第八讲 回归模型

牟克典

2021年5月5日

概要

回归问题是监督学习的主要学习任务之一。

- 不妨设输入空间$\mathcal{X}$就是特征空间且$\mathcal{X} = \mathbf{R}^n$, $\mathcal{D}$ 为$\mathcal{X}$上的未知概率分布；
- 输出空间$\mathcal{Y}$是$\mathbf{R}$的一个可测子集,且目标回归（标记）函数为$f : \mathcal{X} \mapsto \mathcal{Y}$.
- 设训练样本集$T = \{(x_i, y_i)\}_{i=1}^N$,其中$x_1, x_2, \cdots, x_N$为依据$\mathcal{D}$分布独立同分布抽样得到,且对所有$1 \leq i \leq N$, $y_i = f(x_i)$.
- 回归学习任务的目标就是基于训练样本集T从假设空间$\mathcal{H} = \{h|h : \mathcal{X} \mapsto \mathcal{Y}\}$中学得回归函数

$$\hat{h} : \mathcal{X} \mapsto \mathcal{Y},$$

- 并基于学得的回归函数$\hat{h}$对输入实例x对应的输出值作出如下预测：

$$\hat{y} = \hat{h}(x).$$

- 在评价回归学习算法性能时希望$\hat{h}$对x的输出值的预测$\hat{y} = \hat{h}(x)$与$y = f(x)$之间的差异幅度$|\hat{y} - y|$尽可能小而不是强求$\hat{y} = y$.
- 因此在回归学习任务中，我们使用的损失函数通常都与$|\hat{y} - y|$有关.
- 用$L : \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}$表示度量回归误差的损失函数.
  - 平方损失函数L_2: 即对任意$y, y' \in \mathcal{Y}$,

$$L_2(y, y') = |y - y'|^2.$$

  - L_p $(p \geq 1)$损失函数: 即对任意$y, y' \in \mathcal{Y}$,

$$L_p(y, y') = |y - y'|^p,$$

- 为了简单起见将$L(y, y')$写成$L(|y - y'|)$的形式.

- 给定损失函数L
  - 定义假设$h \in \mathcal{H}$在训练样本集T上的经验误差$\hat{R}(h)$为

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(x_i)).$$

  - 定义h(相对于f)的泛化误差$R(h)$为

$$R(h) = E_{x \sim \mathcal{D}}[L(h(x), f(x))]$$

- 在经验误差相同的情况下，我们更关心计算学习复杂度（如Rademacher复杂度）更小的假设集.
- 由线性函数构成的假设集的计算学习复杂度相对比较低.

概要

- 考虑线性假设集

$$\mathcal{H}=\left\{h(x)=w\cdot x+b|w=(w^{(1)},w^{(2)},\cdots,w^{(n)})^T\in\mathbf{R}^n,b\in\mathbf{R}\right\}$$

其中每个假设（线性回归模型(w,b)）

$$h(x)=w\cdot x+b$$

表示输入x的输出值是特征向量x的线性函数.

- 线性回归算法的目标就是在$\mathcal{H}$中选择具有最小均方误差的线性假设.
- 给定训练样本集$T=\{(x_i,y_i)\}_{i=1}^N$，则线性回归算法对应于下面的最优化问题：

$$\min_{(w,b)}\frac{1}{N}\sum_{i=1}^{N}(w\cdot x_i+b-y_i)^2$$

- 对$w = (w^{(1)}, w^{(2)}, \cdots, w^{(n)})^T$进行扩充，把$b$作为其最后一维元素得到

$$W = (w^{(1)}, w^{(2)}, \cdots, w^{(n)}, b)^T.$$

- 同时对每个$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T$进行扩充，把1作为其最后一维元素得到

$$x_i^+ = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}, 1)^T.$$

- 进而构造以每个x_i^+为列向量的矩阵

$$X = [x_1^+, x_2^+, \cdots, x_N^+]$$

- 以所有样本标记y_i为元素的向量

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

- 上述最优化问题的目标函数

$$\frac{1}{N}\sum_{i=1}^{N}(w \cdot x_i + b - y_i)^2$$

可以写成

$$F(W) = \frac{1}{N} \parallel X^T W - Y \parallel_2^2$$

- 则线性回归算法对应的最优化问题可以表示成

$$\min_W F(W).$$

令$\frac{\partial F(W)}{\partial W} = 0$则得到

$$\frac{2}{N}X(X^TW - Y) = 0$$

即

$$XX^TW = XY.$$

- 如果XX^T可逆，则W有惟一解

$$\hat{W} = \left(XX^T\right)^{-1}XY.$$

- 如果XX^T不可逆，则

$$XX^T W = XY.$$

可以求解出一组$\hat{W}$，一般选择范数最小的解

$$\hat{W} = \left(XX^T\right)^{\dagger} XY,$$

其中$\left(XX^T\right)^{\dagger}$为$XX^T$的Moore-Penrose广义逆.

在得到对参数W的估计$\hat{W}$后，得到相应的线性回归模型

$$\hat{h}(x) = \hat{W} \cdot x^{+}.$$

特别地，如果特征向量的维数$n = 1$，且$\{x_i\}$不全相同，则

- $h(x) = wx + b$为直线.
- $F(W) = \frac{1}{N}\sum_{i=1}^{N}(y_i - wx_i - b)^2$.
- $|y_i - wx_i - b|$为点(x_i, y_i)沿平行于y轴方向到直线$h(x) = wx + b$的距离.
- $\sum_{i=1}^{N}(y_i - wx_i - b)^2$为这些距离的平方和.
- 此时最小化目标函数相当于求一条直线使得样本点沿y轴方向距离该直线平均最近.
- 目标函数是关于w和b的二元二次函数，其最小值唯一存在.
- $\{x_i\}_{i=1}^{N}$和$\{y_i\}_{i=1}^{N}$的样本均值分别为$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$和$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

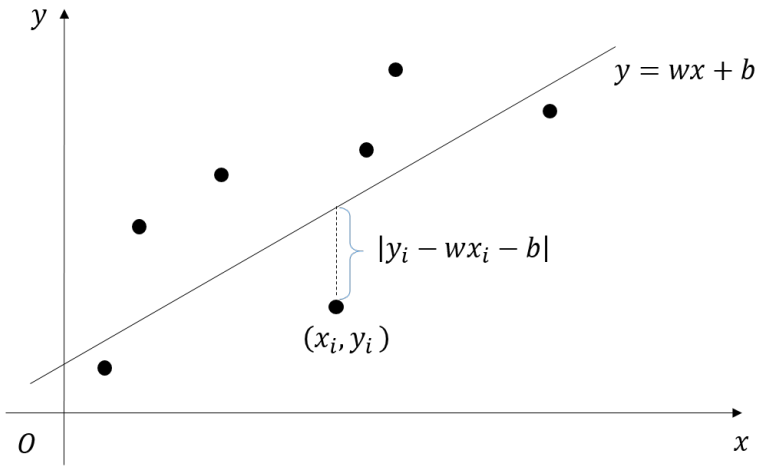

Figure: 回归直线

通过最小化$F(W)$得到

$$\hat{W} = \begin{bmatrix} \hat{w} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{i=1}^{N} y_i x_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} \\ \bar{y} - \hat{w}\bar{x} \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \\ \bar{y} - \hat{w}\bar{x} \end{bmatrix}.$$

如果我们引进

- $\{x_i\}_{i=1}^N$的样本方差

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- 样本协方差

$$s_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}),$$

则$\hat{W} = \begin{bmatrix} \hat{w} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \frac{s_{xy}}{s_x^2} \\ \bar{y} - \hat{w}\bar{x} \end{bmatrix}$.

由于目标函数是二次函数的缘故，上述方法被称为最小二乘法。

对训练样本(x_i, y_i)来说,我们以回归直线$(\hat{w}, \hat{b})$上的

$$\hat{y}_i = \hat{w}x_i + \hat{b}$$

作为y_i的预测值, 则相应的预测误差(也称残差)为

$$\hat{\epsilon}_i = y_i - \hat{w}x_i - \hat{b}.$$

定义残差平方和

$$Q(\hat{w}, \hat{b}) = \sum_{i=1}^{N} \hat{\epsilon}_i^2.$$

显然，

$$y_i = \hat{w}x_i + \hat{b} + \hat{\epsilon}_i.$$

如果考虑误差的随机性，(x_i, y_i)满足一元线性回归模型

$$Y_i = wx_i + b + \epsilon_i, \quad i = 1, 2, \cdots, N.$$

- 这里将输入变量x_i不做随机变量处理，看作常量.
- 参数w和b称为回归系数.
- $\{\epsilon_i\}_{i=1}^N$是独立同分布的随机变量，服从高斯分布$N(0, \sigma^2)$，其中参数σ^2代表了随机误差的强弱.
- Y_i服从高斯分布$N(wx_i + b, \sigma^2)$.

给定训练数据集$T = \{(x_i, y_i)\}_{i=1}^N$，得到回归系数的最大似然估计如下：

$$\begin{aligned} \hat{w} &= \frac{s_{xy}}{s_x^2}, \\ \hat{b} &= \bar{y} - \hat{w}\bar{x} \end{aligned}$$

对σ^2来说，其最大似然估计为

$$\hat{\sigma}_2 = \frac{1}{N} Q(\hat{w}, \hat{b}).$$

但这不是σ^2的无偏估计，我们一般采用σ^2的如下无偏估计：

$$\hat{\sigma}^2 = \frac{1}{N-2} Q(\hat{w}, \hat{b}).$$

概要

- 线性回归算法通过最小化残差平方和来对参数进行估计.
- 基于最小二乘估计得到的线性回归模型具有预测偏差小而方差大的特点.
- 为了增强模型对数据扰动的容忍能力，需要对模型的预测偏差和方差进行权衡.
- 以权重参数w的函数$J(w)$作为惩罚项，通过最小化正则化的残差平方和

$$\min_{w,b} \sum_{i=1}^{N} (y_i - b - w \cdot x_i)^2 + \lambda J(w)$$

来改善线性回归模型的泛化性能，这里$\lambda \geq 0$用于权衡$J(w)$和残差平方和.

岭回归(Ridge Regression)方法就是通过所谓的收缩方法来缩减某些维特征对应的权重系数来降低预测方差的线性回归方法.

- 采用权重参数向量w的L_2范数的平方作为惩罚项，通过最小化如下目标函数

$$F_r(w,b)=\sum_{i=1}^{N}(y_i-b-\ w\cdot x_i)^2+\lambda\parallel w\parallel_2^2$$

来估计权重参数，即

$$(\hat{w},\hat{b})=\underset{w,b}{\operatorname{argmin}}\,F_r(w,b).$$

这里$\lambda\geq 0$用于权衡正则化项和残差平方和.

令

$$\frac{\partial F_r(w,b)}{\partial b} = 0$$

则得到

$$b = \frac{1}{N}\sum_{i=1}^{N}(y_i - w \cdot x_i) = \bar{y} - w \cdot \bar{x}.$$

代入$F_r(\mathbf{w}, b)$中得到

$$F_r(w,b) = \sum_{i=1}^{N}((y_i - \bar{y}) - w \cdot (x_i - \bar{x}))^2 + \lambda \parallel w \parallel_2^2$$

这相当于

- 对输入x_i中心化为（$x_i - \bar{x}$）的情形下用$\bar{y}$估计b.
- 剩下的参数w可以基于中心化的输入数据采用没有偏置的岭回归模型来估计.

- 不妨设数据中心化已经完成，相应的数据矩阵为$X_{n\times N}$，则将最优化问题写成

$$\min_{w} \| X^T w - Y \|_2^2 + \lambda \| w \|_2^2 .$$

这里$\lambda > 0$是正则化系数.

- 将目标函数关于的w偏导数置为0则得到

$$2X(X^T w - Y) + 2\lambda w = 0$$

即$(XX^T + \lambda \mathbf{I})w = XY$, 这里$\mathbf{I}$是$n \times n$单位矩阵。

- 注意到$XX^T + \lambda \mathbf{I}$是可逆的，则得到岭回归模型参数的估计

$$\hat{w} = (XX^T + \lambda \mathbf{I})^{-1} XY.$$

显然岭回归算法给出的权值估计是输出值的线性组合.

- 可以采用如下的带约束的最优化问题来定义岭回归模型：

$$\min_{w,b} \sum_{i=1}^{N} (y_i - b - w \cdot x_i)^2$$
$$s.t. \quad \| w \|_2^2 \leq s,$$

这里s是一个正参数.

- 从这个等价定义来看，约束条件$\| w \|_2^2 \leq s$将w的可能取值限制在$\mathbf{0}$为中心，半径为s的n维球体内.

特别地，对$n = 2$来说，

- w的可能取值限制在**0**为中心，半径为s的球面内，
- 作为目标函数的残差平方和的等值线为椭圆形，二者相交的点即为权值向量的估计.

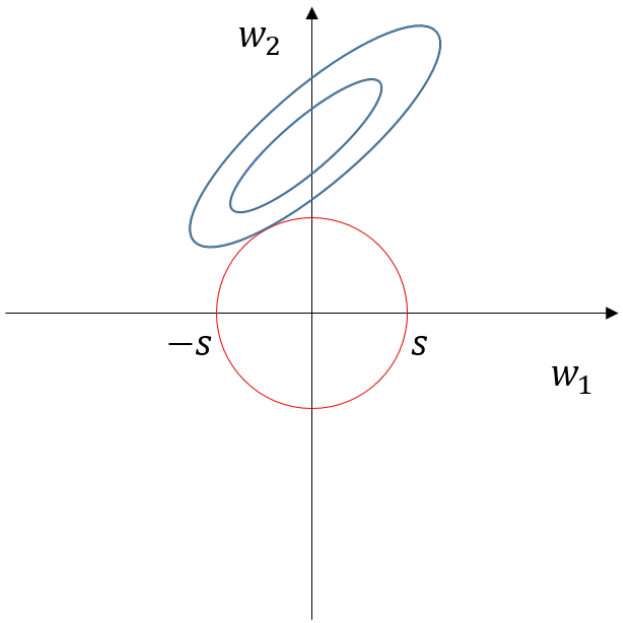

Figure: 岭回归模型的等价定义

概要

Lasso

Lasso(**L**east **a**bsolute **s**hrinkage and **s**election **o**perator)算法

- 引进权重参数向量w的L_1范数作为惩罚项.
- 通过最小化如下目标函数

$$F_l(w,b)=\sum_{i=1}^{N}(y_i-b-w\cdot x_i)^2+\lambda\parallel w\parallel_1$$

来估计权重参数，即

$$(\hat{w},\hat{b})=\underset{w,b}{\operatorname{argmin}}\,F_l(w,b).$$

这里$\lambda\geq 0$用于权衡正则化项和残差平方和.

- 也可以采用如下带约束的最优化问题来表示Lasso算法：

$$\min_{w,b} \sum_{i=1}^{N} (y_i - b - w \cdot x_i)^2$$
$$s.t. \quad \| w \|_1 \leq s,$$

这里s是一个正参数.

- 从Lasso算法的这个等价表示来看，约束条件$\| w \|_1 \leq s$相当于将w的可能取值限制在$\sum_{i=1}^{n} |w^{(i)}| \leq s$的$L_1$-球体内，以该区域和残差平方和的等值线的交点作为$w$的估计.

二维空间中的Lasso算法的刻画示意:

- 在二维空间中，L_1-球就是以**0**为中心，对角线分别在$w^{(1)}$和$w^{(2)}$轴上且都长为$2s$的菱形.

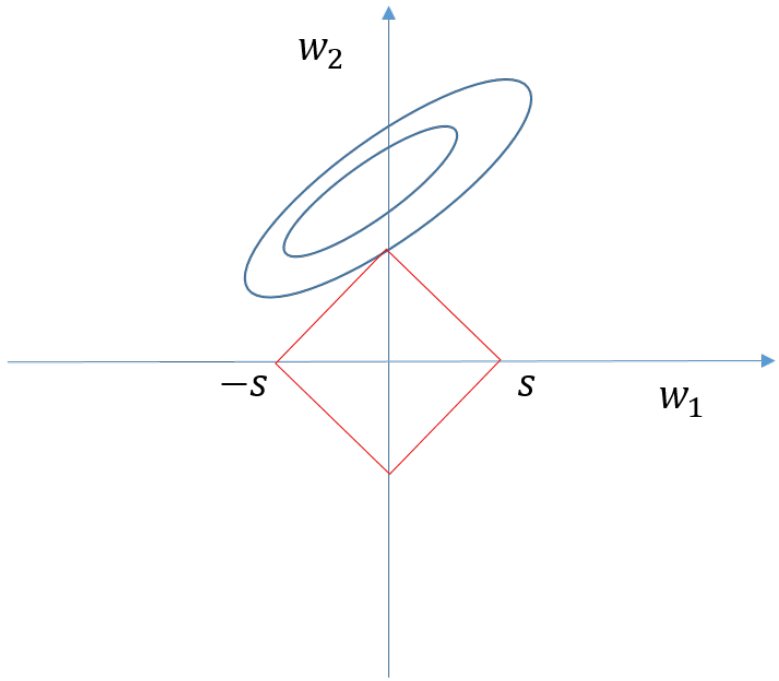

Figure: Lasso算法的等价表示

- Lasso算法对应的最优化问题没有闭式解.
- 但Efron等提出的LAR(Least Angle Regression)算法 可以对所有正则化参数λ求出对应的Lasso估计.
- 此外，还可以将Lasso算法对应的最优化问题转化为一个二次规划问题来求解:
  - 首先将w表示成如下形式

$$w = w_+ - w_-$$

其中$w_+ \geq 0$， $w_- \geq 0$，且对任一$i \in [1, n]$来说， $w_+^{(i)} = 0$或者$w_-^{(i)} = 0$.

  - 通过这种表示方式，可以将权值向量w的L_1 范数表示成如下没有绝对值的形式:

$$\| w \|_1 = \sum_{i=1}^{n} w_+^{(i)} + w_-^{(i)}$$

- 将上述两式代入Lasso算法对应的最优化问题，得到

$$\min_{w^+\geq 0, w^-\geq 0, b} \sum_{i=1}^{N}(y_i - b - (w^+ - w^-)\cdot x_i)^2 + \lambda \sum_{j=1}^{n} w_+^{(j)} + w_-^{(j)}$$

这是一个二次规划问题，可以利用任何一个有效的二次规划求解器来求解该问题.

- Lasso算法一方面与岭回归算法比较类似，压缩回归系数的估计接近0.
- 但L_1范数作为正则化项更倾向于比较稀疏的权值向量w，即非零元素比较少的权值向量.
- 这对数据特征的维数比较大的情形很有吸引力.

概要

- 在线性回归算法、岭回归算法和Lasso算法中，所采用的损失函数都是平方损失

$$L_2(y, h(x)) = (y - h(x))^2 = (y - b - w \cdot x)^2.$$

- 平方损失函数只有在回归函数预测完全正确的时候，即$y = h(x)$时损失$L_2(y, h(x)) = 0$.
- 支持向量回归模型则容忍回归函数$h(x)$与真实输出y之间最多有ϵ的偏差，即将损失为0 的情况放宽到$|y - h(x)| \leq \epsilon$的情形.

这相当于将认为被$h(x)$正确预测的样本点的范围由$h(x)$上的样本点扩展到如下图所示的一条以$h(x)$ 为中心，与y轴平行方向宽度为2ϵ 的间隔带中的样本点.

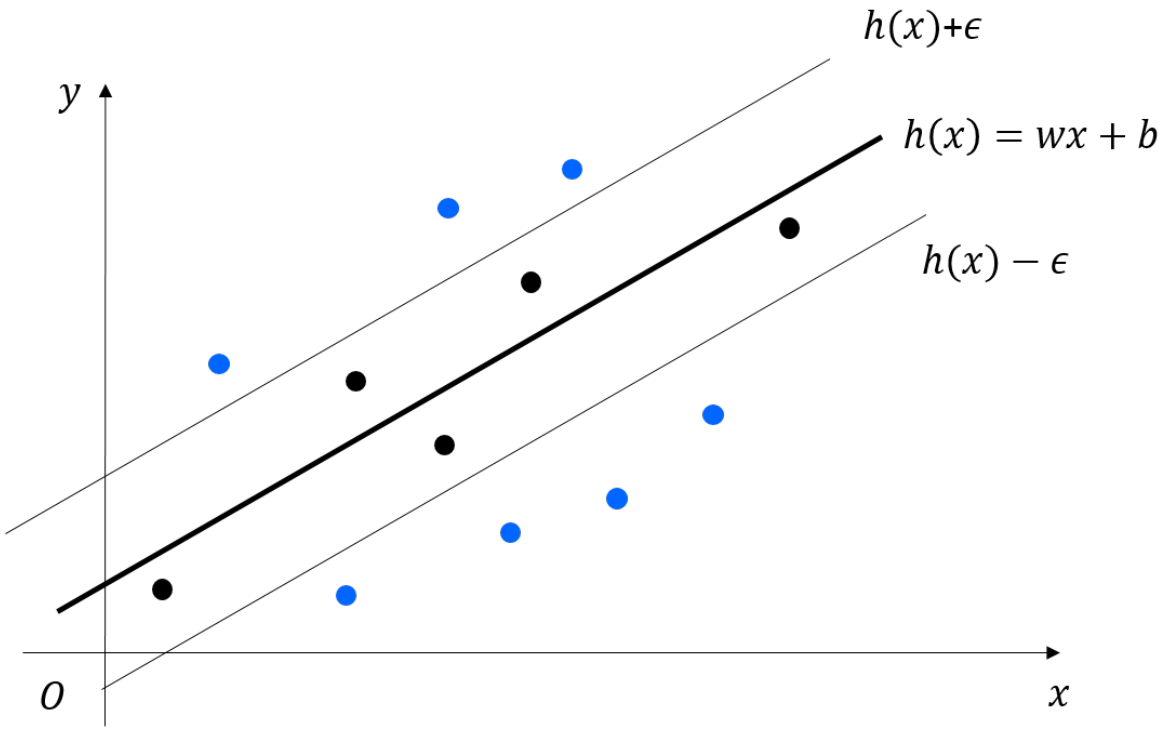

Figure: 支持向量回归

- ϵ不敏感损失函数(ϵ-insensitive loss)L_ϵ: 即对任意$y, y' \in \mathcal{Y}$,

$$L_\epsilon(|y - y'|) = \max(0, |y - y'| - \epsilon).$$

- 与$L_2(y, y')$相比， L_ϵ只有在y和y'之间的差异幅度超过ϵ时才有损失，而且以差异幅度超过ϵ 的部分作为损失度量.

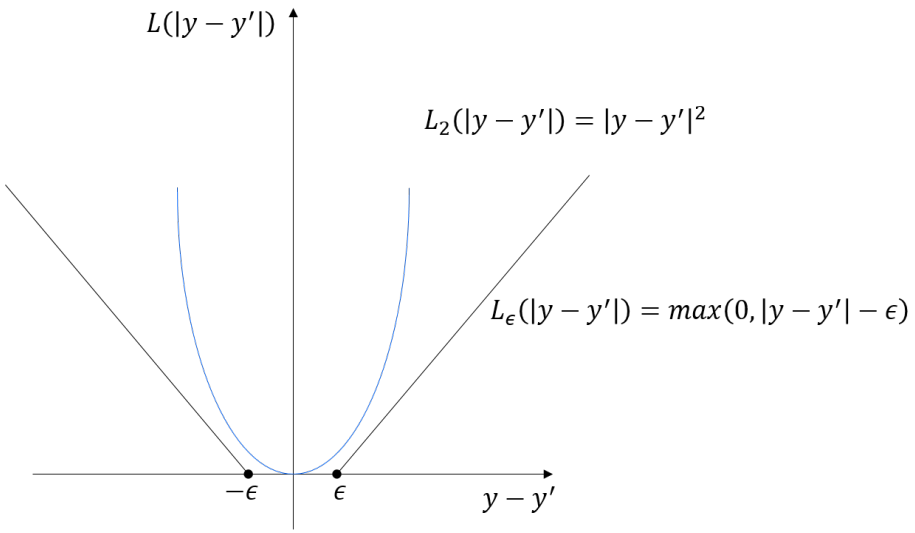

Figure: ϵ不敏感损失函数与平方损失函数

- 基于ϵ不敏感损失函数，我们将支持向量回归问题表示为：

$$\min_{w,b} \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{N} L_\epsilon(|y_i - w \cdot x_i - b|)$$

其中$C > 0$为正则化常数.

- 对每个训练样本(x_i, y_i)引入两个松弛变量$\xi_i \geq 0$ 和 $\xi_i' \geq 0$，将上述最优化问题重新表示为如下的等价形式：

$$\min_{w,b,\xi,\xi'} \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i')$$

$$\begin{aligned} s.t. \quad & (w \cdot x_i + b) - y_i \leq \epsilon + \xi_i, \\ & y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i', \\ & \xi_i \geq 0, \ \xi_i' \geq 0, \forall i \in [1, N]. \end{aligned}$$

这里$\xi = (\xi_1, \xi_2, \cdots, \xi_N)^T$ 和 $\xi' = (\xi_1', \xi_2', \cdots, \xi_N')^T$.

引入拉格朗日乘子α_i，α_i'，μ_i，μ_i'，$i \in [1, N]$，构造拉格朗日函数

$$L(w, b, \xi, \xi', \alpha, \alpha', \mu, \mu') = \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i')$$

$$+ \sum_{i=1}^{N} \alpha_i ((w \cdot x_i + b) - y_i - \epsilon - \xi_i)$$

$$+ \sum_{i=1}^{N} \alpha_i' (y_i - (w \cdot x_i + b) - \epsilon - \xi_i')$$

$$- \sum_{i=1}^{N} \mu_i \xi - \sum_{i=1}^{N} \mu_i' \xi'.$$

令$\frac{\partial L(w,b,\xi,\xi',\alpha,\alpha',\mu,\mu')}{\partial w}=0$得到

$$w=\sum_{i=1}^{N}(\alpha_i'-\alpha_i)x_i.$$

令$\frac{\partial L(w,b,\xi,\xi',\alpha,\alpha',\mu,\mu')}{\partial b}=0$得到

$$\sum_{i=1}^{N}(\alpha_i'-\alpha_i)=0.$$

令$\frac{\partial L(w,b,\xi,\xi',\alpha,\alpha',\mu,\mu')}{\partial \xi_i}=0$得到

$$C=\alpha_i+\mu_i.$$

令$\frac{\partial L(w,b,\xi,\xi',\alpha,\alpha',\mu,\mu')}{\partial \xi_i'}=0$得到

$$C=\alpha_i'+\mu_i'.$$

将上面四式代入拉格朗日函数，得到支持向量回归的对偶问题：

$$\max_{\alpha,\alpha'}(\alpha'-\alpha)\cdot Y-\epsilon(\alpha'+\alpha)\cdot \mathbf{1}-\frac{1}{2}(\alpha'-\alpha)^T G(\alpha'-\alpha)$$

$$s.t.\quad \mathbf{0}\le \alpha,\alpha'\le \mathbf{C}\wedge(\alpha'-\alpha)\cdot\mathbf{1}=0$$

其中$G=[(x_i\cdot x_j)]_{N\times N}$为Gram矩阵，$\mathbf{0}$，$\mathbf{1}$ 和 $\mathbf{C}$ 分别为元素全为 0，1 和 C 的向量.

不妨设$\hat{\alpha}$和$\hat{\alpha}'$为对偶问题的解，则我们可以得到支持向量回归模型权值系数向量的估计

$$\hat{w} = \sum_{i=1}^{N}(\hat{\alpha}_i' - \hat{\alpha}_i)x_i.$$

和支持向量机模型类似，偏置b的估计可以基于使得$0 < \hat{\alpha}_j < C$的x_j得到：

$$\hat{b} = y_j + \epsilon - \hat{w} \cdot x_j.$$

或者基于使得$0 < \hat{\alpha}_j' < C$的x_j得到：

$$\hat{b} = y_j - \epsilon - \hat{w} \cdot x_j.$$

当然在实际学习任务中，一般多计算出几个偏置b的估计，然后取其均值作为最终估计.

- 进一步得到支持向量回归模型

$$\begin{aligned}\hat{h}(x) &= \hat{w}\cdot x+\hat{b}\\ &= \sum_{i=1}^{N}(\hat{\alpha}_i'-\hat{\alpha}_i)x_i\cdot x+\hat{b}\end{aligned}$$

- 我们称$(\hat{\alpha}_i'-\hat{\alpha}_i)\neq 0$的样本为支持向量回归模型的支持向量.
- 支持向量这些点究竟位于什么区域?
- 由KKT条件，对任一$i\in[1,N]$，都有

$$\alpha_i((w\cdot x_i+b)-y_i-\epsilon-\xi_i)=0;$$
$$\alpha_i'(y_i-(w\cdot x_i+b)-\epsilon-\xi_i')=0.$$

- 注意到样本(x_i,y_i)只能在$h(x)$的一侧，因此约束$(w\cdot x_i+b)-y_i-\epsilon-\xi_i=0$ 和 $y_i-(w\cdot x_i+b)-\epsilon-\xi_i'=0$不能同时成立，故$\alpha_i$ 和α_i'至少其一为0.

- 如果样本(x_i, y_i)在以$h(x)$为中心、宽为2ϵ的间隔内，即

$$h(x_i) - \epsilon < y_i < h(x_i) + \epsilon,$$

则约束$(w \cdot x_i + b) - y_i - \epsilon - \xi_i = 0$ 和
$y_i - (w \cdot x_i + b) - \epsilon - \xi_i' = 0$ 都不成立，因此$\alpha_i = \alpha_i' = 0$.

- 这就意味着使得$(\hat{\alpha}_i' - \hat{\alpha}_i) \neq 0$的样本点只能在间隔之外.
- 如果ϵ越大，则在间隔内的样本点数量就相对比较大，那么支持向量回归算法得到的回归函数相对比较稀疏.
- ϵ的选择需要在预测的准确性和解的稀疏性之间进行权衡.

小结

- 回归问题概述
- 线性回归算法
- 岭回归
- Lasso算法
- 支持向量回归