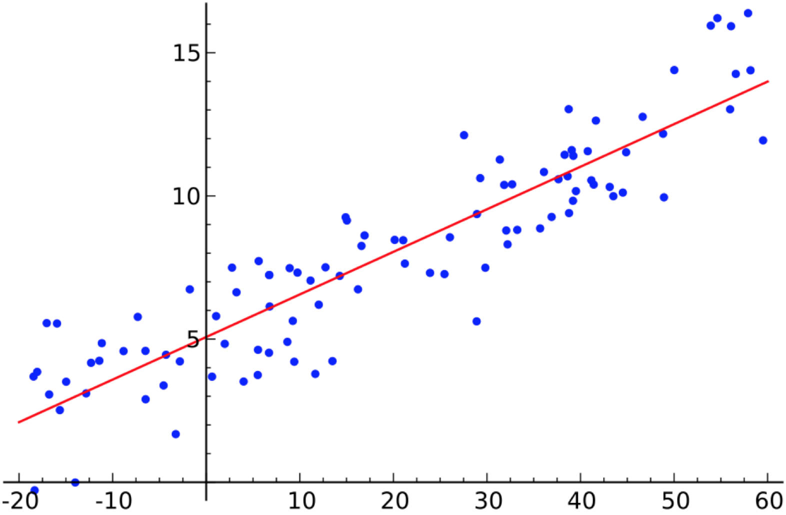
Regression

Evgeny Burnaev

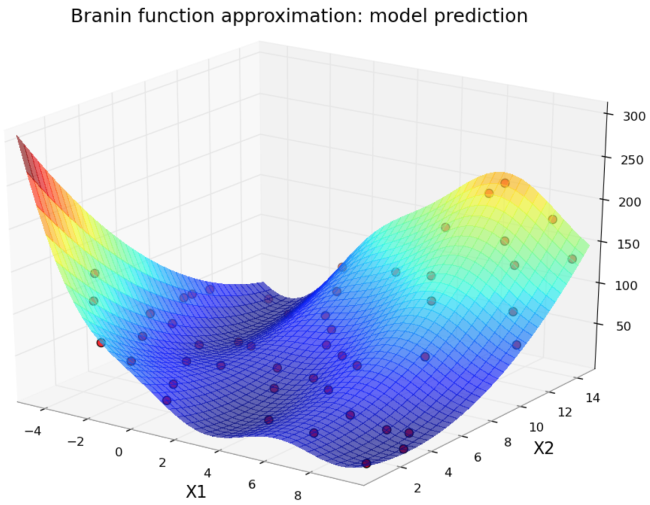
Skoltech, Moscow, Russia

Outline

Regression



Regression


Branin function approximation: model prediction
X1
X2
−4
−2
0
2
4
6
8
2
4
6
8
10
12
14
50
100
150
200
250
300

Regression Problem Statement

- **Training data**: sample drawn i.i.d. from set X according to some distribution D

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in X \times Y,$$

with $Y \subseteq \mathbb{R}$ is a measurable set

- **Loss function**: $L : Y \times Y \to \mathbb{R}_+$ a measure of closeness, e.g. $L(y, y') = (y - y')^2$ or $L(y, y') = |y - y'|^p$ for some $p \geq 1$

- **Problem**: find hypothesis $h : X \to \mathbb{R}$ in $\mathbb{H}$ with small generalization error w.r.t. target f

$$R_D(h) = \mathbb{E}_{\mathbf{x} \sim D}[L(h(\mathbf{x}), f(\mathbf{x}))]$$

Regression Problem

- Empirical error:

$$\hat{R}_D(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(\mathbf{x}_i), y_i)$$

- In much of what follows:
  — $Y = \mathbb{R}$ or $Y = [-M, M]$ for some $M > 0$
  — $L(y, y') = (y - y')^2$ is a mean squared error

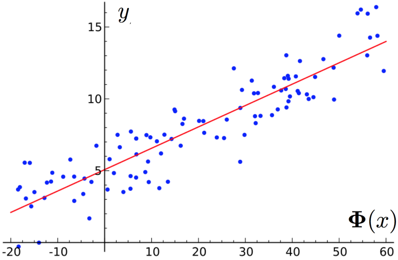
Linear Regression

- Feature mapping: $\Phi: X \to \mathbb{R}^N$
- Hypothesis set: linear functions

$$\mathbb{H} = \{\mathbf{x} \to \mathbf{w} \cdot \Phi(\mathbf{x}) + b : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$$

- **Optimization problem**: empirical risk minimization

$$\min_{\mathbf{w},b} F(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{w} \cdot \Phi(\mathbf{x}) + b - y_i)^2$$



Linear Regression: Solution I

- Rewrite objective function as $F(\mathbf{W}) = \frac{1}{m}\left\|\mathbf{X}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\right\|^2$, where $\mathbf{X} = \begin{bmatrix} \Phi(\mathbf{x}_1) & \dots & \Phi(\mathbf{x}_m) \\ 1 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{(N+1)\times m}$ with

$$\mathbf{X} = \begin{bmatrix} \Phi(\mathbf{x}_1)^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \Phi(\mathbf{x}_m)^{\mathrm{T}} & 1 \end{bmatrix}, \mathbf{W} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \\ b \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

- Convex and differentiable function

$$\nabla F(\mathbf{W}) = \frac{2}{m}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\right)$$

$$\nabla F(\mathbf{W}) = 0 \Leftrightarrow \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\right) = 0 \Leftrightarrow \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \mathbf{X}\mathbf{Y}$$

Linear Regression: Solution II

- **Solution**:

$$\mathbf{W} = \begin{cases} (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\mathbf{Y}, & \text{if } \ \mathbf{X}\mathbf{X}^{\mathrm{T}} \ \text{ invertible} \\ (\mathbf{X}\mathbf{X}^{\mathrm{T}})^{\dagger}\mathbf{X}\mathbf{Y}, & \text{in general} \end{cases}$$

- Computational complexity: $O(mN + N^3)$ if matrix inversion is in $O(N^3)$
- Poor guarantees in general, no regularization
- For output labels in $\mathbb{R}^p$, $p > 1$, solve p distinct linear regression problems

Ridge Regression

- **Optimization problem**:

$$\min_{\mathbf{w},b} F(\mathbf{w}, b) = \sum_{i=1}^{m} (\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b - y_i)^2 + \lambda \|\mathbf{w}\|^2,$$

where $\lambda \geq 0$ is a regularization parameter

- **Benefits**:
  - — directly based on generalization bound (to be proved soon!)
  - — generalization of linear regression
  - — closed-form solution
  - — can be used with kernels (next slides!)

Ridge Regression: Solution

- Assume $b = 0$: often constant feature is used (but not equivalent to the use of original offset!)

- Rewrite objective function as

$$F(\mathbf{W}) = \|\mathbf{X}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\|^2 + \lambda\|\mathbf{W}\|^2$$

- Convex and differentiable function

$$\nabla F(\mathbf{W}) = 2\lambda\mathbf{W} + 2\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W} - \mathbf{Y})$$
$$\nabla F(\mathbf{W}) = 0 \Leftrightarrow (\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})\mathbf{W} = \mathbf{X}\mathbf{Y}$$

- **Solution**:

$$\mathbf{W} = \underbrace{(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}}_{\text{always invertible!}}\ \mathbf{X}\mathbf{Y}$$

Ridge Regression: Equivalent Formulations

- **Optimization problem I**:

$$\min_{\mathbf{w},b} \sum_{i=1}^{m} \left(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b - y_i\right)^2$$

$$\text{subject to: } \|\mathbf{w}\|^2 \leq \Lambda^2$$

- **Optimization problem II**

$$\min_{\mathbf{w},b} \sum_{i=1}^{m} \xi_i^2$$

$$\text{subject to: } \xi_i = \mathbf{w} \cdot \Phi(\mathbf{x}_i) + b - y_i$$

$$\|\mathbf{w}\|^2 \leq \Lambda^2$$

Ridge Regression Equations

- **Lagrangian**: assume $b = 0$. For all ξ, $\mathbf{w}$, $\boldsymbol{\alpha}'$, $\lambda \geq 0$

$$L(\xi, \mathbf{w}, \boldsymbol{\alpha}', \lambda) = \sum_{i=1}^{m} \xi_i^2 + \sum_{i=1}^{m} \alpha_i'(y_i - \xi_i - \mathbf{w} \cdot \Phi(\mathbf{x}_i)) + \lambda(\|\mathbf{w}\|^2 - \Lambda^2)$$

- **KKT**:

$$\nabla_{\mathbf{w}} L = -\sum_{i=1}^{m} \alpha_i' \Phi(\mathbf{x}_i) + 2\lambda \mathbf{w} = 0 \Leftrightarrow \mathbf{w} = \frac{1}{2\lambda} \sum_{i=1}^{m} \alpha_i' \Phi(\mathbf{x}_i)$$

$$\nabla_{\xi_i} L = 2\xi_i - \alpha_i' = 0 \Leftrightarrow \xi_i = \alpha_i'/2$$

$$\forall i \in [1, m],\ \alpha_i'(y_i - \xi_i - \mathbf{w} \cdot \Phi(\mathbf{x}_i)) = 0$$

$$\lambda(\|\mathbf{w}\|^2 - \Lambda^2) = 0$$

Dual Formulation

- Using expressions of $\mathbf{w}$ and ξ_i we get that

$$L = \sum_{i=1}^{m} \frac{(\alpha_i')^2}{4} + \sum_{\alpha_i' y_i} - \sum_{i=1}^{m} \frac{(\alpha_i')^2}{2} - \frac{1}{2\lambda} \sum_{i,j=1}^{m} \alpha_i' \alpha_j' \Phi(\mathbf{x}_i)^{\mathrm{T}} \Phi(\mathbf{x}_j)$$

$$+ \lambda \left(\frac{1}{4\lambda^2} \left\| \sum_{i=1}^{m} \alpha_i' \Phi(\mathbf{x}_i) \right\|^2 - \Lambda^2 \right)$$

- Thus

$$L = -\frac{1}{4} \sum_{i=1}^{m} (\alpha_i')^2 + \sum_{i=1}^{m} \alpha_i' y_i - \frac{1}{4\lambda} \sum_{i,j=1}^{m} \alpha_i' \alpha_j' \Phi(\mathbf{x}_i)^{\mathrm{T}} \Phi(\mathbf{x}_j) - \lambda \Lambda^2$$

$$= -\lambda \sum_{i=1}^{m} \alpha_i^2 + 2 \sum_{i=1}^{m} \alpha_i y_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j \Phi(\mathbf{x}_i)^{\mathrm{T}} \Phi(\mathbf{x}_j) - \lambda \Lambda^2$$

with $\alpha_i' = 2\lambda \alpha_i$

Dual Optimization Problem

- **Optimization problem**:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m} -\lambda\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y} - \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\alpha}$$

or $$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m} -\boldsymbol{\alpha}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}$$

- **Solution**

$$h(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i)\cdot\Phi(\mathbf{x})$$

with

$$\boldsymbol{\alpha} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{Y}$$

Direct Dual Solution

- **Lemma**: The following matrix identity always holds

$$(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\mathbf{X} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}$$

- **Proof**: Observe that $(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})\mathbf{X} = \mathbf{x}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})$. Left-multiplying by $(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}$ and right-multiplying by $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}$ yields the statement
- **Dual solution**: $\boldsymbol{\alpha}$ such that

$$\mathbf{W} = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}_i, \cdot) = \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i) = \mathbf{X}\boldsymbol{\alpha}$$

By lemma,
$\mathbf{W} = (\mathbf{X}\mathbf{X}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{Y}$. Thus we get that

$$\boldsymbol{\alpha} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{Y}$$

Computational Complexity

Type	Solution	Prediction
Primal	$O(mN^2 + N^3)$	$O(N)$
Dual	$O(\kappa m^2 + m^3)$	$O(\kappa m)$

Here κ denotes the time complexity of computing a kernel value; for polynomial and Gaussian kernels, $\kappa = O(N)$

Repr. Kernel Hilbert Space I (Aronszajn, 1950)

- Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$. We consider the space of functions $\mathbb{H}$ generated by the linear span of $\{K(\cdot, \mathbf{z}), \mathbf{z} \in \mathbb{R}^N\}$; i.e. arbitrary linear combinations of the form

$$h(\mathbf{x}) = \sum_i \alpha_i K(\mathbf{x}, \mathbf{z}_i),$$

  where each kernel term is viewed as a function of the first argument, and indexed by the second
- Suppose K has an eigen-expansion

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} a_i \phi_i(\mathbf{x}) \phi_i(\mathbf{z})$$

  with $a_i > 0$, $\sum_{i=1}^{\infty} a_i^2 < \infty$

Repr. Kernel Hilbert Space II

- Elements of $\mathbb{H}$ have an expansion in terms of these eigen-functions

$$h(\mathbf{x}) = \sum_{i=1}^{\infty} c_i \phi_i(\mathbf{x}),$$

with the constraint that

$$\|h\|_{\mathbb{H}}^2 := \sum_{i=1}^{\infty} \frac{c_i^2}{a_i} < \infty$$

- For $h \in \mathbb{H}$ it can be easily seen that

$$\langle K(\cdot, \mathbf{x}_i), h \rangle = h(\mathbf{x}_i),\ \langle K(\cdot, \mathbf{x}_i), K(\cdot, \mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j)$$

- Thus for $h(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ we get that

$$\|h\|_{\mathbb{H}}^2 = \sum_{i,j=1}^{m} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

General regularization problem statement I

- A general class of regularization problems has the form

$$\min_{h \in \mathbb{H}} \left[\sum_{i=1}^{m} L(y_i, h(\mathbf{x}_i)) + \lambda P(h) \right]$$

  where $L(y, h(\mathbf{x}))$ is a loss function, $P(h)$ is a penalty functional, $\mathbb{H}$ is a space of functions

- In case of RKHS $\mathbb{H}_K$, induced by the kernel K we use $P(h) = \|h\|^2_{\mathbb{H}_K}$ and get

$$\min_{h \in \mathbb{H}_K} \left[\sum_{i=1}^{m} L(y_i, h(\mathbf{x}_i)) + \lambda \|h\|^2_{\mathbb{H}_K} \right]$$

General regularization problem statement II

- Using RKHS basis representation we get equivalent problem formulation

$$\min_{\{c_j\}_{j=1}^{\infty}} \left[\sum_{i=1}^{m} L\left(y_i, \sum_{j=1}^{\infty} c_j \phi_j(\mathbf{x}_i) \right) + \lambda \sum_{j=1}^{\infty} \frac{c_j^2}{a_j} \right]$$

- It the Representer Theorem it is shown that the solution is finite-dimensional, and has the form

$$h(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

Finite-dimensional representation

- Kernel ridge regression

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^m} (\mathbf{Y} - \mathrm{K}\boldsymbol{\alpha})^{\mathrm{T}}(\mathbf{Y} - \mathrm{K}\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{K}\boldsymbol{\alpha}$$

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m} -\lambda\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y} - \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{K}\boldsymbol{\alpha}$$

$$\text{or} \quad \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} -\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{K} + \lambda\mathbf{I})\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}$$

- Solution:

$$h(\mathbf{x}) = \sum_{i=1}^{m} K(\mathbf{x}_i, \mathbf{x})$$

  with $\boldsymbol{\alpha} = (\mathrm{K} + \lambda\mathbf{I})^{-1}\mathbf{Y}$
- Fitted values

$$\hat{\mathbf{Y}} = \mathrm{K}\boldsymbol{\alpha} = (\mathbf{I} + \lambda\mathrm{K}^{-1})^{-1}\mathbf{Y}$$

Comments

- Advantages
  — strong theoretical guarantees
  — generalization to outputs in $\mathbb{R}^p$: single matrix inversion
  — use of kernels
- Disadvantages
  — solution is not sparse
  — training time for large matrices: low-rank approximations of kernel matrix, e.g., Nyström approximation, partial Cholesky decomposition

Support Vector Regression I

- Hypothesis set

$$\{x \to \mathbf{w} \cdot \Phi(\mathbf{x}) + b : \; \mathbf{w} \in \mathbb{R}^N, \, b \in \mathbb{R}\}$$

- Loss function: ϵ-insensitive loss

$$L(y, y') = |y - y'|_\epsilon = \max\left(0, |y' - y| - \epsilon\right)$$

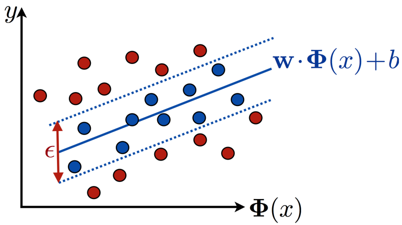


Figure : Fit "tube" with width ϵ to data

Support Vector Regression II

- **Optimization problem**: similar to that of SVM

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m} |y_i - (\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b)|_\epsilon$$

- Equivalent formulation

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

$$\text{subject to} \quad \begin{aligned} &(\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b) - y_i \le \epsilon + \xi_i \\ &y_i - (\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b) \le \epsilon + \xi_i' \\ &\xi_i \ge 0,\ \xi_i' \ge 0 \end{aligned}$$

SVR: Dual formulation

- **Optimization problem**:

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

s.t. $(\mathbf{0}\leq\boldsymbol{\alpha}\leq\mathbf{C})$ or $(\mathbf{0}\leq\boldsymbol{\alpha}'\leq\mathbf{C})$ or $((\boldsymbol{\alpha}'-\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{1}=0)$

- **Solution**

$$h(\mathbf{x})=\sum_{i=1}^{m}(\alpha_i'-\alpha_i)K(\mathbf{x}_i,\mathbf{x})+b$$

with $b =$
$$\begin{cases} -\sum_{i=1}^{m}(\alpha_j'-\alpha_j)K(\mathbf{x}_j,\mathbf{x}_i)+y_i+\epsilon, & \text{when } 0<\alpha_i<C \\ -\sum_{i=1}^{m}(\alpha_j'-\alpha_j)K(\mathbf{x}_j,\mathbf{x}_i)+y_i-\epsilon, & \text{when } 0<\alpha_i'<C \end{cases}$$

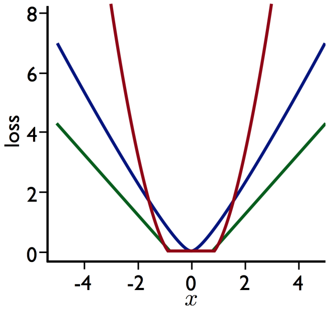
- Support vectors: points strictly outside the tube

Comments

- Advantages
  — strong theoretical guarantees (for that loss)
  — sparser solution
  — use of kernels

- Disadvantages
  — selection of two parameters: C and ϵ. Heuristics for that:
    * search C new maximum y, ϵ new average difference of ys, measure of no. of SVs
  — large matrices: low-rank approximations of kernel matrix

Alternative Loss Functions



- quadratic ϵ-insensitive

$$x \to \max(0, |x| - \epsilon)^2$$

- Huber

$$x \to \begin{cases} x^2, & \text{if } \ |x| \leq c \\ 2c|x| - c^2, & \text{otherwise} \end{cases}$$

- ϵ-insensitive

$$x \to \max(0, |x| - \epsilon)$$

SVR: Quadratic loss

- **Optimization problem**:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'} -\epsilon(\boldsymbol{\alpha}' + \boldsymbol{\alpha})^{\mathrm{T}}\mathbf{1} + (\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\mathrm{T}}\mathbf{Y}$$
$$-\frac{1}{2}(\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\mathrm{T}}\left(\mathrm{K} + \frac{1}{C}\mathbf{I}\right)(\boldsymbol{\alpha}' - \boldsymbol{\alpha})$$
s.t. $(\boldsymbol{\alpha} \geq \mathbf{C})$ or $(\boldsymbol{\alpha}' \geq \mathbf{C})$ or $((\boldsymbol{\alpha}' - \boldsymbol{\alpha})^{\mathrm{T}}\mathbf{1} = 0)$

- **Solution**

$$h(\mathbf{x}) = \sum_{i=1}^{m}(\alpha_i' - \alpha_i)K(\mathbf{x}_i, \mathbf{x}) + b$$

with $b =$
$$\begin{cases} -\sum_{i=1}^{m}(\alpha_j' - \alpha_j)K(\mathbf{x}_j, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i \text{ or } \xi_i = 0 \\ -\sum_{i=1}^{m}(\alpha_j' - \alpha_j)K(\mathbf{x}_j, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha_i' \text{ or } \xi_i' = 0 \end{cases}$$

- Support vectors: points strictly outside the tube
- For $\epsilon = 0$ coincides with KRR

On-line Regression

- On-line version of batch algorithm
  - stochastic gradient descent
  - primal or dual
- Example
  - Mean squared error function: Widrow-Howw (or LMS) algorithm
  - SVR ϵ-insensitive (dual) linear or quadratic function: on-line SVR

Widrow-Hoff

WidrowHoff($\mathbf{w}_0$)

1. $\mathbf{w}_1 \Leftarrow \mathbf{w}_0$ (usually $\mathbf{w}_0 = \mathbf{0}$ is used)
2. for $t \Leftarrow 1$ to T do
3. $\quad$ RECEIVE($\mathbf{x}_t$)
4. $\quad$ $\hat{y}_t \Leftarrow \mathbf{w}_t \cdot \mathbf{x}_t$
5. $\quad$ RECEIVE(y_t)
6. $\quad$ $\mathbf{w}_{t+1} \Leftarrow \mathbf{w}_t + 2\eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)\mathbf{x}_t$ ($\eta > 0$)
7. return $\mathbf{w}_{T+1}$

Dual On-Line SVR

$(b = 0)$ DualSVR

1. $\boldsymbol{\alpha} \Leftarrow \mathbf{0}$
2. $\boldsymbol{\alpha}' \Leftarrow \mathbf{0}$
3. for $t \Leftarrow 1$ to T do
4. $\quad \text{RECEIVE}(\mathbf{x}_t)$
5. $\quad \hat{y}_t \Leftarrow \sum_{s=1}^{T} (\alpha'_s - \alpha_s) K(\mathbf{x}_s, \mathbf{x}_t)$
6. $\quad \text{RECEIVE}(y_t)$
7. $\quad \alpha'_{t+1} \Leftarrow \alpha'_t + \min(\max(\eta(y_t - \hat{y}_t - \epsilon), -\alpha'_t), C - \alpha'_t)$
8. $\quad \alpha_{t+1} \Leftarrow \alpha_t + \min(\max(\eta(\hat{y}_t - y_t - \epsilon), -\alpha_t), C - \alpha_t)$
9. return $\sum_{t=1}^{T} \alpha_t K(\mathbf{x}_t, \cdot)$

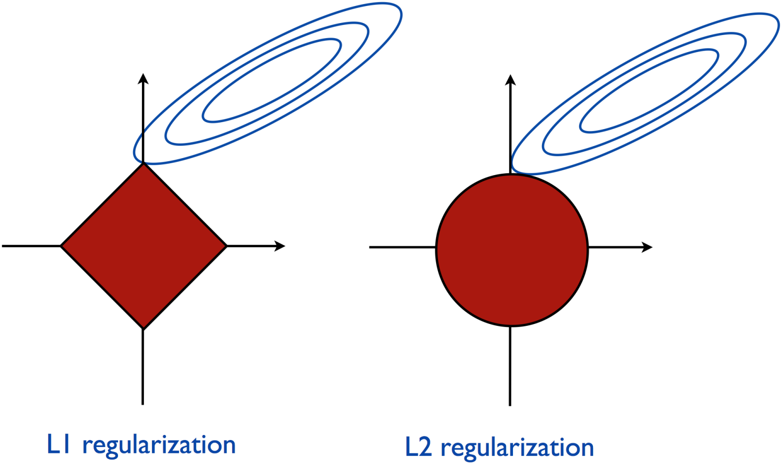
LASSO

- **Optimization problem**: "least absolute shrinkage and selection operator"

$$\min_{\mathbf{w},b} F(\mathbf{w}, b) = \lambda \|\mathbf{w}\|_1 + \sum_{i=1}^{m} (\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b - y_i)^2,$$

  where $\lambda \geq 0$ is a regularization parameter
- **Solution**: equivalent convex quadratic programming (QP)
  — general: standard QP solvers
  — specific algorithms: LARS (least angle regression procedure), entire path of solution

Sparsity of L_1 regularization



Comments

- Advantages
  - strong theoretical guarantees
  - sparse solution
  - feature selection
- Disadvantages
  - no natural use of kernels
  - no closed-form solution (not necessary, but can be convenient for theoretical analysis)
- Many other families of algorithms include
  - neural networks, GPs
  - decision trees
  - boosting trees for regression

Model Selection

- Occam's razor: among competing hypotheses, the one with the fewest assumptions should be selected
- Too much variables/parameters $\Rightarrow$ significant prediction variance and small bias on the training sample, and vice versa
- We have two interrelated problems
  — to estimate value of a target function, characterizing generalization ability of the considered model
  — select an optimal model w.r.t. to the constructed accuracy criterion

Notations

- We consider a linear model $h(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b$, $\mathbf{w} \in \mathbb{R}^N$, $\Phi(\mathbf{x}) \in \mathbb{R}^N$ in a stochastic white noise setting
- Let $J \subseteq \{1, \ldots, N\}$ be a subset of features from $\Phi(\mathbf{x})$ we use to construct a linear model
- We denote by
  — $\mathbf{X}_J$ a submatrix of the full feature matrix $\mathbf{X}$, selected according to the specified subset of feature
  — $\mathbf{w}_J$ linear model coefficients, corresponding to $\mathbf{X}_J$, $\hat{\mathbf{w}}_J$ are their estimates by the least squares method
  - $\hat{h}_J(\mathbf{x}) = \hat{\mathbf{w}}_J \cdot \Phi_J(\mathbf{x}) + \hat{b}$ a regression function, $\hat{y}_i(J) = \hat{h}_J(\mathbf{x}_i)$

Regression Risk I

- Risk of a prediction

$$R(J) = \sum_{i=1}^{m} \mathbb{E}(\hat{y}_i(J) - y_i^*)^2,$$

where y_i^* is a newly randomly generated y_i (with independently generated noise value) for the same $\mathbf{x}_i$

- The problem is to select J, such that $R(J)$ is small
- Risk estimate on the training set is equal to

$$\hat{R}_{\mathrm{tr}}(J) = \sum_{i=1}^{m} (\hat{y}_i(J) - y_i)^2$$

- **Theorem**: $\mathbb{E}(\hat{R}_{\mathrm{tr}}(J)) < R(J)$ and

$$\mathrm{bias}(\hat{R}_{\mathrm{tr}}(J)) = \mathbb{E}(\hat{R}_{\mathrm{tr}}(J)) - R(J) = -2\sum_{i=1}^{m} \mathrm{Cov}(\hat{y}_i, y_i)$$

Regression Risk II

- It can be proved, that in the linear case

$$2\sum_{i=1}^{m}\mathrm{Cov}(\hat{y}_i, y_i) \sim 2|J|\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is an estimate of an output noise standard deviation σ^2, obtained using residuals on the training set, calculated by fitting the model

- Thus, we get C_p Mallow statistics, representing asymptotically unbiased estimate of the regression risk

$$\hat{R}(J) = \hat{R}_{\mathrm{tr}}(J) + 2|J|\hat{\sigma}^2.$$

The second term here penalizes complexity

Regression Risk II

- AIC (Akaike Information Criterion) provides estimate of the risk in case of more general models. It has the form

$$\mathcal{L}_J - |J|,$$

  where
  - $\mathcal{L}_J$ is a model log-likelihood
  - $|J|$ is a number of model parameters
- AIC is equivalent to Mallow C_p in case of linear regression model with a Gaussian noise

Regression Risk III

- Another possibility to estimate riks: leave-one-out cross-validation

$$\hat{R}_{CV}(J) = \sum_{i=1}^{m} \left(y_i - \hat{y}_{(-i)}\right)^2,$$

where $\hat{y}_{(-i)}$ is a prediction, obtained by a model, constructed using a sample $S \setminus \{(\mathbf{x}_i, y_i)\}$

- Increase computational efficiency using formula

$$\hat{R}_{CV}(J) = \sum_{i=1}^{m} \left(\frac{y_i - \hat{y}_i(J)}{1 - U_{ii}(J)}\right)^2$$

$$U(J) = \mathbf{X}_J(\mathbf{X}_J^{\mathrm{T}}\mathbf{X}_J)^{-1}\mathbf{X}_J^{\mathrm{T}}$$