figure(1)
colors = cycle(['b', 'r', 'g', 'c', 'k'])
neg_log_alphas_lasso = -np.log10(alphas_lasso)
for coef_l, c in zip(coefs_lasso, colors):
    l1 = plt.plot(neg_log_alphas_lasso, coef_l, c=c,  linewidth=0.5)

plt.xlabel('-Log(alpha)')
plt.ylabel('coefficients')
plt.title('Lasso Paths')
plt.axis('tight')
```

[4]: (-2.123141518845042, 0.4079914763853374, -4.613041187632075, 90.80173345049359)



我们看到，随着惩罚值的减小，非零特征系数接连出现，并且不断增长。在本实验生成的数据集中，

由于 make_regression() 函数生成了大量无关特征，因此这里对于更多的特征，LASSO 方法不会进行选取。