11 Regression 回归

鸢尾花书有关回归算法模型的综述

卓越从来都不是偶然。卓越永远都是志存高远、百折不挠、有勇有谋的结果；它代表了明智之选。选择，而不是机会，决定了你的命运。

Excellence is never an accident. It is always the result of high intention, sincere effort, and intelligent execution; it represents the wise choice of many alternatives. Choice, not chance, determines your destiny.

—— 亚里士多德 (Aristotle) | 古希腊哲学家 | 384 ~ 322 BC

- `sklearn.linear_model.ElasticNet()` 求解弹性网络回归问题
- `sklearn.linear_model.lars_path()` 生成 Lasso 回归参数轨迹图
- `sklearn.linear_model.Lasso()` 求解套索回归问题
- `sklearn.linear_model.LinearRegression()` 最小二乘法回归
- `sklearn.linear_model.LogisticRegression()` 逻辑回归函数，也可以用来分类
- `sklearn.linear_model.Ridge()` 求解岭回归问题
- `sklearn.neighbors.NearestCentroid` 最近质心分类算法函数
- `sklearn.preprocessing.PolynomialFeatures()` 建模过程中构造多项式特征
- `statsmodels.api.add_constant()` 线性回归增加一列常数 1
- `statsmodels.api.OLS()` 最小二乘法函数

11.1 一张“回归”版图

机器学习中的回归 (regression) 是一种利用已知数据来预测未知数据的方法，常用于预测数值型的结果。回归模型通过分析数据中的变量之间的关系，得到一种数学模型，可以对未知数据进行预测。回归模型可以分为线性回归和非线性回归两种，其中线性回归模型假设变量之间存在线性关系，而非线性回归模型则允许变量之间存在更为复杂的关系。

图 1 总结本系列丛书介绍的各种回归算法。本章就按照这个“回归”版图展开讲解。这幅图也就是本章的思维导图。

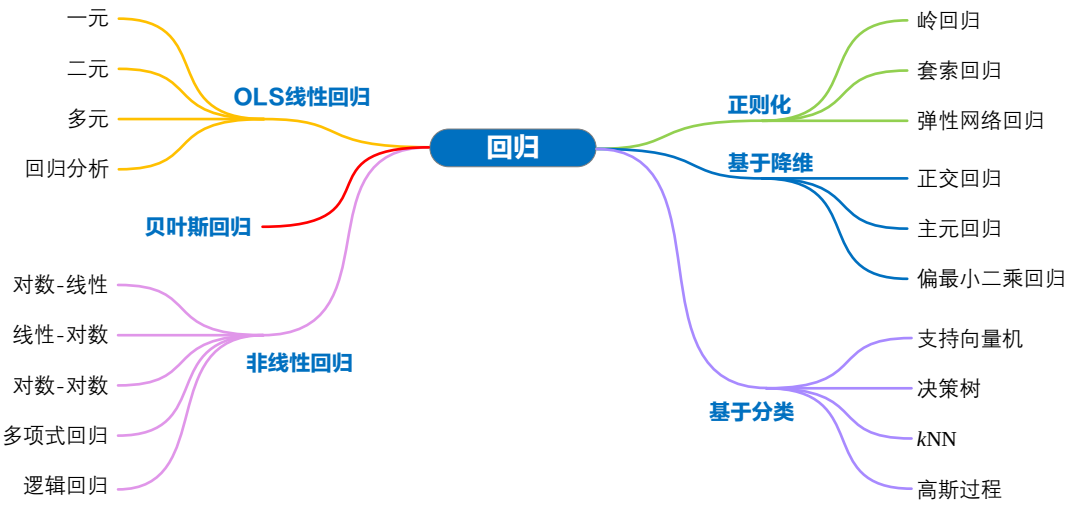

图 1. 回归方法分类

最小二乘算法

相信大家对于最小二乘 (Ordinary Least Squares, OLS) 线性回归已经烂熟于心。本章想强调如下几点。

首先，希望大家能够从多重视角理解 OLS 线性回归，比如优化 (图 2)、条件概率 (图 3)、几何 (图 4)、投影 (图 5)、数据、线性组合、SVD 分解、QR 分解、最大似然 MLE、最大后验 MAP 等视角。

此外，回归模型不能拿来就用，需要通过严格的回归分析。

再提到 OLS 线性回归时，希望大家闭上眼睛，脑中不仅仅浮现各种多彩的图像，而且能够用 OLS 线性回归把代数、几何、线性代数、概率统计、优化等数学板块有机地联结起来！

丛书讲解 OLS 线性回归时可谓抽丝剥茧、层层叠叠。对于这些视角感到生疏的话，请回归《数学要素》第 24 章、《矩阵力量》第 9、25 两章、《统计至简》第 24 章、《数据有道》第 10、11 两章。

线性回归常用来预测。代码 Bk7_Ch11_01.py 给出一个例子，根据股票日收益率协方差矩阵，推断如果股票指数上涨或下跌，其他股票价格变动情况。这个回归分析是单变量，但是多输出，也就是说一个 x、多个 y。请大家自行学习。

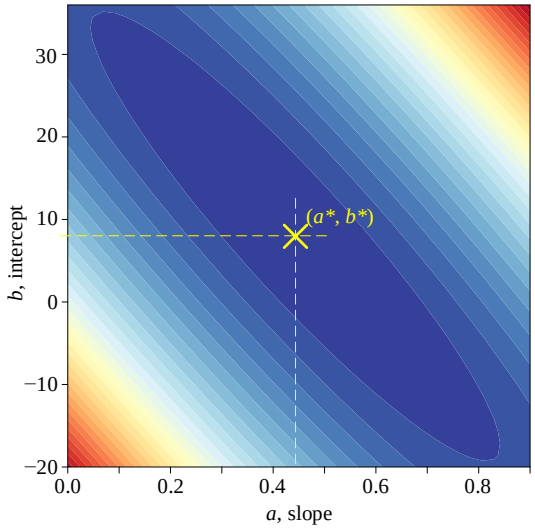

图 2. 一元 OLS 回归目标函数，图片来自《数学要素》第 24 章

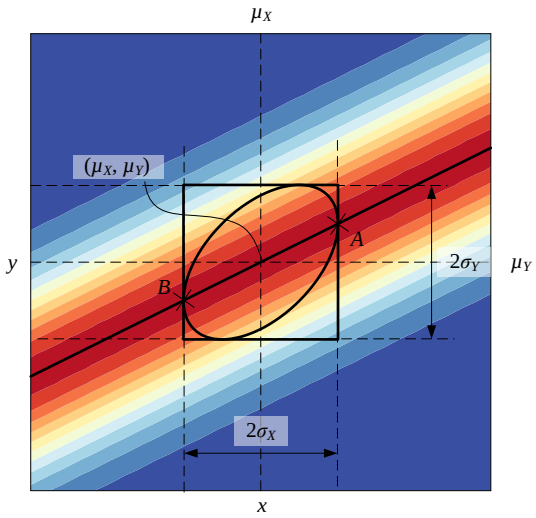

图 3. 条件期望视角看 OLS 线性回归，图片来自《统计至简》第 12 章

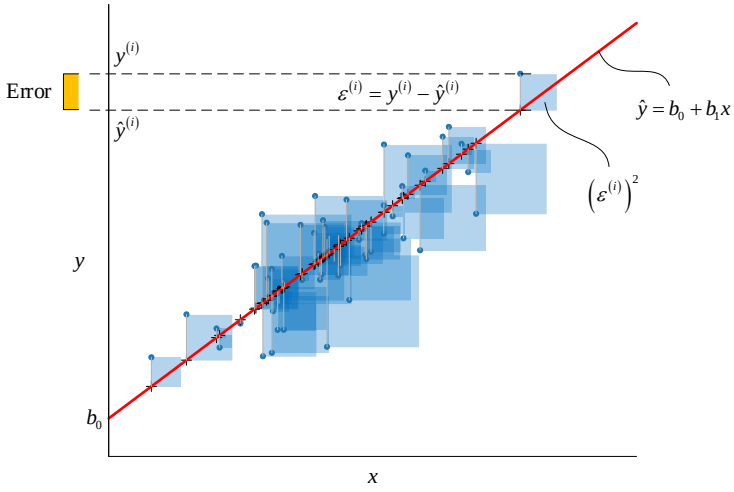

图 4. 残差平方和的几何意义，图片来自《统计至简》第 24 章

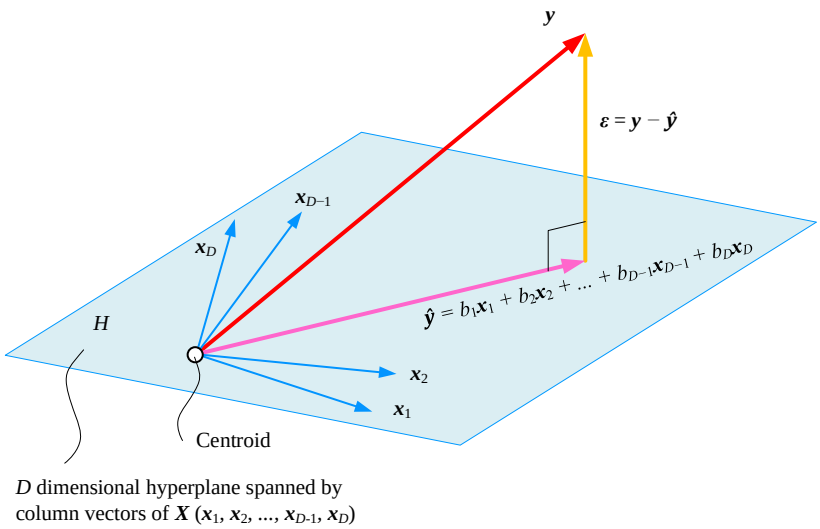

图 5. 投影角度解释多元最小二乘法线性回归，图片来自《数据有道》第 11 章

贝叶斯回归

贝叶斯回归基于贝叶斯推断 (Bayesian inference)。

贝叶斯推断把模型参数看作随机变量。根据主观经验和既有知识给出未知参数的概率分布，称为先验分布。从总体中得到样本数据后，根据贝叶斯定理，基于给定的样本数据，得出模型参数的后验分布。

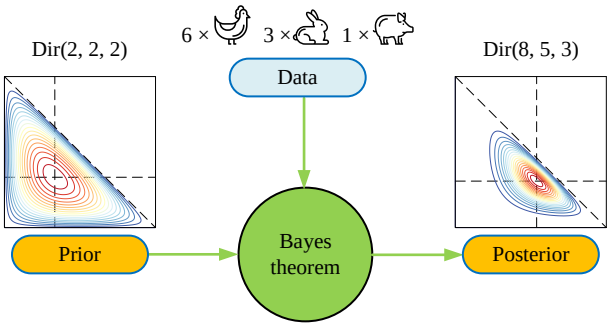

图 6. 先验 Dir(2, 2, 2) + 样本 → 后验 Dir(8, 5, 3)，图片来自《统计至简》第 22 章

贝叶斯回归的优化问题对应最大后验 MAP。贝叶斯推断中，后验 ∝ 似然 × 先验，是最重要的关系，希望大家牢记。希望大家掌握利用 PyMC3 完成贝叶斯回归参数模拟。

欢迎大家回顾《统计至简》第 20、21、22 三章有关贝叶斯推断的内容。此外，请大家回顾《数据有道》第 13 章有关贝叶斯回归。

非线性回归

非线性回归 (nonlinear regression) 寻找因变量和自变量之间关系的非线性模型的方法。

《数据有道》第 14、15 章专门介绍非线性回归。

请大家特别注意，逻辑回归不但可以用来回归，也可以用来分类。

此外，大家会发现 *k*-NN、高斯过程算法完成的回归也都可以归类为非线性回归。

正则化

正则化是多元回归的重要技术之一。正则化能够降低多重共线性，有效解决模型过拟合问题，提高泛化能力。请大家注意正则化和范数的关系。

对这部分内容感到生疏的话，请回顾《数据有道》第 2 章。此外，《数据有道》第 13 章还介绍如何从贝叶斯推断角度理解正则化。

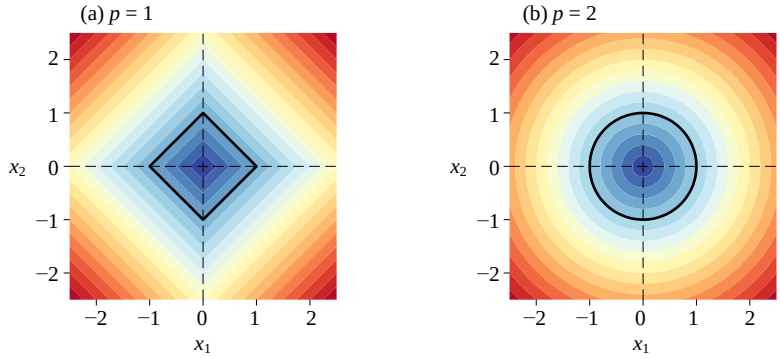

图 7. 两个范数

基于降维算法的回归

《数据有道》第 18、19 章特别介绍两种基于主成分分析的回归方法——正交回归、主元回归。

平面上，最小二乘法线性回归 OLS 仅考虑纵坐标方向上误差，如图 8 (a) 所示；而正交回归 TLS 同时考虑横纵两个方向误差，如图 8 (b) 所示。

主元回归的因变量则来自于主成分分析结果。

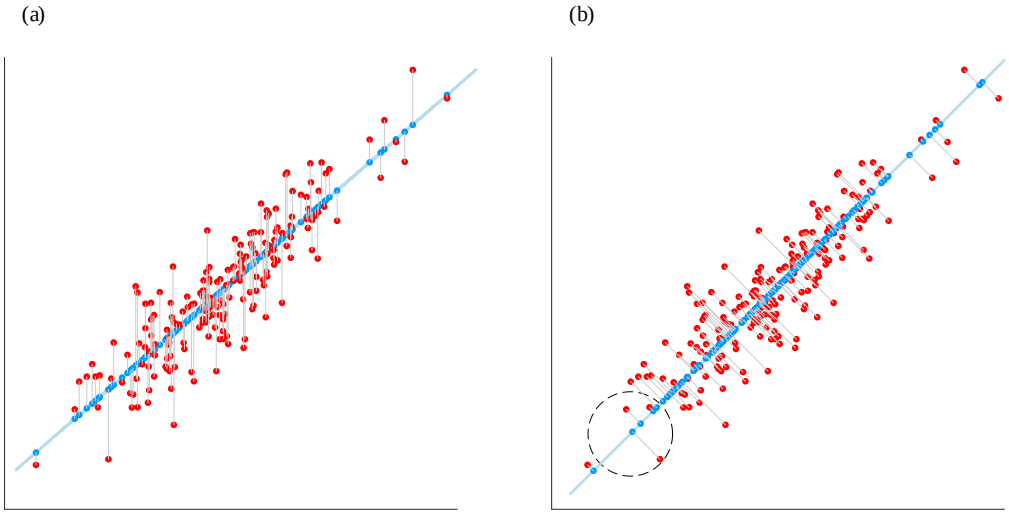

图 8. 对比 OLS 和 TLS 线性回归，图片来自《数据有道》第 18 章

基于分类算法的回归

实际上，监督学习的很多算法都兼顾分类、回归两项任务，比如逻辑回归、*k*-NN、支持向量机、高斯过程等等。本章下一节给出一个 *k*-NN 回归的例子。

11.2 *k*-NN 回归：非参数回归

本书第 2 章介绍的 *k*-NN 分类算法针对离散标签，比如 C_1 (红色 ●) 和 C_2 (蓝色 ●)。当输出值 y 为连续数据时，监督学习便是回归问题。本节讲解如何利用 *k*-NN 求解回归问题。

对分类问题，一个查询点的标签预测是由它附近 k 个近邻中占多数的标签决定；同样，某个查询点的回归值，也是由其附近 k 个近邻的输出值决定。

采用等权重条件下，查询点 $\boldsymbol{q}$ 回归值 $\hat{y}$ 可以通过下式计算获得：

$$\hat{y}(\boldsymbol{q}) = \frac{1}{k} \sum_{i \in kNN(\boldsymbol{q})} y^{(i)} \tag{1}$$

其中，$k\text{NN}(\boldsymbol{q})$ 为查询点 $\boldsymbol{q}$ 的 k 个近邻构成的集合。

举个例子

如图 9 所示，当 $k = 3$ 时，查询点 Q 附近三个近邻 $\boldsymbol{x}^{(1)}$、$\boldsymbol{x}^{(2)}$ 和 $\boldsymbol{x}^{(3)}$ 标记为蓝色 ●。这三个点对应的连续输出值分别为 $y^{(1)}$、$y^{(2)}$ 和 $y^{(3)}$。根据 (1) 计算 $y^{(1)}$、$y^{(2)}$ 和 $y^{(3)}$ 平均值，得到查询点回归预测值 $\hat{y}$：

$$\hat{y}(\boldsymbol{q}) = \frac{1}{3}\left(y^{(1)} + y^{(2)} + y^{(3)}\right) = \frac{1}{3}(5+3+4) = 4 \tag{2}$$

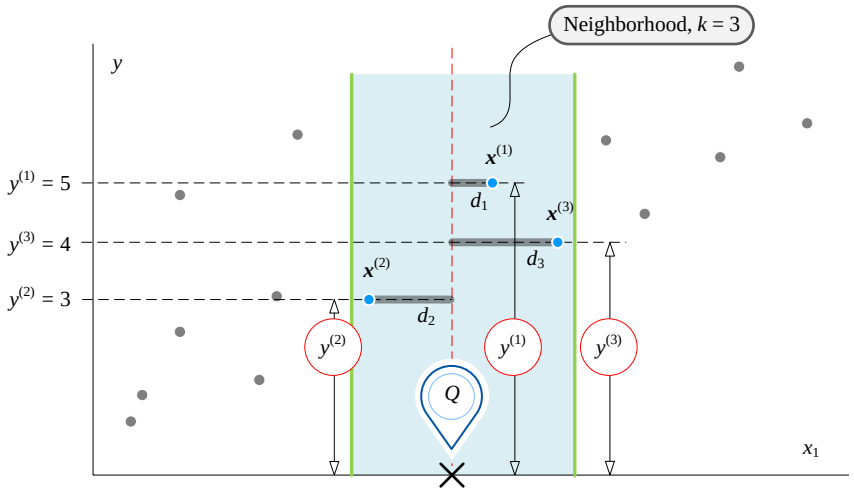

图 9. *k*-NN 回归算法原理

函数

`sklearn.neighbors.KNeighborsRegressor` 函数完成 *k*-NN 回归问题求解。默认等权重投票，`weights = 'uniform'`。

如果 *k*-NN 回归中考虑近邻投票权重，查询点 $\boldsymbol{q}$ 回归值 $\hat{y}$ 可以通过下式计算获得：

$$\hat{y}(\boldsymbol{q}) = \frac{1}{\sum_{i \in kNN(\boldsymbol{q})} w_i} \sum_{i \in kNN(\boldsymbol{q})} w_i y^{(i)} \tag{3}$$

类似 *k*-NN 分类，`weights = 'distance'`设置样本数据权重与到查询点距离成反比。

图 10 所示为利用 *k*-NN 回归得到的不同种类鸢尾花花萼长度 x_1 和花萼宽度 x_2 回归关系。花萼宽度 x_2 相当于 (3) 中 y。图 10 (a) 采用等权重投票，图 10 (b) 中投票权重与查询点距离成反比。

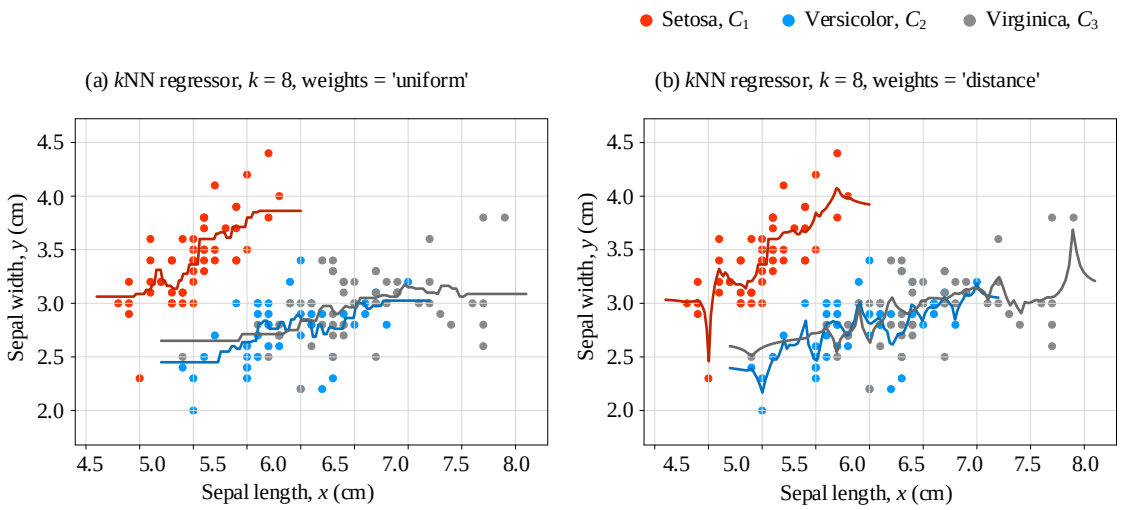

图 10. *k*-NN 回归，不同种类鸢尾花花萼长度和花萼宽度回归关系

代码 Bk7_Ch11_02.py 完成 *k*-NN 回归，并绘制图 10 两幅图像。

回归算法是一种机器学习技术，用于预测一个或多个连续型变量的值。其中 OLS 线性回归是最常见的回归算法之一，它基于最小二乘法来建立一个线性方程，使得预测值和实际值之间的差异最小化。

贝叶斯回归是一种基于贝叶斯定理的概率模型，能够通过考虑先验知识来提高预测准确性。非线性回归能够建立非线性模型来更准确地预测目标变量。正则化方法包括 L1 和 L2 正则化，可以在模型训练时加入惩罚项来避免过拟合。

基于 PCA 回归是一种使用主成分分析来降维和建立回归模型的技术。基于分类的回归方法如支持向量机、*k*NN 和决策树，通过将回归问题转化为分类问题来预测目标变量的值。每种回归算法都有其优缺点和适用场景，需要根据具体情况选择合适的方法。

利用支持向量机完成回归，请参考：

https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html

利用决策树完成回归，请参考：

https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html