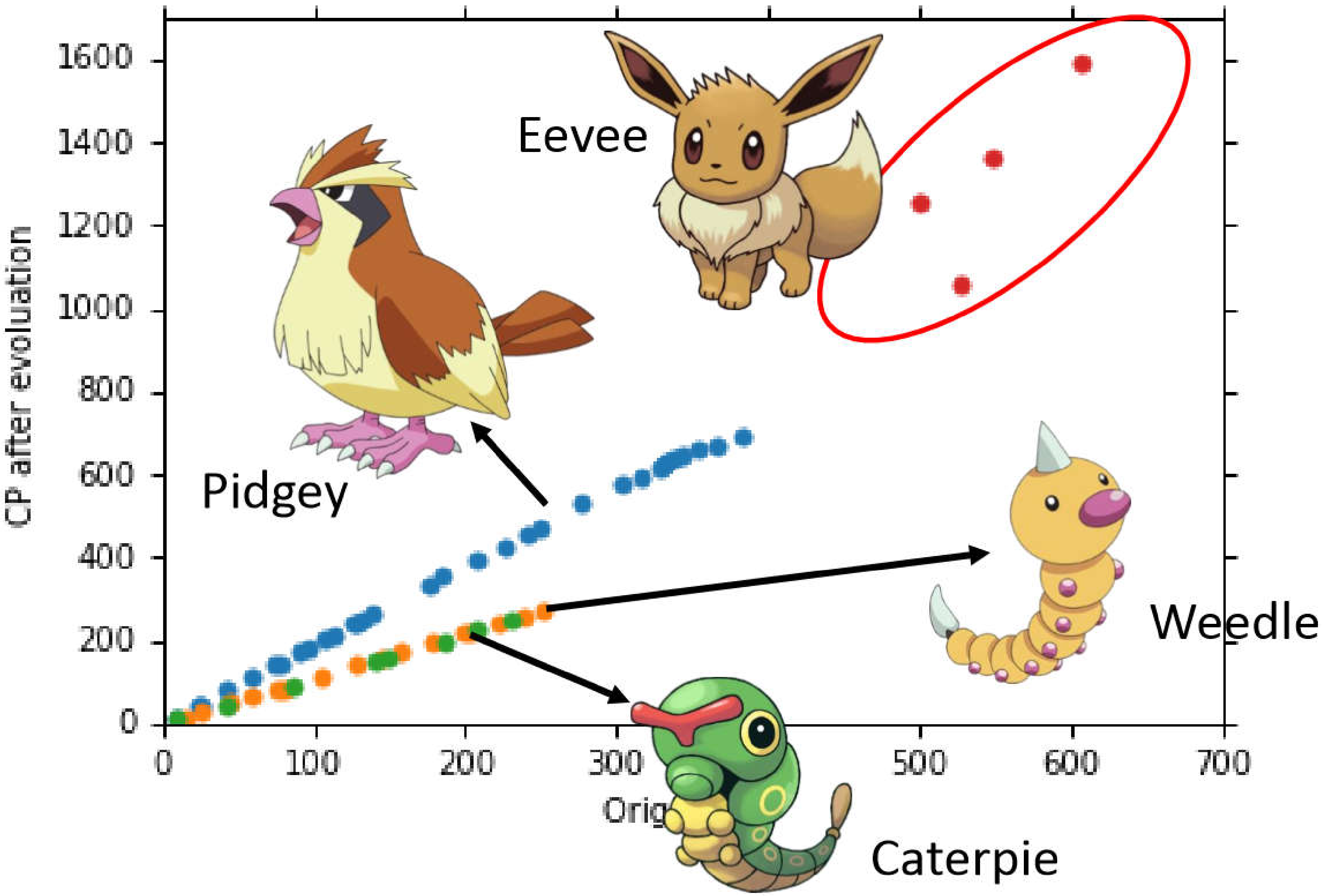
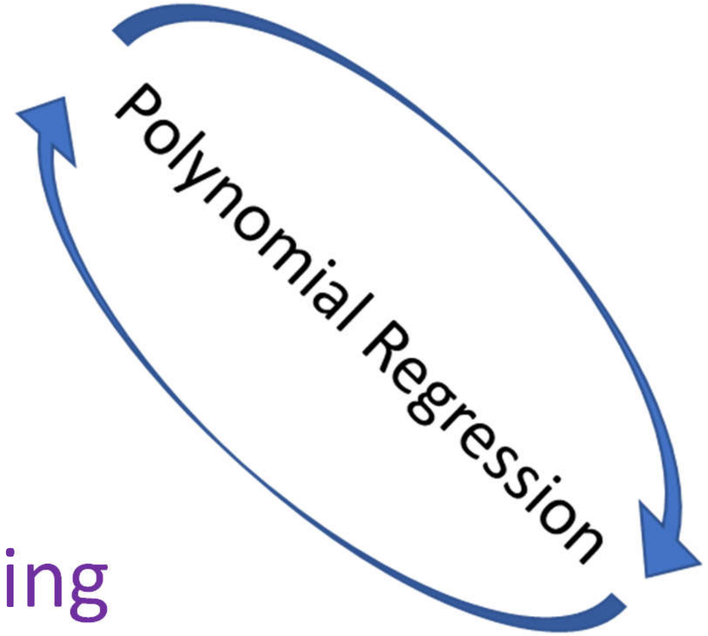
Linear Regression



Linear Regression

- Step 1: Model: Linear Regression
- Step 2: Goodness of Function: MSE
- Step 3: Gradient Descent
- How's the results?
  - –Generalization: Underfitting/Overfitting
- Back to Step 1
  - More Data → Indictive Function → More Features → Overfitting
- Back to Step 2
  - Regularization



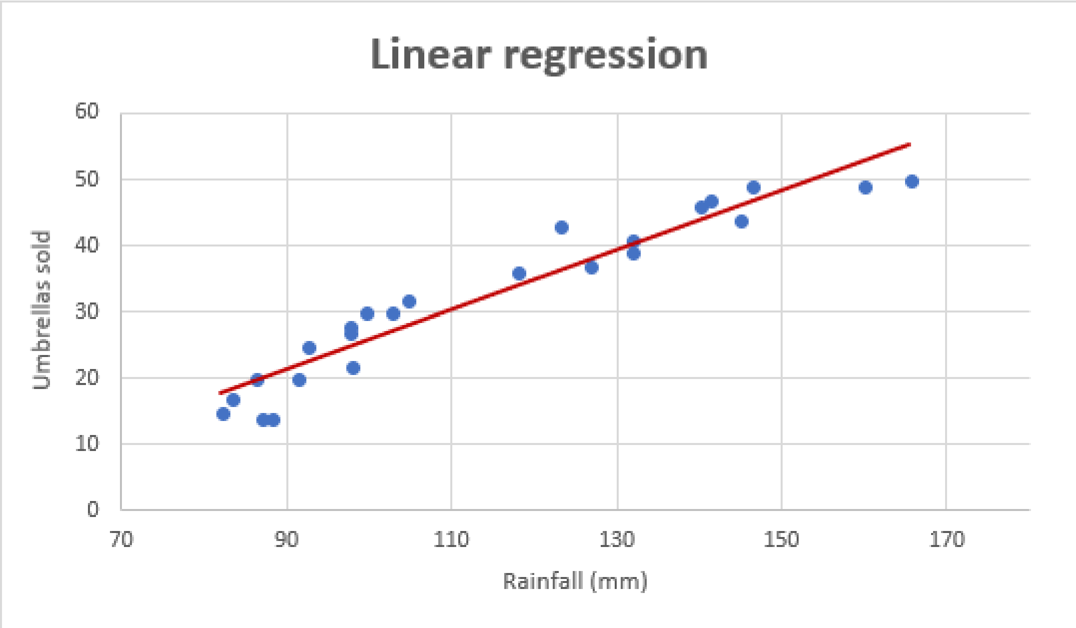
Linear Regression

- 经验风险最小化
  - 矩阵表示
  - 几何意义
  - 梯度下降法
- →正态分布+MLE
- 结构风险最小化
- →正态分布+MAP

线性回归详解

Peng Li
https://simplelp.github.io/
2019/06/04

本文为 机器学习-白板推导系列（三）-线性回归（Linear Regression） 的学习笔记，具体内容请参考原视频，感谢UP主的分享。



一、最小二乘法的矩阵表示与几何意义

1. 数据集

$$D = (x_1, y_1), (x_2, y_2), \dots (x_n, y_n), x_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, 2, \dots N$$

$$X = (x_1, x_2, \dots, x_N)^T_{N \times p}$$

$$Y = (y_1, y_2, \dots, y_N)^T_{N \times 1}$$

2. 最小二乘估计的矩阵表示

线性回归拟合函数为

$$f(w,b)=w^Tx+b, w\in\mathbb{R}^p, b\in\mathbb{R}$$

令$w=(w^1,w^2,\ldots,w^p,b)^T, x=(x^1,x^2,\ldots,x^p,1)^T$,有

$$f(w)=w^Tx, w\in\mathbb{R}^{p+1}$$

最小二乘估计损失函数为

$$\begin{aligned}
L(w) &= \sum_{i=1}^{N}||w^Tx_i-y_i||^2 \\
&= \sum_{i=1}^{N}(w^Tx_i-y_i)^2 \\
&= (w^Tx_1-y_1, w^Tx_2-y_2,\ldots,w^Tx_N-y_N)\begin{pmatrix} w^Tx_1-y_1 \\ w^Tx_2-y_2 \\ \ldots \\ w^Tx_N-y_N \end{pmatrix} \\
&= (w^TX^T-Y^T)(w^TX^T-Y^T)^T \\
&= (w^TX^T-Y^T)(Xw-Y) \\
&= w^TX^TXw-w^TX^TY-Y^TXw-Y^TY \\
&= w^TX^TXw-2w^TX^TY-Y^TY
\end{aligned}$$

因为最小二乘损失函数是关于w的凸函数，直接对损失函数求导

$$\begin{aligned}
\frac{\partial L(w)}{\partial w} &= \frac{\partial(w^TX^TXw-2w^TX^TY-Y^TY)}{\partial w} \\
&= 2X^TXw-2X^TY
\end{aligned}$$

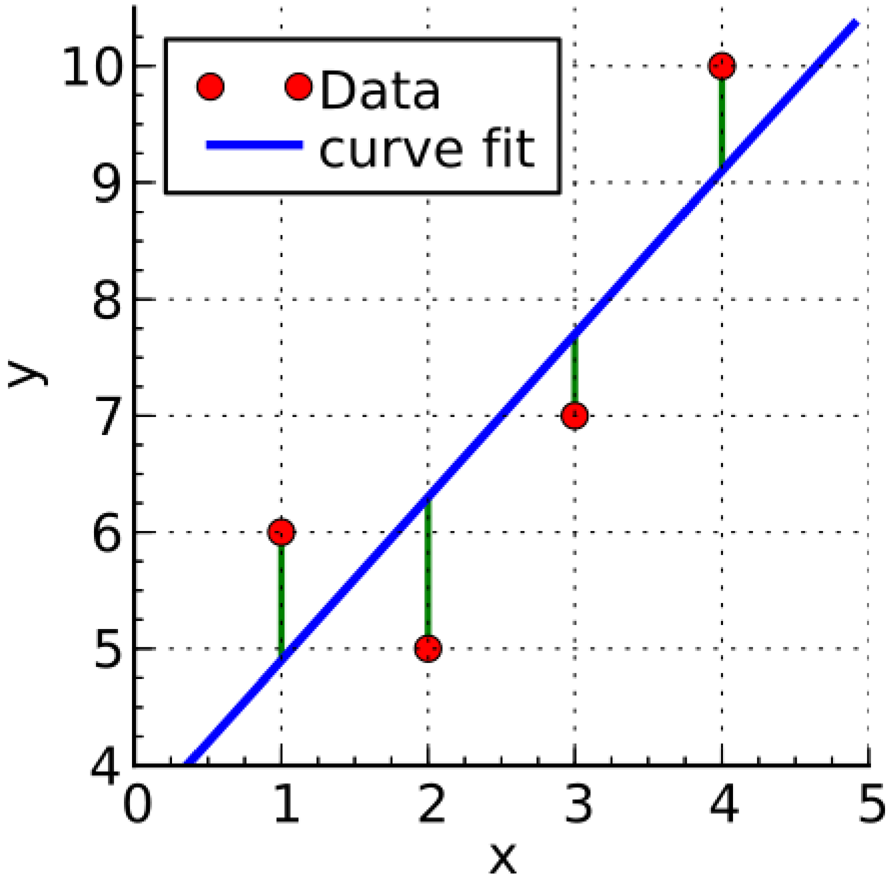
令导数等于0得

$$w=(X^TX)^{-1}X^TY$$

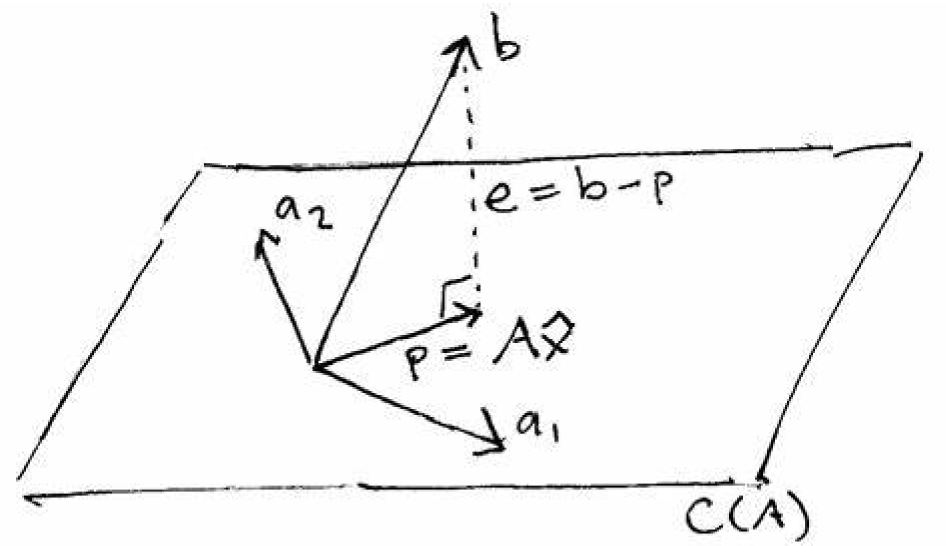
注意，此处的前提是X^TX可导。$(X^TX)^{-1}X^T$被称为伪逆。

3. 最小二乘法的几何意义

(1) 欧式距离角度



(2) 投影角度



从投影角度来看，要最小化的函数 $L(w) = (Xw - Y)^2$ 可以看作n维空间中，让

$$Y = (y_1, y_2, \ldots, y_N)^T$$

这个向量与

$$
\begin{aligned}
Xw = & \quad (x_1, x_2, \ldots, x_N)^T_{N\times p}(w_1, w_2, \ldots 2_p)^T_{p\times 1} \\
= & \quad (p_1, p_2, \ldots, p_p)(w_1, w_2, \ldots 2_p)^T_{p\times 1}
\end{aligned}
$$

这个向量的距离最小，其中 $(p_1, p_2, \ldots, p_p)(w_1, w_2, \ldots 2_p)^T_{p\times 1}$ 可以看作由$p_1, p_2, \ldots, p_p$ 构成的超平面，那么最小距离就应该是Y在这个超平面中的投影，设为$X\beta$，因为垂直关系，有

$$X^T(Y - X\beta) = 0$$

$$\beta = (X^TX)^{-1}X^TY$$

二、最小二乘法—贝叶斯学派视角

贝叶斯学派认为，因为真实数据中存在噪音，设真实的 y 满足

$$y = f(w) + \varepsilon = w^Tx + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

那么

$$p(y|w;x) \sim N(w^Tx, \sigma^2)$$

$$p(y|w;x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-w^Tx)^2}{2\sigma^2}}$$

因此可以利用最大似然估计(MLE)估计 w 的值

$$\hat{w}_{MLE} = argmax(logP(Y|X;w))$$

因为样本都是独立同分布的，所以

$$
\begin{aligned}
\hat{w}_{MLE} = & \quad arg\max_w(logP(Y|X;w)) \\
= & \quad arg\max_w(log\prod_{i=1}^{N}P(y_i|x_i;w)) \\
= & \quad arg\max_w\sum_{i=1}^{N}logP(y_i|x_i;w) \\
= & \quad arg\max_w\sum_{i=1}^{N}log(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-w^Tx_i)^2}{2\sigma^2}}) \\
= & \quad arg\max_w\sum_{i=1}^{N}(-\frac{(y_i-w^Tx_i)^2}{2\sigma^2} - log(\sqrt{2\pi}\sigma))
\end{aligned}
$$

不考虑常数，上式等价于

$$\hat{w}_{MLE} = arg\,\min_{w} \sum_{i=1}^{N}(y_i - w^T x_i)^2$$

综上，**线性回归的最小二乘估计等价于噪声是Gauss分布的最大似然估计**

三、正则化

1. 过拟合

通过最小二乘估计，我们得到线性回归的参数为

$$w = (X^T X)^{-1} X^T Y$$

但是这个的前提是 $X_{p\times N}^T X_{N\times p}$ 是可逆的，如果X的列向量线性无关，$X^T X$是可逆的（参考）。但是在 $N \leq q$时，N的列向量往往是线性相关的，导致$X^T X$并不可逆，这样就无法由上式得到w（其实，这里的w有很多种情况）。从另一个角度讲，是因为参数过多，出现了**过拟合**的状况。

解决过拟合的策略：

- 增加数据
- 特征选择/特征提取（PCA）
- 正则化（对w进行约束）

2. 正则化

正则化的通用表示为

$$arg\,\min_{w}[L(w) + \lambda P(w)]$$

$P(w) = \|w\|_1$时，成为Lasso回归，会产生比较多的为0的参数，是一种特征选择方法。但是往往不容易计算，因此人们往往采用$P(w) = \|w\|_2$的方式，使用这种正则化方式的线性回归称为岭回归（Ridge Regression），也称作权值衰减

$$arg\,\min_{w}(L(w) + \lambda \|w\|_2)$$

下面推导w的矩阵表示，令

$$\begin{aligned} J(w) = &\quad w^T X^T X w - 2w^T X^T Y - Y^T Y + \lambda w^T w \\ = &\quad w^T (X^T X + \lambda I) w - 2w^T X^T Y - Y^T Y \end{aligned}$$

则

$$\frac{\partial J(w)}{\partial w} = 2(X^T X + \lambda I)w - 2X^T Y = 0$$

得

$$\hat{w} = (X^TX + \lambda I)^{-1}X^TY$$

这样 $X^TX + \lambda I$ 就可逆了

四、正则化—贝叶斯学派视角

贝叶斯学派认为，w的先验分布为

$$w \sim N(0, \sigma_0^2)$$

即

$$p(w) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{||w||^2}{2\sigma_0^2}}$$

根据贝叶斯定理

$$p(w|y) = \frac{p(y|w)p(w)}{p(y)}$$

同时，根据第二节的假设

$$p(y|w;x) \sim N(w^Tx, \sigma^2)$$

$$p(y|w;x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-w^Tx)^2}{2\sigma^2}}$$

利用最大后验概率估计

$$\begin{aligned}
\hat{w}_{MAP} &= \underset{w}{arg\,max}\, P(w|Y;X) \\
&= \underset{w}{arg\,max} \frac{P(Y|X;w)P(w)}{P(Y|X)} \\
&\sim \underset{w}{arg\,max}\, P(Y|X;w)P(w) \\
&= \underset{w}{arg\,max} \prod_{i=1}^{N} p(y_i|x_i;w)p(w) \\
&\sim \underset{w}{arg\,max}(\sum_{i=1}^{N} -\frac{(y_i - w^Tx)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_0^2}) \\
&\sim \underset{w}{arg\,min}(\sum_{i=1}^{N}(y_i - w^Tx)^2 + \frac{\sigma^2}{\sigma_0^2}||w||^2)
\end{aligned}$$

综上,**正则化的最小二乘估计等价于噪声是Gauss分布、权重w的先验分布是Gauss分布的最大后验概率估计**

五、思考

- 为什么最小二乘损失函数是w的凸函数?

延伸阅读

- 正态分布的前世今生 (上)
- 正态分布的前世今生 (下)
- 什么是龙格现象(Runge phenomenon)？如何避免龙格现象?