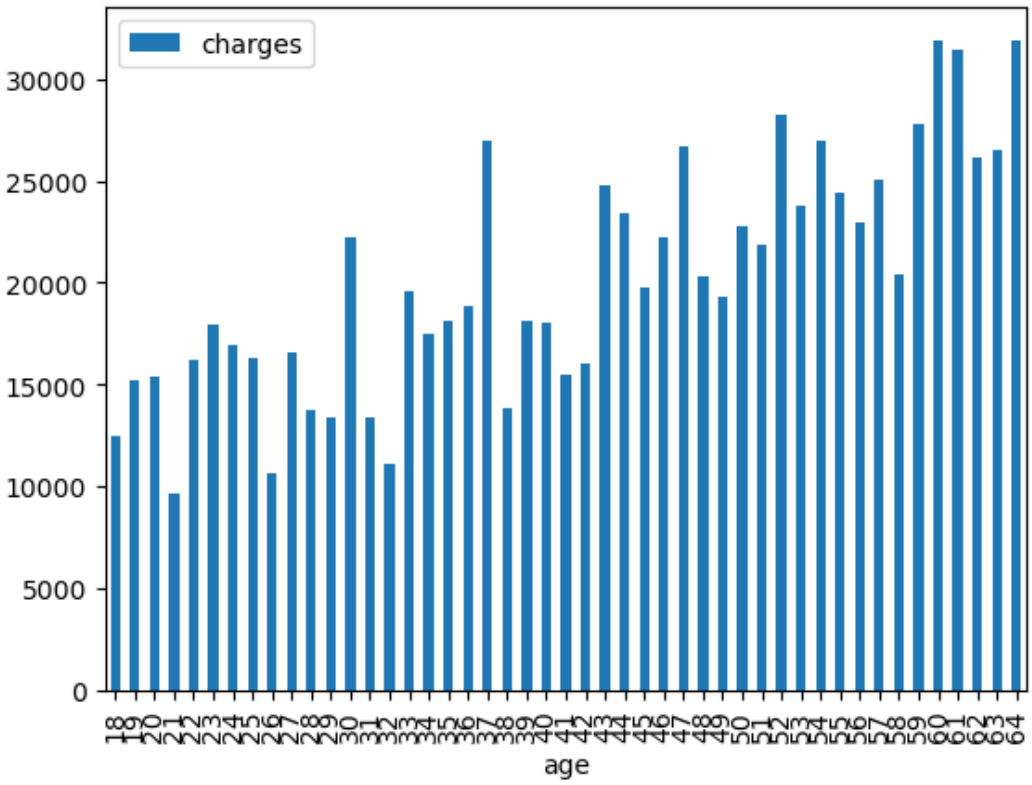
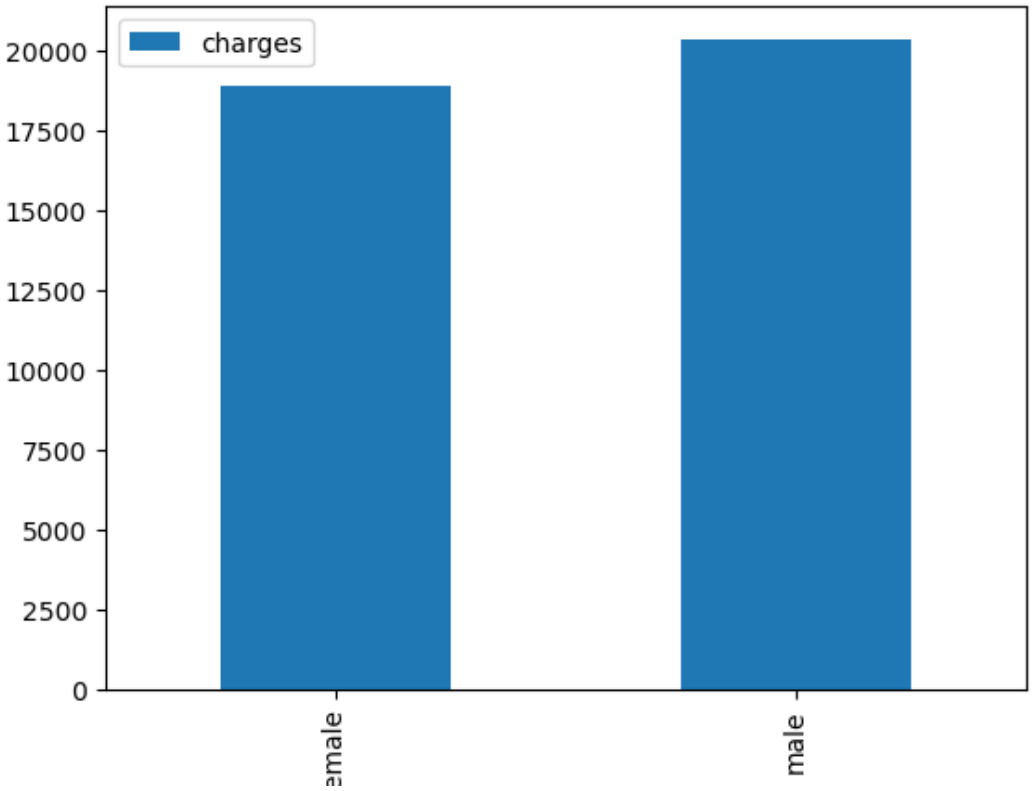
医疗花费预测

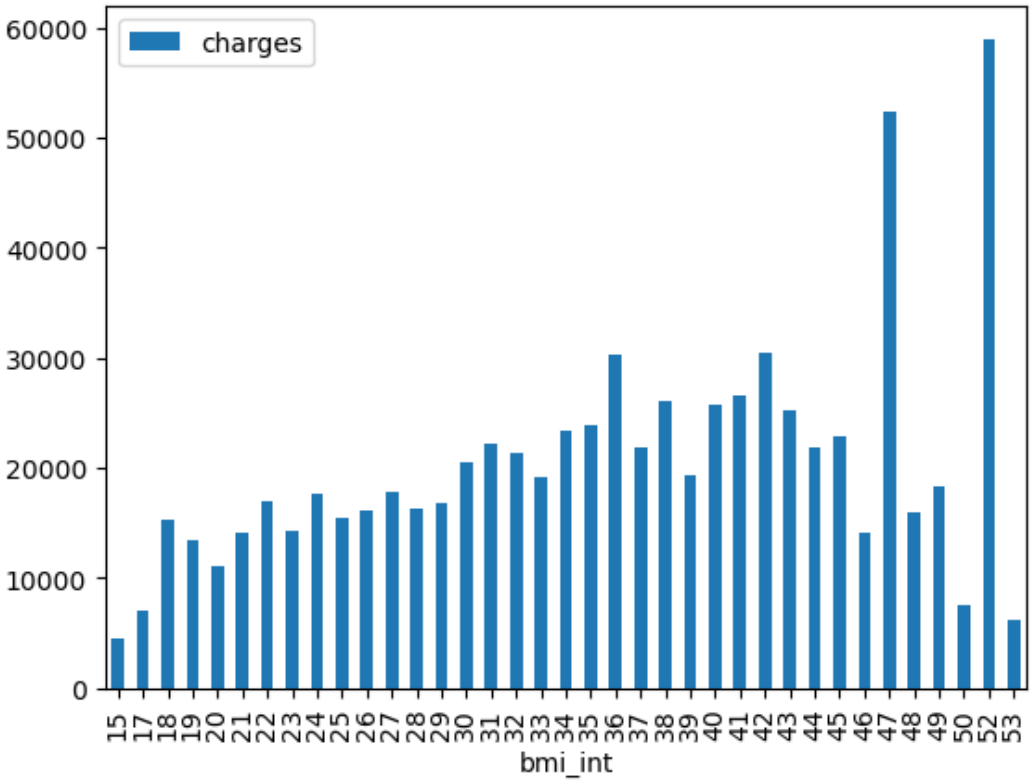
数据特征分析

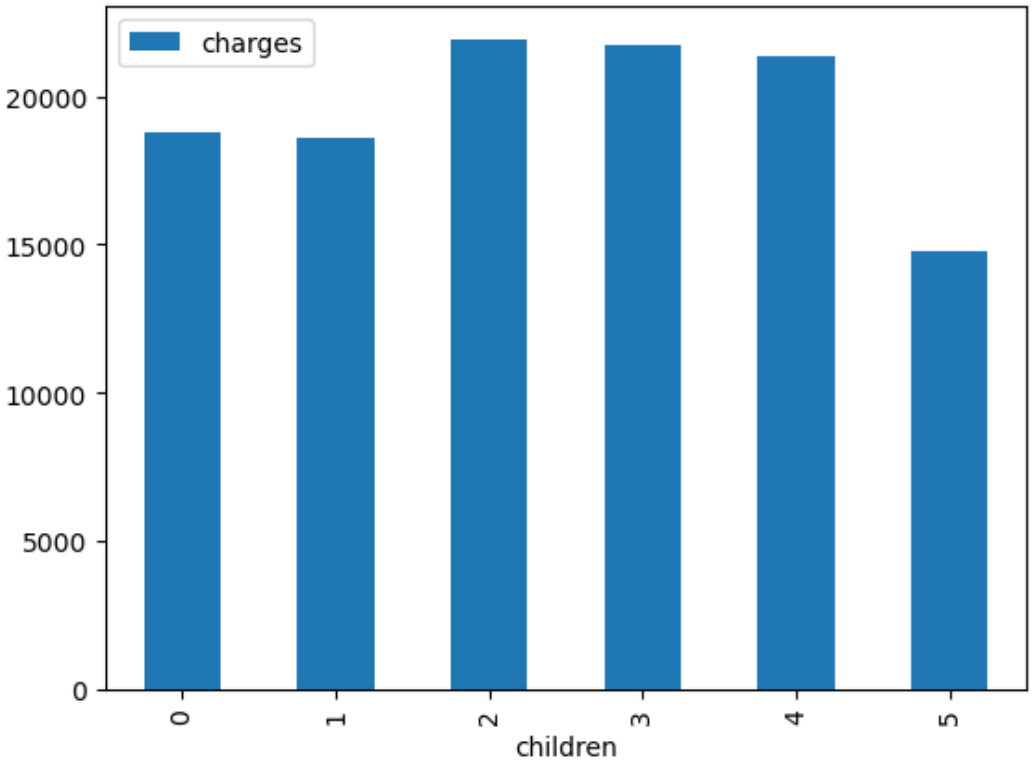
题目要求预测一个连续值，显然是一个回归问题。题目输入中含有“sex”、“smoker”和“region”三个标签。前两个标签只有两种可能的取值，因此可以直接转换为0、1；后一个标签有四种可能的取值，而且并没有明显的大小关系，因此，采取四位独热编码的策略进行编码。

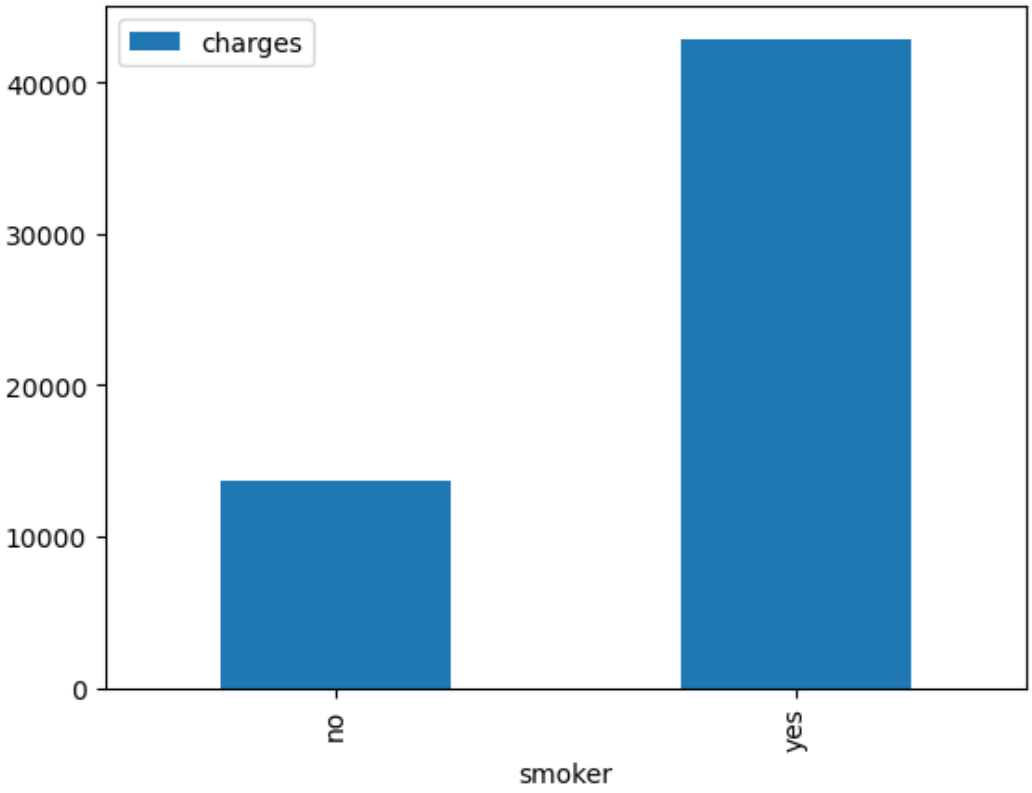
下面对每个特征对应的平均医疗花费进行可视化：

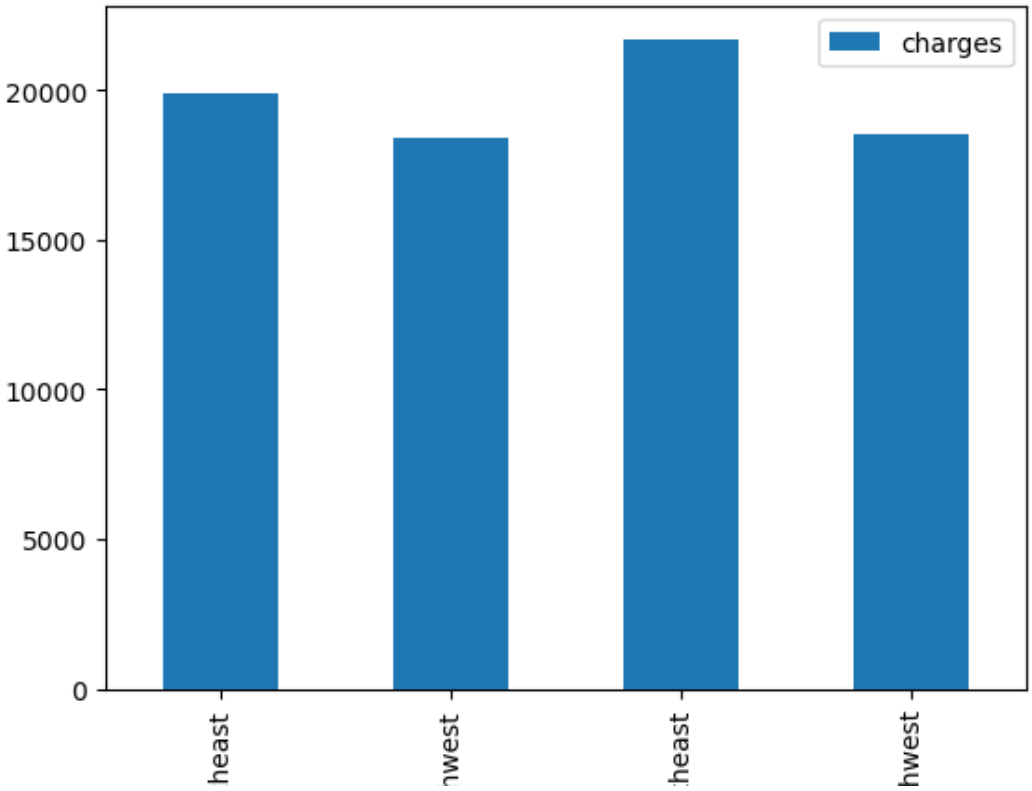



charges
20000
17500
15000
12500
10000
7500
5000
2500
0
male


charges
60000
50000
40000
30000
20000
10000
0
15 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 52 53
bmi_int


charges
20000
15000
10000
5000
0
0
1
2
3
4
5
children


charges
40000
30000
20000
10000
0
no
yes
smoker



由观察可以发现，大部分特征对医疗花费的影响大致为线性关系，但也并非绝对；有些特征对医疗花费的影响较大，而另一些影响较小。基于这个特点，下文选取了三种方法进行医疗花费的预测。

方法1：简单线性回归

方法介绍

假设一个输入样本为一个列向量$\mathbf{x}'$，为其增加一个截距项x_0，得到新的向量$\mathbf{x}$。假设结果与输入呈线性相关，我们要预测的函数就是

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

。将每一个样本作为一个行向量，按列拼接得到矩阵X，则损失函数可以写为

$$J(\mathbf{w}) = (X\mathbf{w} - \mathbf{y})^T(X\mathbf{w} - \mathbf{y})$$

。由于样本数较少，可以不使用梯度下降，而直接使用解析法对其进行求解。将其对$\mathbf{w}$求导并置0，得

$$\hat{\mathbf{w}} = (X^TX)^{-1}X^T\mathbf{y}$$

。

方法实现

将输入数据进行数值化，并添加一列截距项1，使用

$$\hat{\mathbf{w}} = (X^TX)^{-1}X^T\mathbf{y}$$

直接求解权向量即可。注意由于X^TX可能不可逆，故需要求其伪逆。无需对输入数据进行标准化。

结果分析

提交后的R^2为0.747989。通过对数据的观察，有些特征对医疗花费的影响较大，而另一些影响较小。由此，采取了可以将权向量稀疏化的下一种方法。

方法2：Lasso

方法介绍

Lasso的预测函数与简单线性回归相同，只是损失函数变为

$$J(\mathbf{w}) = \frac{1}{2n_{samples}}(X\mathbf{w} - \mathbf{y})^T(X\mathbf{w} - \mathbf{y}) + \alpha||\mathbf{w}||_1$$

。由于优化L1范数会倾向于将$\mathbf{w}$稀疏化，故Lasso会倾向于选择最为重要的几个特征。

方法实现

采用坐标下降的方法对损失函数进行优化。在每轮优化中，会假设$\mathbf{w}$的其他分量固定，只对其中一个分量求偏导，并优化一个步长。对$\mathbf{w}$的每个分量依次执行前述操作，就完成了一轮优化。

调参过程

α控制权向量稀疏化的程度。使用5折交叉验证法对α进行调整，并选择交叉验证过程中均方误差最小值对应的α。选择的范围为相差1000倍的100个值。

结果分析

提交后的R^2为0.748294。结合数据特征，猜测非线性因素对医疗花费的影响可能较大。由此，采取了非线性的下一种方法。

方法3：ϵ-SVR

方法介绍

该方法采用软间隔的支持向量机的思想，通过优化问题

$$\begin{gathered}\min_{\mathbf{w},b,\xi,\xi^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\xi_i + C\sum\xi_i^* \\ subject\ to \quad \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \xi_i, \\ y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0\end{gathered}$$

来实现回归。其中，C和ϵ为超参数，分别代表惩罚代价和容忍的错误范围；$\phi(\mathbf{x})$为核函数。使用拉格朗日乘子法，可将原问题转化为其对偶问题

$$\begin{gathered}\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha - \alpha^*)^TQ(\alpha - \alpha^*) + \epsilon\sum(\alpha_i + \alpha_i^*) + \sum y_i(\alpha_i - \alpha_i^*) \\ subject\ to \quad \mathbf{e}^T(\alpha - \alpha^*) = 0, \\ 0 \leq \alpha_i, \alpha_i^* \leq C\end{gathered}$$

。该问题是一个二次规划问题。其中，$\mathbf{e}$为各分量全为1的向量，$Q_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$。可以使用

$$\sum(\alpha_i^* - \alpha_i)K(\mathbf{x}_i, \mathbf{x}) + b$$

来预测回归值。

方法实现

在本问题中，核函数选择了高斯径向基函数Radial Basis Function（RBF）。此时，有

$$K(\mathbf{x}_1,\mathbf{x}_2)=\exp(-\frac{(\mathbf{x}_1-\mathbf{x}_2)^T(\mathbf{x}_1-\mathbf{x}_2)}{2\sigma^2})$$

。采用此核函数的优势在于其可以对原始数据中非线性的因素进行拟合。

采用SkLearn自带的SVR模型进行训练。

调参过程

ϵ控制回归函数对样本数据的不敏感区域的宽度，影响支持向量的数目，其值和样本噪声有密切关系。ϵ越大，支持向量数就少，可能导致模型过于简单，学习精度不够；ϵ越小，回归精度就较高，但可能导致模型过于复杂，得不到好的泛化性能。

惩罚系数C反映了算法对超出ϵ管道的样本数据的惩罚程度，其值影响模型的复杂性和稳定性。C越小，对超出ϵ管道的样本数据的惩罚就小，训练误差变大；C越大，学习精度相应提高，但模型的泛化性能可能变差。另外，C的值影响到对样本中"离群点"（噪声影响下非正常数据点）的处理，选取合适的C就能在一定程度上抗干扰，从而保证模型的稳定性。

由此可见：若能选取到合适的(C,ϵ)对，就能得到比较精确、稳定的回归模型。

使用SkLearn自带的GridSearchCV中的5折交叉验证法对C和ϵ进行调整，并选择交叉验证过程中R^2的最小值对应的C和ϵ。采用4路并行计算。最后得到的最优值为$C=2.15,\epsilon=0.13$。此时交叉验证的$R^2=0.84$。

遇到的问题及解决方法

在训练过程中，一开始没有对y值进行标准化，导致C增大时，训练收敛十分缓慢。后将y标准化后进行训练，大大增加了收敛速度。

结果分析

提交后的R^2为0.857448。由此，可以发现样本中存在一定的非线性因素影响。

评测排名

12名（截至2020年12月22日17:28分）。

#	用户名	登录	上次登录日期	R-square ▲
1	18377290_孙亦琦	6	11/29/20	0.8686 (1)
2	18373636_田昶尧	14	12/10/20	0.8669 (2)
3	18373202_刘勇	8	12/07/20	0.8666 (3)
4	18376161李明昕	2	12/09/20	0.8644 (4)
5	18231047_王肇凯	10	12/07/20	0.8637 (5)
6	18231217_吴陶然	24	11/28/20	0.8631 (6)
7	18373805_杨再飞	8	12/11/20	0.8619 (7)
8	18231174_任杰瑞	23	12/22/20	0.8618 (8)
9	18373054_陆晓东	12	12/10/20	0.8611 (9)
10	18373384-欧卓健	3	12/08/20	0.8600 (10)
11	陈天异	2	11/25/20	0.8575 (11)
12	18373109_孔祥浩	5	12/04/20	0.8574 (12)
13	18373488_袁劭涵	3	12/16/20	0.8570 (13)

源代码

https://github.com/refkxh/BUAA_ML_2020Autumn