计算机科学与技术学院

本科生《机器学习》课程作业

第二次上机报告

学　　号 **22281052**

姓　　名 张鳃沣

班　　级 **AI2201**

日　　期 **2024 年 12 月 16 日**

《机器学习》实验报告

第二次大作业 实验报告

张鲴沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

任务解决：基因变异分类

任务数据集下载地址：https://www.kaggle.com/datasets/kevinarvai/clinvar-conflicting

任务目的

本次任务的目的是利用 ClinVar 数据集，**通过 LASSO 回归建模，分析基因变异特征与其临床后果之间的关系。**我们旨在识别与基因变异临床后果显著相关的特征，并量化这些特征对临床后果的影响。通过 LASSO 回归的特征选择能力，我们能够筛选出最重要的基因特征，同时去除冗余和不相关的特征，从而提高模型的解释性和预测性能。该分析有助于临床医生和研究人员更好地理解哪些基因特征对遗传疾病的临床后果具有重要影响，为遗传变异的临床解读提供有价值的参考。

任务要求

对**ClinVar基因变异数据集**进行分析，对**基因变异特征与临床后果之间的关系进行建模，筛选出显著特征，评估模型性能，助力临床解读。**

需使用**至少两种优化方法**（如坐标梯度下降、近端梯度下降等）实现“**最小二乘法+LASSO**”的求解（建议自己至少独立实现一种方法），并用其解决一个实际问题（应用于实际任务数据集）。

图1 ClinVar 十周年时间线

上图展示了 **ClinVar 数据库**在2013年至2019年间的主要发展历程。ClinVar 是由 美国国家医学图书馆 (National Library of Medicine, NIH) 运营的数据库，提供有关人类基因变异及其临床意义的信息。ClinVar 在这十年内从最初的 27,000个变异增长到 500,000个变异，并提供了更便捷的提交工具、更频繁的数据更新和更详细的变异数据格式，为临床和研究人员提供了更强大的基因变异数据资源。

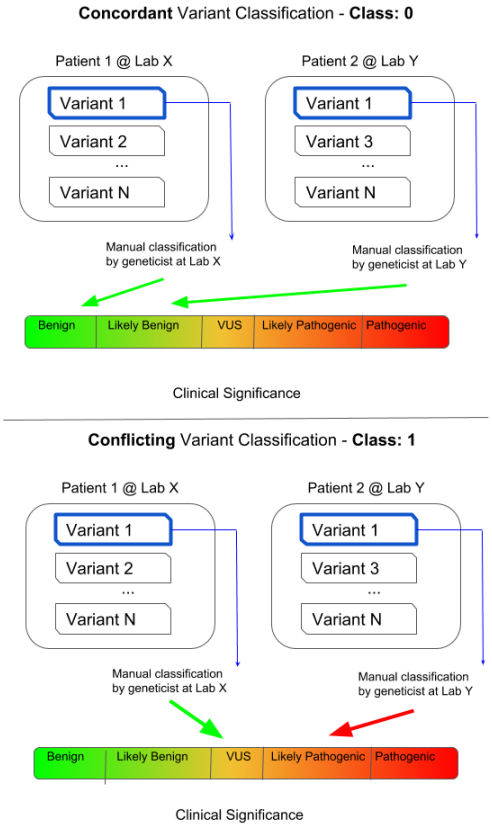


图2 基因变异分类

基因分类案例

1. Concordant Variant Classification - Class: 0

Variant 1（变异1）在两个不同的实验室（Lab X 和 Lab Y）被检测到。

- **Lab X** 和 **Lab Y** 的遗传学家分别对 **Variant 1** 进行了手动分类。
- 这两个实验室对 **Variant 1** 的分类是一致的（都指向绿色或浅绿色区域，如 **Benign（良性）** 或 **Likely Benign（可能良性）**）。

因为分类结果一致，这种情况被标记为 **Class: 0**（一致分类）。

2. Conflicting Variant Classification - Class: 1

在下半部分中，**Variant 1** 同样在两个不同的实验室（Lab X 和 Lab Y）被检测到。

- **Lab X** 的遗传学家对 **Variant 1** 进行了手动分类，结果为 **Benign（良性）**（绿色箭头所示）。
- **Lab Y** 的遗传学家对同一 **Variant 1** 进行了手动分类，但结果为 **Likely Pathogenic（可能致病）**（红色箭头所示）。

由于两个实验室对同一变异的分类不同，这种情况被标记为 **Class: 1**（冲突分类）。

这种分类冲突可能对临床医生和研究人员在解释患者的疾病风险时产生困惑和挑战。

张鲴沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第一部分　数据清洗

数据清洗流程

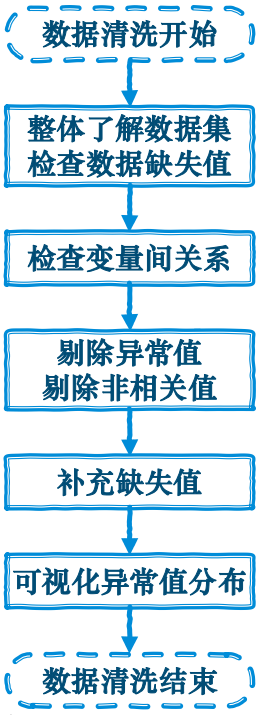


图3 数据预处理流程

1. 整体了解数据集和检查数据缺失值

使用 df.head() 查看前几行数据，df.info() 检查数据类型和记录数，**快速了解数据结构和每列的数据类型**。通过 df.isnull().sum() 统计每列缺失值，识别缺失数据的分布情况，为后续清洗提供依据。**如果缺失值较多，可以考虑删除相应列或行，或者在后续步骤中进行填补，以确保数据的完整性。**

2. 检查变量间关系

通过 df.describe() 获取数值特征的统计信息，如均值、方差、最小值和最大值，初步了解数据分布情况。利用 df.corr() 计算特征之间的相关性矩阵，**通过 sns.heatmap() 进行热图可视化，识别强相关或弱相关的特征。**这有助于特征选择和降维，剔除冗余信息，优化数据集结构，为建模提供更有效的特征。

3. 剔除异常值和非相关值

通过统计方法与实际案例关系，分析数据频率情况，剔除偏离正常分布的极端值、缺失值过多的数据，以减少其对模型的影响并删除非相关或冗余特征，提升数据质量和模型的泛化能力，确保数据更具代表性。

表1 遗传突变数据集特征汇总（**红色部分是后续分析保留特征**）

#	字段名	描述	数据类型	缺失率 (%)
1	CHROM	变异所在的染色体	object	0.000
2	POS	变异在染色体上的位置	int64	0.000
3	REF	参考等位基因	object	0.000
4	ALT	替代等位基因	object	0.000
5	AF_ESP	ESP 数据库中的等位基因频率	float64	0.000
6	AF_EXAC	ExAC 数据库中的等位基因频率	float64	0.000
7	AF_TGP	TGP 数据库中的等位基因频率	float64	0.000
8	CLNDN	临床疾病名称	object	0.000
9	CLNVC	变异的分类类型	object	0.000
10	ORIGIN	变异的来源（如生殖系、体细胞）	int64	0.000
11	CLASS	分类结果（0：一致，1：冲突）	int64	0.000
12	Allele	等位基因描述	object	0.000
13	Consequence	变异的生物学后果	object	0.000
14	IMPACT	变异的影响等级（如高、中、低）	object	0.000
15	STRAND	链信息（1：正链，-1：负链）	float64	0.000
16	Feature	特征标识符	object	0.021
17	Feature_type	特征类型（如转录本、调控区域）	object	0.021
18	BIOTYPE	特征的生物类型（如蛋白编码）	object	0.025
19	SYMBOL	基因符号	object	0.025
20	MC	变异的分子后果	object	1.298
21	CADD_RAW	CADD 原始分数	float64	1.675
22	CADD_PHRED	CADD Phred 分数	float64	1.675
23	LoFtool	基因耐受度的 LoFtool 分数	float64	6.463
24	cDNA_position	变异在 cDNA 上的位置	object	13.628
25	EXON	外显子信息	object	13.642
26	CDS_position	变异在 CDS 上的位置	object	15.271
27	Protein_position	变异在蛋白质上的位置	object	15.271
28	Amino_acids	氨基酸变化	object	15.346
29	Codons	密码子变化	object	15.346
30	BAM_EDIT	BAM 编辑信息	object	50.959
31	CLNVI	临床变异解释	object	57.570
32	BLOSUM62	BLOSUM62 替换矩阵分数	float64	60.740
33	SIFT	SIFT 预测的蛋白影响	object	61.901
34	PolyPhen	PolyPhen 预测的蛋白影响	object	61.962
35	INTRON	内含子信息	object	86.496
36	CLNDISDBINCL	包含的临床疾病数据库 ID	object	99.744
37	CLNDNINCL	包含的临床疾病名称	object	99.744
38	CLNSIGINCL	包含的临床意义分类	object	99.744
39	SSR	简单双核苷酸重复信息	float64	99.801
40	DISTANCE	到最近特征的距离	float64	99.834
41	MOTIF_NAME	模体名称	object	99.997
42	MOTIF_POS	模体内的位置	float64	99.997
43	HIGH_INF_POS	高信息含量的位置	object	99.997
44	MOTIF_SCORE_CHANGE	模体分数的变化	float64	99.997
45	DISTANCE	到最近特征的距离	float64	99.834
46	SSR	简单双核苷酸重复信息	float64	99.801

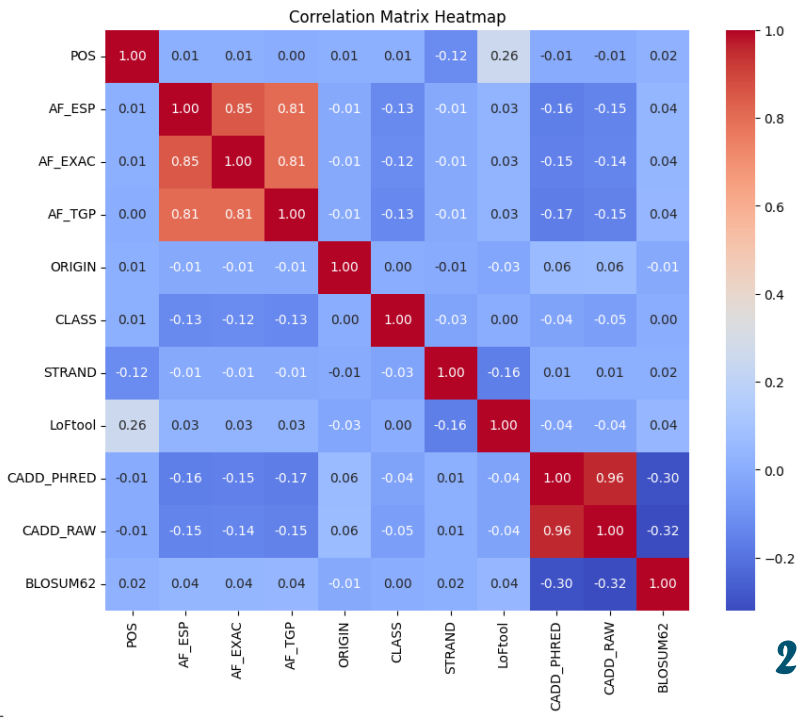
4. 补充缺失值

特征类型**选择合适的方法填补缺失值**。对于数值型特征，使用均值或中位数填补，如 df.fillna(df.mean())；对于类别型特征，使用众数填补，如 df.fillna(df.mode()[0])。也可以通过插值法（df.interpolate()）补全缺失数据。

5. 可视化异常值分布

通过箱线图（sns.boxplot()）可视化异常值的分布，查看数据的离散点，进一步确认异常值的位置和影响，为剔除或调整数据处理策略提供支持，确保数据质量。

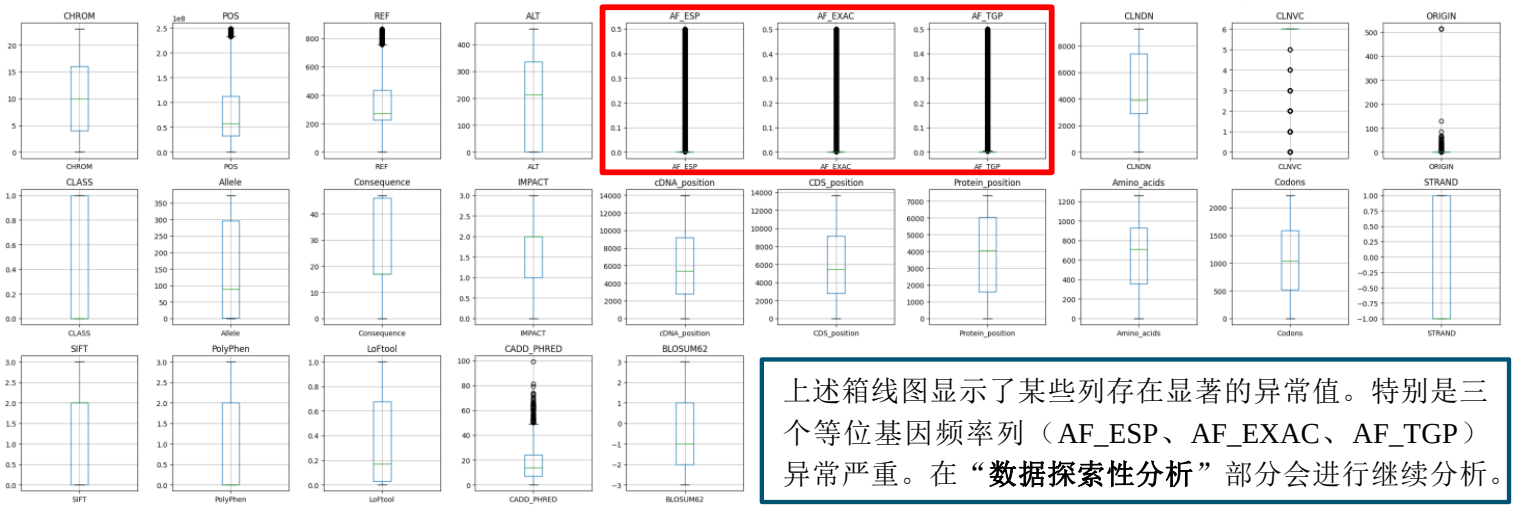
图4 检查变量间关系（热力图）



张鳃沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第二部分　数据探索性分析



上述箱线图显示了某些列存在显著的异常值。特别是三个等位基因频率列（AF_ESP、AF_EXAC、AF_TGP）异常严重。在“**数据探索性分析**”部分会进行继续分析。

图5 可视化异常值分布

数据探索性分析（EDA）的目的

通过EDA, 可以用直观的统计图表和数值分析**深入分析数据集的结构和特征分布**。具体包括识别数据的分布规律、检测和处理缺失值及异常值，分析特征间的相关性，以及确定数据中潜在的模式和关系。通过直方图、箱线图、散点图和相关性矩阵等工具，可以发现数据的偏态分布、异常点和强相关特征，为后续的数据清洗、特征工程和建模提供依据。EDA还帮助选择合适的缺失值填补方法和特征变换策略，以提升数据质量和模型的性能。

AF_ESP：来自 NHLBI 外显子测序项目的等位基因频率数据，涵盖多种疾病样本。

AF_EXAC：由 Exome Aggregation Consortium 提供，基于大规模外显子组数据的等位基因频率。

AF_TGP：源自 1000 基因组计划，提供全球不同人群的等位基因频率数据。

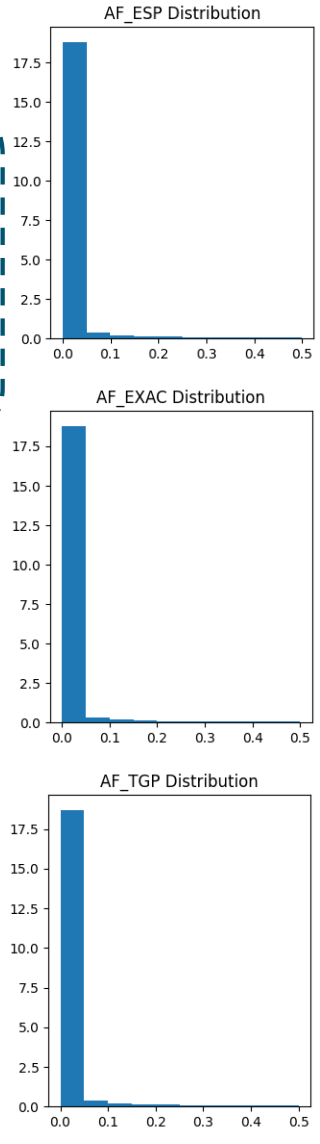


图6 等位基因频率偏态分布

EDA流程

1. 绘制特征分布图

使用直方图 (plt.hist()) 展示特征的分布情况，观察数据的分布形态和可能存在的偏态或异常值。通过将 AF_ESP、AF_EXAC 和 AF_TGP 三个特征的分布放在同一图中，可以直观地了解等位基因频率的分布特征，为后续的数据清洗和特征处理提供参考。

2. 统计、填充缺失值

通过统计 AF_ESP、AF_EXAC 和 AF_TGP 列中零值的数量，判断零值是否代表缺失值，并进一步分析各列缺失值的比例，明确需要填补的数据量。根据缺失值的数量和数据类型，选择合适的方法（均值、中位数和插值法等）进行填补。通过比较填补后的偏度和峰度来评估不同方法的效果，保持数据的连续性和分布特性。

3. 特征变换

采用 Box-Cox 变换、对数变换和 Winsorization 等方法，探究最优方法以可以改善数据分布，减少异常值对模型的影响，能够使数据更接近正态分布。

4. 特征工程和特征选择

通过计算特征间的相关性矩阵，使用热力图 (sns.heatmap()) 进行可视化，识别强相关和弱相关的特征，以剔除冗余特征，减少模型复杂度。高相关性可能导致多重共线性问题，通过散点图进一步验证特征间的关系，为特征选择提供支持。

张鲴沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第二部分　数据探索性分析

数据探索性分析过程及结论

由图6（报告第3页），**等位基因频率分布显示出相似的右偏态分布。**结合 0% 的缺失值和高相关性，这些列之间并无显著差异。但是这些列的缺失值被标记为 "0"，Python 在迭代时将其计为非空值。

进一步探查这些 “0” 后，发现零值数量各不相同。ESP 和 TGP 的缺失百分比分别为 54.89% 和 58.25%，而 EXAC 仅为 36.89%。EXAC 似乎更全面。**然而，关联不等于因果关系，需要进一步验证。**

运用三种填充缺失值方法（均值、中位数和插值法）的偏度和峰度。**插值法在数据上引入了最小的偏度和峰度。**

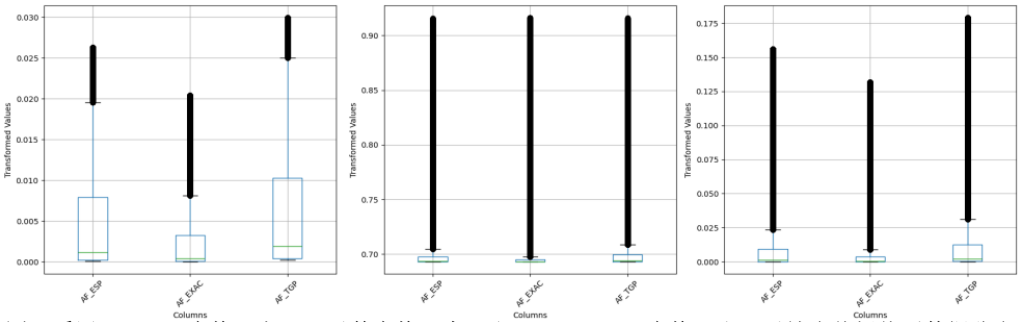


图7 采用Box-Cox 变换（左）、对数变换（中）和 Winsorization变换（右）对缺失值插值后数据分布

从箱线图的表现来看，Box-Cox 变换对三列的异常值处理效果最好。因此选用Box-Cox 变换对三列的异常值进行处理。

通过分析发现，一些变量之间存在很高的相关性。这对模型来说既可能是好事，也可能是坏事。**目标变量与特征之间的高相关性可能意味着这些变量是良好的预测因子，也可能表示存在多重共线性问题。**因此引入特征工程和特征选择。

通过散点图描绘与相关系数探寻，我们找到了相关特征的关联强度。其中，图8-10 显示 AF_ESP、AF_EXAC 和 AF_TGP 之间存在较强的正相关性，说明它们在总体上反映了相似的等位基因频率分布。后续部分以这些分析为基础开展。

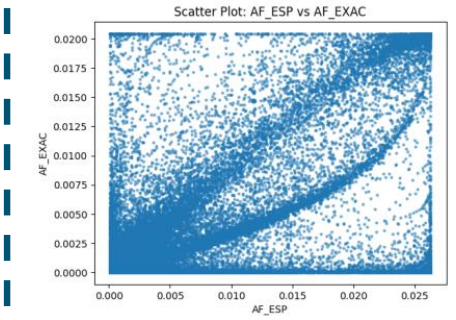


图8 AF_ESP和AF_EXAC 特征之间的关系

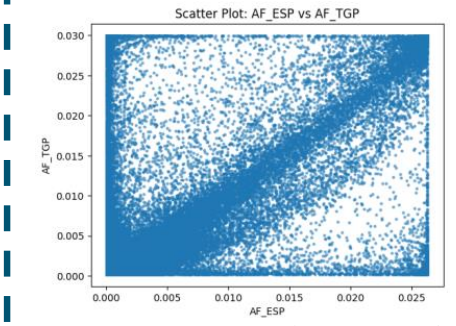


图9 AF_ESP和 AF_TGP特征之间的关系

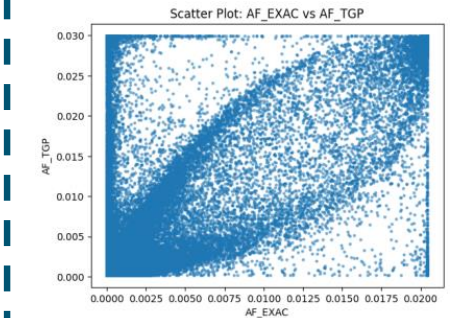


图10 AF_EXAC 和 AF_TGP特征之间的关系

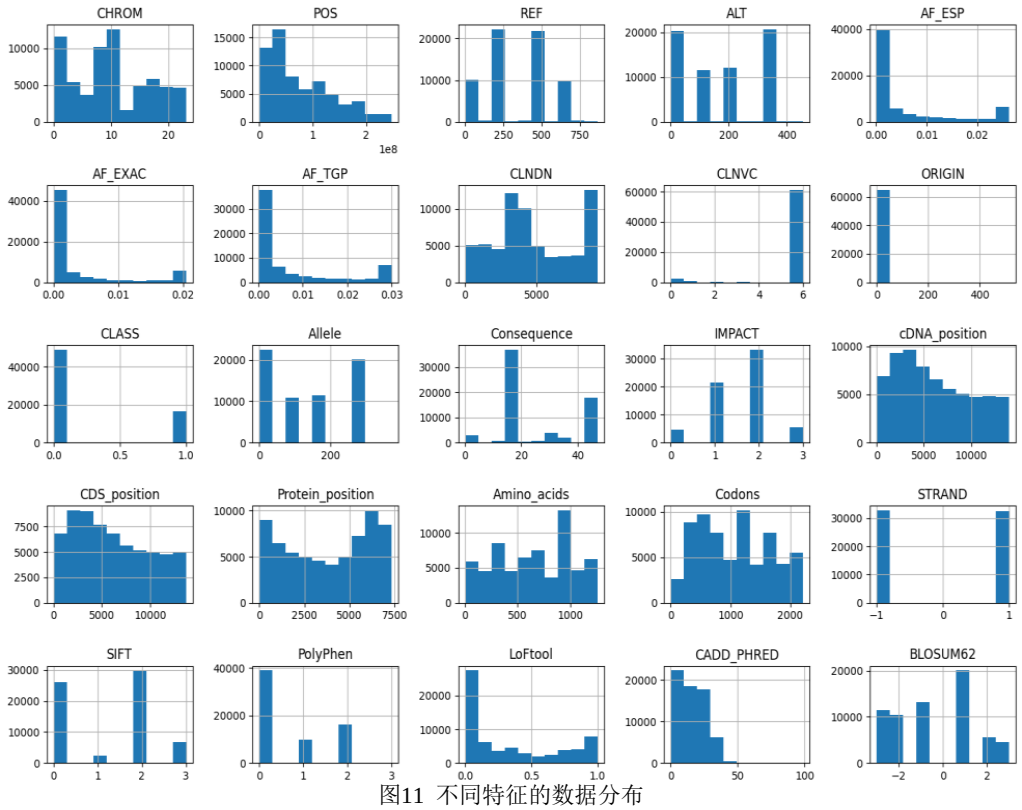


图11 不同特征的数据分布

表2 部分特征的相关系数（绝对值≥0.20）

字段 1	字段 2	相关系数
ALT	Allele	0.94
AF_ESP	AF_EXAC	0.77
AF_ESP	AF_TGP	0.73
AF_EXAC	AF_TGP	0.71
SIFT	PolyPhen	-0.62
Consequence	IMPACT	-0.61
cDNA_position	CDS_position	0.56
CDS_position	Protein_position	-0.48
REF	Allele	-0.46
cDNA_position	Protein_position	-0.44
REF	ALT	-0.40
Consequence	Codons	0.36
Consequence	CADD_PHRED	-0.31
PolyPhen	CADD_PHRED	0.28
CLNVC	Consequence	0.27
SIFT	CADD_PHRED	-0.25
POS	LoFtool	0.24
IMPACT	Codons	-0.20
CLNVC	IMPACT	0.20

张鲖沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第三部分　模型构建

LASSO 模型

模型介绍

LASSO 回归（Least Absolute Shrinkage and Selection Operator）是一种经典的线性回归模型，通常用于回归任务，尤其在高维数据中表现出色。它的基本思想是在最小二乘法的基础上，通过加入 L1 正则化项来约束模型的参数，从而达到特征选择和防止过拟合的目的。LASSO 回归的核心目标是找到一组最佳的参数，使模型在拟合数据的同时保持稀疏性，即将一些特征的权重收缩到 0。其损失函数由最小二乘项和 L1 正则化项构成，优化方法可以使用**坐标梯度下降法**、**近端梯度下降法**等。由于其特征选择能力和良好的泛化性能，LASSO 回归在特征筛选和稀疏建模中被广泛应用。

数学表示

LASSO回归的数学模型可以表示为：

$$\min_{\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-x_i^T\beta\right)^2+\lambda\parallel\beta\parallel_1\right\}$$

其中，y_i是响应变量，x_i是预测变量，β是系数向量，$\parallel\beta\parallel_1$表示系数向量的L1范数（即系数的绝对值之和），$\lambda$是正则化参数，控制着对系数的惩罚强度。

坐标梯度下降法

模型介绍

坐标梯度下降（Coordinate Descent）是一种分块优化方法，它通过在每次迭代时仅优化一个或多个坐标方向（维度），而不是同时优化所有坐标方向。这种方法适合高维问题，特别是在一些维度上计算梯度相对容易的场合。

主要步骤

1. **初始化参数：**
   设定初始参数向量$\beta=(\beta_1,\beta_2,\ldots,\beta_p)$。
2. **循环迭代：**依次更新每个参数β_j:
   对于每个参数β_j,保持其他参数固定，最小化目标函数$f(\beta)$关于β_j的子问题。
3. **检查收敛条件：**
   如果参数更新的变化量小于设定的阈值或者达到最大迭代次数，停止迭代

Code1 自行实现坐标梯度下降法LASSO模型代码

```
def lasso_coordinate_descent(X, y, alpha, max_iter=1000, tol=1e-4):
    m, n = X.shape
    beta = np.zeros(n)  # 初始化系数向量为零
    for iteration in range(max_iter):
        beta_old = beta.copy()
        for j in range(n): # 计算残差（排除当前 j 特征的影响）
            residual = y - (X @ beta - X[:, j] * beta[j])
            rho_j = np.dot(X[:, j], residual) / n
            if j == 0:          # 如果是截距项（第0列），不施加正则化
                beta[j] = rho_j
            else:
                beta[j] = np.sign(z) * max(abs(z) - alpha / n, 0)  # 软阈值函数
            if np.linalg.norm(beta - beta_old, ord=2) < tol:# 判断是否收敛
                break
    return beta
```

近端梯度下降法

模型介绍

近端梯度下降（Proximal Gradient Descent）是一种适用于非光滑优化问题的梯度下降方法，尤其适合目标函数由光滑部分和非光滑正则化部分组成的情况。它通过将问题分解为两个步骤：**梯度下降更新**和**近端映射**。

主要步骤

1. **初始化参数：**
   设定初始参数向量$\beta=(\beta_1,\beta_2,\ldots,\beta_p)$。
2. **梯度下降更新：**
   对目标函数的光滑部分（如最小二乘项）进行梯度下降更新：

$$\beta^{(t+1/2)}=\beta^{(t)}-\eta\nabla f\left(\beta^{(t)}\right)$$

其中，η是学习率，$\nabla f\left(\beta^{(t)}\right)$是光滑部分的梯度。

3. **近端映射 (Proximal Mapping)：**
   对更新后的参数应用近端算子来处理非光滑正则化项：

$$\beta^{(t+1)}=\mathrm{prox}_{\lambda\eta g}\left(\beta^{(t+1/2)}\right)$$

对于L1正则化，**近端映射是软阈值函数：**

$$\mathrm{prox}_{\lambda\eta g}(z)=S(z,\lambda\eta)=\mathrm{sign}(z)\cdot\max(|z|-\lambda\eta,0)$$

4. **检查收敛条件：**
   如果参数更新的变化量小于设定的阈值或者达到最大迭代次数，停止迭代。

张鲖沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第三部分　模型构建

最优参数搜寻

对 alpha 进行参数搜索

alpha 是正则化强度的控制参数，它决定了模型对特征权重的惩罚程度。**合适的 alpha 值可以在防止过拟合和保持模型准确性之间找到平衡。**较大的alpha会使更多特征的权重收缩为 0，有助于特征选择；较小的 alpha 则保留更多特征。在任务背景下，找到最佳 alpha 可以确保模型具备良好的泛化能力。

对学习率衰减参数进行参数搜索

在使用近端梯度下降法训练 LASSO 回归模型时，**学习率衰减参数会影响模型的收敛速度和最终性能。**较高的学习率衰减可以帮助模型快速收敛，但可能导致收敛到局部最优；较低的学习率衰减则有助于更精细地调整模型参数，但可能使训练过程变得缓慢或陷入未收敛状态。在此任务背景下，为了在不同数据特征和应用场景中取得最佳回归效果，如减少模型的误差并提高稳定性，需要对学习率衰减参数进行搜索，以权衡收敛速度与收敛精度，最终优化模型性能。

Code2 自行实现近端梯度下降法LASSO模型代码

```python
def soft_thresholding(x, lambda_):  # 软阈值函数定义
    return np.sign(x) * np.maximum(np.abs(x) - lambda_, 0)
# 使用近端梯度下降法进行 Lasso 回归
def proximal_gradient_descent(X, y, alpha, lr=0.01, decay=0.99, max_iter=5000,
tol=1e-6):
    m, n = X.shape
    beta = np.zeros(n)      # 初始化系数向量为零
    bias = 0                # 初始化截距项为零
    for iteration in range(max_iter):
        beta_old = beta.copy()
        bias_old = bias
        y_pred = X.dot(beta) + bias
        gradient = -X.T.dot(y - y_pred) / m
        bias_gradient = -np.sum(y - y_pred) / m
        beta -= lr * gradient
        bias -= lr * bias_gradient
        beta = soft_thresholding(beta, alpha * lr)
        if np.linalg.norm(beta - beta_old, ord=2) < tol and abs(bias - bias_old) < tol:
            print(f"Converged in {iteration + 1} iterations.")
            break
        lr *= decay
    return beta, bias
```

表3 LASSO模型训练参数

参数	量值						
alpha	0.00001	0.0001	0.001	0.01	0.1	1	10
学习率衰减参数（近端梯度下降法适用）	1	0.999	0.9	0.95	0.9	0.85	

表4 使用坐标梯度下降法的 LASSO 回归结果

Alpha	MSE	RMSE	MAE	R^2	Pearson Correlation
1×10^{-5}	58.4159	7.6430	5.8202	0.6907	0.8311
0.0001	58.4156	7.6430	5.8202	0.6907	0.8311
0.001	58.4149	7.6430	5.8203	0.6907	0.8311
0.01	**58.4129**	**7.6428**	**5.8216**	**0.6908**	**0.8311**
0.1	58.5534	7.6520	5.8470	0.6900	0.8309
1	63.7652	7.9853	6.2564	0.6624	0.8250
10	188.8922	13.7438	12.4121	-0.0000	nan

表5 使用手动实现的坐标梯度下降法的 LASSO 回归结果

Alpha	MSE	RMSE	MAE	R^2	Pearson Correlation
1×10^{-5}	58.4205	7.6433	5.8182	0.6907	0.8311
0.0001	58.4205	7.6433	5.8182	0.6907	0.8311
0.001	58.4205	7.6433	5.8182	0.6907	0.8311
0.01	58.4205	7.6433	5.8182	0.6907	0.8311
0.1	58.4205	7.6433	5.8182	0.6907	0.8311
1	58.4204	7.6433	5.8182	0.6907	0.8311
10	**58.4197**	**7.6433**	**5.8182**	**0.6907**	**0.8311**

代码优化

运用F检验，选择 k 个最佳特征

Code3　运用F检验选择 k 个最佳特征

```python
selector = SelectKBest(score_func=f_regression, k=k)
X_new = selector.fit_transform(X, y)
```

在选择 F 检验（f_regression） 进行特征选择，是因为它能**衡量每个特征与目标变量之间的线性相关性**，从而快速筛选出对回归任务贡献显著的特征。F 检验通过计算特征方差与目标变量方差之间的比率，评估特征的预测能力，**有助于减少冗余特征，降低模型复杂度，从而避免过拟合。**此外，F 检验计算简单高效，适用于连续型数据的回归任务，特别是在特征与目标变量存在线性关系时，能够显著提高模型性能。

引入截距项

Code4　坐标梯度下降法LASSO模型引入截距项

```python
# 在特征矩阵中添加一列全1作为截距项
X_train_bias = np.hstack([np.ones((X_train.shape[0], 1)), X_train])
X_test_bias = np.hstack([np.ones((X_test.shape[0], 1)), X_test])
```

在特征矩阵中添加一列全 1 作为截距项是为了在回归模型中显式引入截距（或偏置）。这一步确保模型能够在没有输入特征为零的情况下依然预测出合理的结果。回归模型的标准形式包含一个截距项 β_0，**如果不加入截距，模型将强制通过原点，这往往会导致拟合效果变差或出现欠拟合。**通过在特征矩阵左侧拼接一列全 1，等效于为模型添加了一个可以学习的偏置项，这对于自定义实现梯度下降等优化方法非常必要，有助于提高模型的灵活性和准确性，以防出现决定系数 (R^2) 小于 0 或者皮尔逊相关系数为 NaN的情况。

张鳃沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第三部分　模型构建

模型评估指标

1. 均方误差 (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

均方误差是预测值与真实值之间的平方误差的平均值，用于衡量模型预测的整体误差大小。MSE 对误差进行了平方放大，因此对异常值非常敏感。如果模型预测存在较大的偏差，MSE 值会显著增大。**MSE 越小越好，表示模型的预测误差越小。**

2. 均方根误差 (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

均方根误差是均方误差的平方根，具有与原始数据相同的量纲，更便于理解和解释误差的实际大小。与 MSE 一样，RMSE 对较大的误差较为敏感，能够放大模型预测中的大误差点的影响。**RMSE 越小越好，表示模型的预测误差越小。**

3. 平均绝对误差 (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

平均绝对误差是预测值与真实值之间的绝对误差的平均值，不会放大误差，因此对异常值的敏感性较低。MAE 直观地反映了模型预测误差的平均水平，具有良好的可解释性。**MAE 越小越好，表示模型预测的平均误差越小。**

4. 决定系数 (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

决定系数 R^2 衡量了模型对数据变异的解释程度，取值范围通常在 $(-\infty, 1]$ 之间。**R^2 = 1 表示模型完美拟合数据；R^2 = 0 表示模型的预测效果与简单的均值预测相当；负值表示模型预测效果比均值预测更差。**

5. 皮尔逊相关系数 (Pearson Correlation)

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

表6 使用近端梯度下降法的 LASSO 回归结果

Alpha	Decay	MSE	RMSE	MAE	R^2	Pearson Correlation
1×10^{-5}	1	57.8314	7.6047	5.7803	0.6937	0.8329
	0.999	**57.7931**	**7.6022**	**5.7807**	**0.6939**	**0.8331**
	0.99	167.4973	12.9421	10.4385	0.1128	0.8191
	0.95	587.3129	24.2345	21.1072	-2.1109	0.7933
	0.9	704.2320	26.5374	23.2490	-2.7301	0.7859
	0.85	748.7971	27.3642	24.0249	-2.9662	0.7830
0.0001	1	57.8315	7.6047	5.7803	0.6937	0.8329
	0.999	57.7932	7.6022	5.7808	0.6939	0.8330
	0.99	167.4983	12.9421	10.4385	0.1128	0.8191
	0.95	587.3135	24.2346	21.1072	-2.1109	0.7933
	0.9	704.2324	26.5374	23.2490	-2.7301	0.7859
	0.85	748.7974	27.3642	24.0249	-2.9662	0.7831
0.001	1	57.8329	7.6048	5.7805	0.6937	0.8329
	0.999	57.7946	7.6023	5.7810	0.6939	0.8330
	0.99	167.5081	12.9425	10.4386	0.1128	0.8191
	0.95	587.3202	24.2347	21.1072	-2.1109	0.7933
	0.9	704.2365	26.5375	23.2490	-2.7302	0.7860
	0.85	748.8004	27.3642	24.0249	-2.9662	0.7831
0.01	1	57.8422	7.6054	5.7822	0.6936	0.8329
	0.999	57.8099	7.6033	5.7833	0.6938	0.8330
	0.99	167.6059	12.9463	10.4393	0.1122	0.8191
	0.95	587.3867	24.2361	21.1073	-2.1112	0.7937
	0.9	704.2773	26.5382	23.2490	-2.7304	0.7865
	0.85	748.8296	27.3648	24.0249	-2.9664	0.7837
0.1	1	58.0478	7.6189	5.8101	0.6925	0.8325
	0.999	58.0506	7.6191	5.8117	0.6925	0.8325
	0.99	168.6514	12.9866	10.4488	0.1067	0.8189
	0.95	588.0447	24.2496	21.1083	-2.1147	0.7968
	0.9	704.6794	26.5458	23.2486	-2.7325	0.7907
	0.85	749.1177	27.3700	24.0246	-2.9679	0.7883
1	1	63.4857	7.9678	6.2315	0.6637	0.8264
	0.999	63.5004	7.9687	6.2324	0.6637	0.8264
	0.99	179.5749	13.4006	10.5045	0.0488	0.8166
	0.95	594.3686	24.3797	21.1240	-2.1482	0.8045
	0.9	708.4077	26.6159	23.2470	-2.7523	0.8024
	0.85	751.7508	27.4181	24.0232	-2.9818	0.8017
10	1	188.8102	13.7408	12.3971	-0.0001	nan
	0.999	188.8107	13.7408	12.3966	-0.0001	nan
	0.99	278.9929	16.7031	10.9745	-0.4778	nan
	0.95	630.8986	25.1177	21.1976	-2.3417	nan
	0.9	728.4440	26.9897	23.2926	-2.8584	nan
	0.85	765.5404	27.6684	24.0517	-3.0549	nan

皮尔逊相关系数 r 反映了预测值与真实值之间的线性相关性，取值范围为 [-1, 1]。皮尔逊相关系数的绝对值越接近 1 越好， r = 1 表示完全正相关，r = -1 表示完全负相关，r = 0 表示无线性相关性。

张鲖沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第四部分　结果分析

实验结果

表7 3种模型最优参数对比

模型	最优参数	MSE	RMSE	MAE	R^2	皮尔逊相关系数
坐标梯度下降法	$\alpha = 0.01$	58.4129	7.6428	5.8216	0.6908	0.8311
手动实现坐标梯度下降法	$\alpha = 10$	58.4197	7.6433	5.8182	0.6907	0.8311
近端梯度下降法	$\alpha = 1 \times 10^{-5}$, **decay=0.999**	**57.7931**	**7.6022**	**5.7807**	**0.6939**	**0.8331**

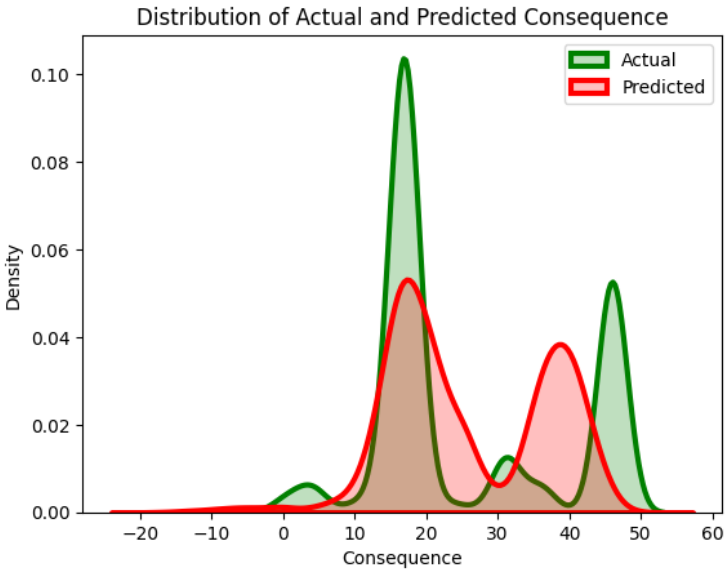


图12 实际值和预测值的分布密度图

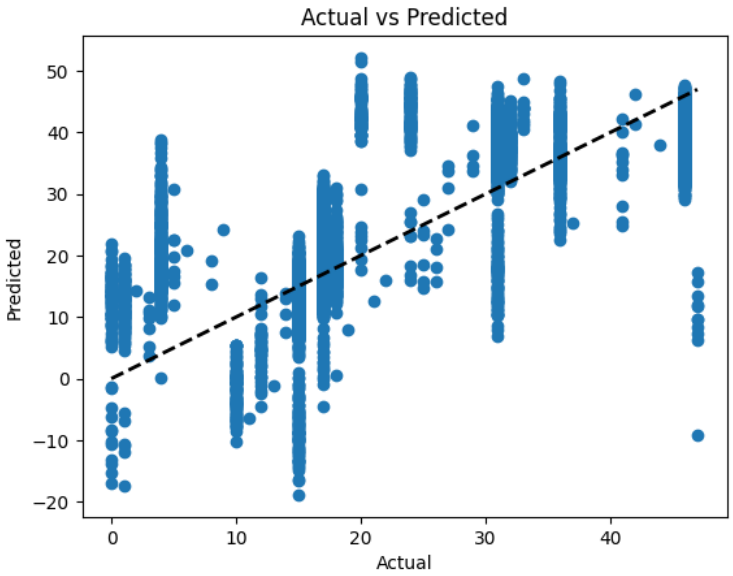


图13 实际值与预测值的散点图

对于图12分布密度图的分析

在实际值（绿色曲线）和预测值（红色曲线）的密度分布中，**两者在主峰附近存在较好的重合**，特别是在 20 左右的峰值处，**模型能够较准确地捕捉数据的主要分布趋势。**这表明在大多数情况下，模型对主流数据的预测较为准确。

然而，**分布曲线也显示出明显的偏差：**在 30 到 50 的区域，实际值的次峰和预测值之间存在较大的差异，预测值分布偏向于低估这一部分的实际值。此外，在分布的尾部区域（例如，0 以下和 50 以上），红色曲线和绿色曲线出现显著的分离，**表明模型在极端值或边缘数据上的预测效果不理想。**这种偏差可能是导致 MSE 较高的原因之一。

对于图13散点图的分析

散点图展示了实际值和预测值之间的线性关系。理想情况下，所有散点应沿着 45 度的虚线均匀分布，这意味着预测值和实际值完全一致。

从图中可以看出，**大部分点分布在虚线附近，说明模型在总体趋势上能够较好地拟合数据。**然而，散点在多个区域存在明显的偏离，这种偏离在实际值较高和较低的部分尤为明显，说明**模型在某些范围内的预测存在系统性误差。**

此外，**离散的散点和极端值进一步揭示了模型在复杂场景下拟合不足的问题。**这些偏差可能是由于特征不足、异常值影响或模型正则化导致的欠拟合。

最优模型函数为：

$$\hat{y} = 25.5790 + 0.1460 \cdot x_{\text{CHROM}} + 0.3758 \cdot x_{\text{POS}} + 0.1213 \cdot x_{\text{REF}} - 0.2042 \cdot x_{\text{ALT}} - 0.0126 \cdot x_{\text{AF\_ESP}} + 0.3862 \cdot x_{\text{AF\_EXAC}}$$
$$- 0.0818 \cdot x_{\text{AF\_TGP}} + 0.2523 \cdot x_{\text{CLNDN}} + 5.1127 \cdot x_{\text{CLNVC}} - 0.1021 \cdot x_{\text{ORIGIN}} + 0.3026 \cdot x_{\text{CLASS}} + 0.4777 \cdot x_{\text{Allele}}$$
$$- 9.4155 \cdot x_{\text{IMPACT}} - 0.0082 \cdot x_{\text{cDNA\_position}} + 0.0151 \cdot x_{\text{CDS\_position}} - 0.0453 \cdot x_{\text{Protein\_position}} + 0.0252 \cdot x_{\text{Amino\_acids}} + 2.1627 \cdot x_{\text{Codons}}$$
$$+ 0.1025 \cdot x_{\text{STRAND}} - 0.4625 \cdot x_{\text{SIFT}} + 0.5999 \cdot x_{\text{PolyPhen}} - 0.2087 \cdot x_{\text{LoFtool}} - 4.4786 \cdot x_{\text{CADD\_PHRED}} + 0.0341 \cdot x_{\text{BLOSUM62}} + 0.0669 \cdot x_{\text{AF\_avg}}$$

张鳃沣 北京交通大学计算机科学与技术学院 AI2201 班 22281052@bjtu.edu.cn

第二次大作业 实验报告

第五部分　结论感想

最终结论

基于近端梯度下降法的 LASSO 回归模型（$\alpha = 1 \times 10^{-5}$, decay=0.999），对linVar 基因变异数据集进行了建模，成功展现了基因变异特征与临床后果 (Consequence) 之间的显著关系。筛选出的关键特征包括 CLNVC （临床变异类别，如错义、无义等）、IMPACT （变异的影响程度，如高影响、低影响等）、Codons（编码子变异信息）和 CADD PHRED（综合变异预测得分，衡量变异的潜在致病性）。模型性能指标显示拟合效果良好（$R^2 = 0.6939$,皮尔逊相关系数= 0.8331），表明模型能够有效捕捉基因变异特征与临床后果之间的联系，为临床解读基因变异、评估变异的潜在致病性提供了有力支持。

结果分析

1. 为什么近端梯度下降法相对比坐标梯度下降来说效果较好?

近端梯度下降法在你的任务中表现更好的主要原因是它在**每次迭代中同时更新所有特征的系数，能够更有效地处理特征间的相关性**，避免了坐标梯度下降法逐一更新特征时可能陷入局部最优或收敛缓慢的问题。此外，通过引入学习率衰减（decay = 0.999），近端梯度下降法在训练初期快速收敛，而在后期通过逐步减小步长实现了更精细的参数调整。这种**全量更新与动态步长结合的特性**，使得近端梯度下降法在高维数据和复杂分布下更具稳定性和鲁棒性，最终在整体误差和拟合效果上优于坐标梯度下降法。

2.结果出现 R^2 <0 或皮尔逊相关系数为 NaN?

当**回归结果出现 R^2 <0时**，说明模型的预测效果比简单使用真实值的均值来预测还要差，**可能是由于模型欠拟合、特征不足、数据噪声过大或异常值影响导致的。**而皮尔逊相关系数为 NaN，通常是因为实际值或预测值的方差为零，这意味着所有数据点都相同或极其接近，无法计算相关性。要解决这些问题，可以尝试增加特征、多样化数据、去除异常值，或者选择更适合数据特征的模型。

3. 为什么在进行特征选择时要对特征进行标准化或归一化?

在进行特征选择时对特征进行标准化或归一化的主要原因是，不同特征的取值范围和尺度可能存在较大差异，这会影响特征选择算法的效果。特征标准化（如使用 MinMaxScaler 或 StandardScaler）将所有特征转换到相同的尺度上，使它们对模型的贡献程度具有可比性。如果不进行标准化，取值范围较大的特征可能会对特征选择过程产生主导作用，而取值较小的特征可能被忽略。

实验感想

本次实验通过对**近端梯度下降法**和**坐标梯度下降法**两种 LASSO 回归模型的实现与优化，探讨了不同优化方法在回归任务中的表现差异。实验结果表明，**近端梯度下降法在处理高维特征和特征间相关性较强的数据时表现出显著优势。**得益于全量更新和学习率衰减机制（decay = 0.999），近端梯度下降法能够快速收敛并获得较优解，最终取得了较高的 R^2（0.6939）和皮尔逊相关系数（0.8331）。通过调整正则化参数 α 和学习率衰减，不仅有效减少了模型误差，还提高了模型的鲁棒性和稳定性。这一过程让我深刻理解了**优化算法的选择和动态步长调整对于回归任务的重要性**。相比之下，坐标梯度下降法的表现相对较差，**主要原因在于逐特征更新的方式难以有效处理特征间的复杂相关性，收敛速度较慢，且容易陷入局部最优。**

整体来看，这次实验强化了我对优化方法、参数调优和特征工程的理解，也让我体会到在回归任务中根据数据特征选择合适的模型和优化策略的重要性。

注：源代码直接下载，就可以使用。请注意，数据集文件需要和代码放到同一目录！