5 - Regression problems

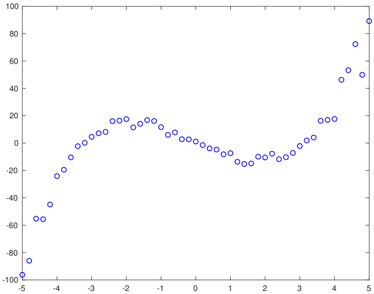
Mauro Passacantando

Department of Computer Science, University of Pisa
mauro.passacantando@unipi.it

Optimization Methods and Game Theory
Master of Science in Artificial Intelligence and Data Engineering
University of Pisa – A.Y. 2020/21

Polynomial regression

We have ℓ experimental data $y_1, y_2, \dots, y_\ell \in \mathbb{R}$ corresponding to observations made on points $x_1, x_2, \dots, x_\ell \in \mathbb{R}$.



We want to find the best approximation of experimental data with a polynomial p of degree $n-1$, with $n \leq \ell$.
Polynomial p has coefficients $z_1, \dots, z_n$:

$$p(x) = z_1 + z_2\, x + z_3\, x^2 + \cdots + z_n\, x^{n-1}$$

Polynomial regression - model

The residual is the vector $r \in \mathbb{R}^\ell$ such that $r_i = p(x_i) - y_i$, with $i = 1, \dots, n$.

We want to find coefficients z of polynomial p such that $\|r\|$ is minimum, i.e., solving the following unconstrained problem

$$\begin{cases} \min \ \|A\,z - y\| \\ z \in \mathbb{R}^n \end{cases}$$

where

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_\ell & x_\ell^2 & \dots & x_\ell^{n-1} \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{pmatrix}$$

For any norm, the objective function $f(z) = \|Az - y\|$ is convex.

We will consider three special norms: $\|\cdot\|_2$, $\|\cdot\|_1$ and $\|\cdot\|_\infty$.

Polynomial regression with $\| \cdot \|_2$

Euclidean norm $\| \cdot \|_2$ (least squares approximation)
$\rightarrow$ unconstrained quadratic programming problem:

$$\begin{cases} \min \ \frac{1}{2}\|A\,z - y\|_2^2 = \frac{1}{2}(Az - y)^{\mathsf{T}}(Az - y) = \frac{1}{2}z^{\mathsf{T}}A^{\mathsf{T}}Az - z^{\mathsf{T}}A^{\mathsf{T}}y + \frac{1}{2}y^{\mathsf{T}}y \\ z \in \mathbb{R}^n \end{cases}$$

It can be proved that $\text{rank}(A) = n$, thus $A^{\mathsf{T}}A$ is positive definite.

Hence, the unique optimal solution is the stationary point of the objective function, i.e., the solution of the system of linear equations:

$$A^{\mathsf{T}}A\,z = A^{\mathsf{T}}y$$

Polynomial regression with $\|\cdot\|_1$

Norm $\|\cdot\|_1 \to$ linear programming problem:

$$\begin{cases} \min \ \|Az-y\|_1 = \sum_{i=1}^{\ell} |A_i z - y_i| \\ z \in \mathbb{R}^n \end{cases}$$

is equivalent to

$$\begin{cases} \min\limits_{z,u} \ \sum_{i=1}^{\ell} u_i \\ u_i = |A_i z - y_i| \\ \quad = \max\{A_i z - y_i, y_i - A_i z\} \end{cases} \quad \to \quad \begin{cases} \min\limits_{z,u} \ \sum_{i=1}^{\ell} u_i \\ u_i \geq \max\{A_i z - y_i, y_i - A_i z\} \end{cases}$$

$$\to \quad \begin{cases} \min\limits_{z,u} \ \sum_{i=1}^{\ell} u_i \\ u_i \geq A_i z - y_i & \forall\ i = 1, \dots, \ell \\ u_i \geq y_i - A_i z & \forall\ i = 1, \dots, \ell \end{cases}$$

Polynomial regression with $\|\cdot\|_\infty$

Norm $\|\cdot\|_\infty \to$ linear programming problem:

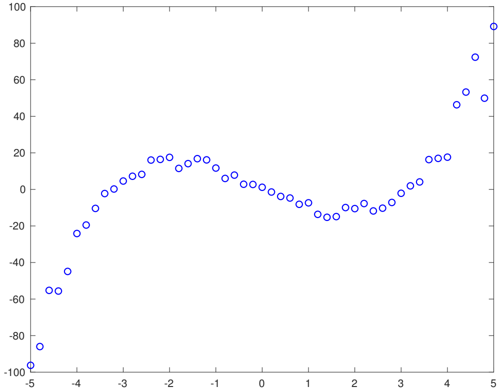
$$\begin{cases} \min \ \|Az - y\|_\infty = \max\limits_{i=1,\dots,\ell} |A_i z - y_i| \\ z \in \mathbb{R}^n \end{cases}$$

is equivalent to

$$\begin{cases} \min \ u \\ u = \max\limits_{i=1,\dots,\ell} |A_i z - y_i| \end{cases} \to \quad \begin{cases} \min\limits_{z,u} \ u \\ u \geq A_i z - y_i \qquad \forall \ i = 1,\dots,\ell \\ u \geq y_i - A_i z \qquad \forall \ i = 1,\dots,\ell \end{cases}$$

Polynomial regression

Exercise 5.1. Consider the experimental data in the file `5-1.txt`.

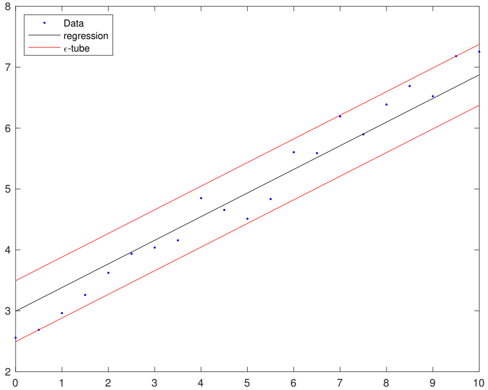


a) Find the best approximating polynomial of degree 3 with respect to $\| \cdot \|_2$.
b) Find the best approximating polynomial of degree 3 w.r.t. $\| \cdot \|_1$.
c) Find the best approximating polynomial of degree 3 w.r.t. $\| \cdot \|_\infty$.

ε-SV regression

We have a set of training data $\{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$, where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. In ε-SV regression we aim to find a function f that

- has at most ε deviation from the targets y_i for all the training data
- is as flat as possible



Linear SVM

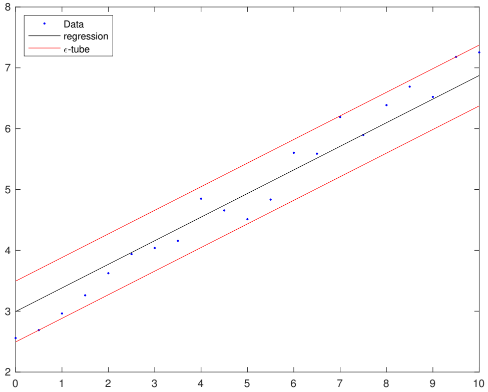
Start with linear regression.
Consider an affine function $f(x) = w^{\mathsf{T}}x + b$ and set a tolerance parameter ε.

Flatness means that one seek a small w, that is we aim to solve the quadratic optimization problem

$$\begin{cases} \min\limits_{w,b} \ \dfrac{1}{2}\|w\|^2 \\ y_i \leq w^{\mathsf{T}}x_i + b + \varepsilon \qquad \forall\ i = 1, \ldots, \ell \\ y_i \geq w^{\mathsf{T}}x_i + b - \varepsilon \qquad \forall\ i = 1, \ldots, \ell \end{cases} \tag{1}$$

Linear SVM

Exercise 5.2. Apply the linear SVM model with $\varepsilon = 0.5$ to the training data contained in the file `5-2.txt`.



Linear SVM with slack variables

If ε is too small, the model (1) could not be feasible.

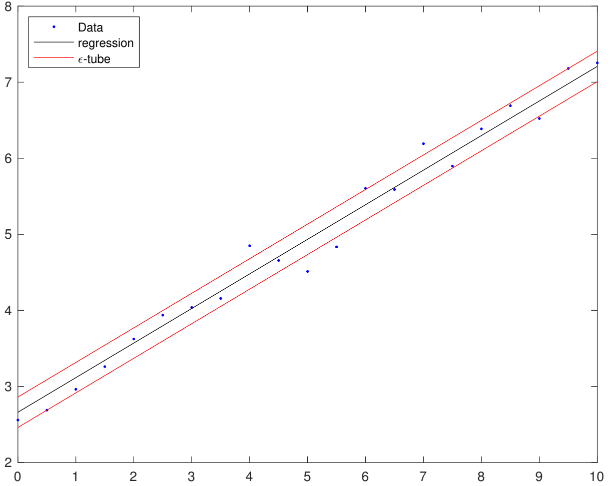
The linear SVM model can be extended by introducing slack variables ξ^+ and ξ^- to relax the constraints of problem (1):

$$\begin{cases} \min\limits_{w,b,\xi^+,\xi^-} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i^+ + \xi_i^-) \\ y_i \leq w^{\mathsf{T}} x_i + b + \varepsilon + \xi_i^+ \qquad \forall\ i = 1,\dots,\ell \\ y_i \geq w^{\mathsf{T}} x_i + b - \varepsilon - \xi_i^- \qquad \forall\ i = 1,\dots,\ell \\ \xi^+ \geq 0 \\ \xi^- \geq 0 \end{cases} \tag{2}$$

where parameter C gives the trade-off between the flatness of f and tolerance to deviations larger than ε.

Linear SVM with slack variables

Exercise 5.3. Apply the linear SVM with slack variables (set $\varepsilon = 0.2$ and $C = 10$) to the training data contained in the file `5-2.txt`.



Linear SVM with slack variables - dual problem

What is the dual of problem (2)? The Lagrangian function is

$$L(\underbrace{w, b, \xi^+, \xi^-}_{\text{primal var.}}, \underbrace{\lambda^+, \lambda^-, \eta^+, \eta^-}_{\text{dual var.}}) = \frac{1}{2}\|w\|^2 - w^\mathsf{T}\left[\sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-)x_i\right] - b\sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-)$$
$$+ \sum_{i=1}^{\ell}\xi_i^+(C - \lambda_i^+ - \eta_i^+) + \sum_{i=1}^{\ell}\xi_i^-(C - \lambda_i^- - \eta_i^-)$$

If $\sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-) \neq 0$ or $C - \lambda_i^+ - \eta_i^+ \neq 0$ for some i or $C - \lambda_i^- - \eta_i^- \neq 0$ for some i, then $\min_{w,b,\xi^+,\xi^-} L = -\infty$. Otherwise,

$$\nabla_w L = w - \sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-)x_i = 0.$$

Linear SVM with slack variables - dual problem

The dual problem is

$$\begin{cases} \max\limits_{\lambda^+,\lambda^-} \; -\dfrac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}(\lambda_i^+ - \lambda_i^-)(\lambda_j^+ - \lambda_j^-)(x_i)^\mathsf{T} x_j \\ \qquad\qquad -\varepsilon \sum_{i=1}^{\ell}(\lambda_i^+ + \lambda_i^-) + \sum_{i=1}^{\ell} y_i(\lambda_i^+ - \lambda_i^-) \\ \sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-) = 0 \\ \lambda_i^+ \in [0, C] \\ \lambda_i^- \in [0, C] \end{cases}$$

Linear SVM with slack variables - dual problem

- Dual problem is a convex quadratic programming problem
- Dual constraints are simpler than primal constraints
- If $\lambda_i^+ > 0$ or $\lambda_i^- > 0$, then x_i is said support vector
- If (λ^+, λ^-) is a dual optimum, then

$$w = \sum_{i=1}^{\ell} (\lambda_i^+ - \lambda_i^-) x_i,$$

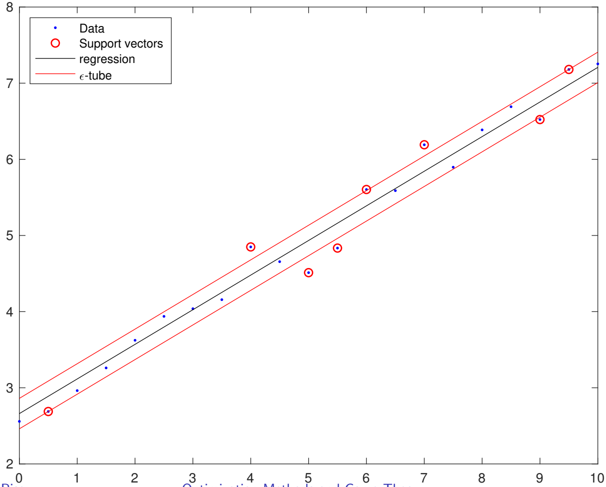
- b is obtained using the complementarity conditions:

$$\begin{array}{l} \lambda_i^+ \left[\varepsilon + \xi_i^+ - y_i + w^\mathsf{T} x_i + b\right] = 0 \\ \lambda_i^- \left[\varepsilon + \xi_i^- + y_i - w^\mathsf{T} x_i - b\right] = 0 \\ \xi_i^+ (C - \lambda_i^+) = 0 \\ \xi_i^- (C - \lambda_i^-) = 0 \end{array}$$

Hence, if there is some i s.t. $0 < \lambda_i^+ < C$, then $b = y_i - w^\mathsf{T} x_i - \varepsilon$;
if there is some i s.t. $0 < \lambda_i^- < C$, then $b = y_i - w^\mathsf{T} x_i + \varepsilon$.

Linear SVM with slack variables - dual problem

Exercise 5.4. Solve the dual problem of the linear SVM with slack variables (set $\varepsilon = 0.2$ and $C = 10$) applied to the training data contained in the file `5-2.txt`. Moreover, find the support vectors.



Nonlinear SVM

How to generate a nonlinear regression function f?

Nonlinear SVM

How to generate a nonlinear regression function f? Use the kernel!

Define a map $\phi : \mathbb{R}^n \to \mathcal{H}$, where $\mathcal{H}$ (features space) is an higher dimensional (maybe infinite) space and find the linear regression for the points $\{(\phi(x_i), y_i)\}$ in the feature space $\mathcal{H}$.

Nonlinear SVM

Primal problem:

$$\begin{cases} \min \ \dfrac{1}{2}\|w\|^2 + C \displaystyle\sum_{i=1}^{\ell}(\xi_i^+ + \xi_i^-) \\ y_i \leq w^\mathsf{T}\phi(x_i) + b + \varepsilon + \xi_i^+ & \forall\ i = 1,\dots,\ell \\ y_i \geq w^\mathsf{T}\phi(x_i) + b - \varepsilon - \xi_i^- & \forall\ i = 1,\dots,\ell \end{cases}$$

w is a vector in a high dimensional space (maybe infinite variables)

Dual problem:

$$\begin{cases} \displaystyle\max_{(\lambda^+,\lambda^-)} \ -\frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}(\lambda_i^+ - \lambda_i^-)(\lambda_j^+ - \lambda_j^-)\phi(x_i)^\mathsf{T}\phi(x_j) \\ \qquad\qquad -\varepsilon\displaystyle\sum_{i=1}^{\ell}(\lambda_i^+ + \lambda_i^-) + \sum_{i=1}^{\ell} y_i(\lambda_i^+ - \lambda_i^-) \\ \displaystyle\sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-) = 0 \\ \lambda_i^+, \lambda_i^- \in [0, C] \end{cases}$$

number of variables $= 2\ell$

Nonlinear SVM

Primal problem:

$$\begin{cases} \min \ \dfrac{1}{2}\|w\|^2 + C\displaystyle\sum_{i=1}^{\ell}(\xi_i^+ + \xi_i^-) \\ y_i \le w^{\mathsf{T}}\phi(x_i) + b + \varepsilon + \xi_i^+ \qquad \forall\ i = 1,\dots,\ell \\ y_i \ge w^{\mathsf{T}}\phi(x_i) + b - \varepsilon - \xi_i^- \qquad \forall\ i = 1,\dots,\ell \end{cases}$$

w is a vector in a high dimensional space (maybe infinite variables)

Dual problem:

$$\begin{cases} \displaystyle\max_{(\lambda^+,\lambda^-)} \ -\frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}(\lambda_i^+ - \lambda_i^-)(\lambda_j^+ - \lambda_j^-)k(x_i, x_j) \\ \qquad\qquad \displaystyle -\varepsilon\sum_{i=1}^{\ell}(\lambda_i^+ + \lambda_i^-) + \sum_{i=1}^{\ell} y_i(\lambda_i^+ - \lambda_i^-) \\ \displaystyle\sum_{i=1}^{\ell}(\lambda_i^+ - \lambda_i^-) = 0 \\ \lambda_i^+, \lambda_i^- \in [0, C] \end{cases}$$

number of variables $= 2\ell$

Nonlinear SVM

Therefore:

- choose a kernel k
- solve the dual $\to$ find (λ^+, λ^-)
- find b:

$$b = y_i - \varepsilon - \sum_{j=1}^{\ell} (\lambda_j^+ - \lambda_j^-) k(x_i, x_j), \qquad \text{for some } i \text{ s.t. } 0 < \lambda_i^+ < C$$

  or

$$b = y_i + \varepsilon - \sum_{j=1}^{\ell} (\lambda_j^+ - \lambda_j^-) k(x_i, x_j), \qquad \text{for some } i \text{ s.t. } 0 < \lambda_i^- < C$$

- Recession function

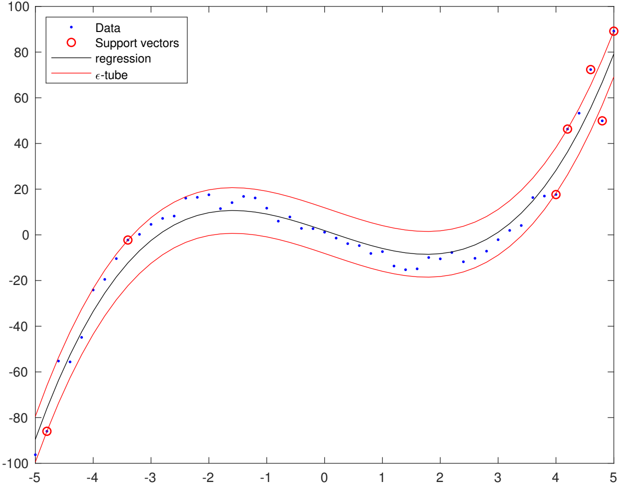
$$f(x) = \sum_{i=1}^{\ell} (\lambda_i^+ - \lambda_i^-) k(x_i, x) + b$$

Recession function is

- linear in the features space
- nonlinear in the input space

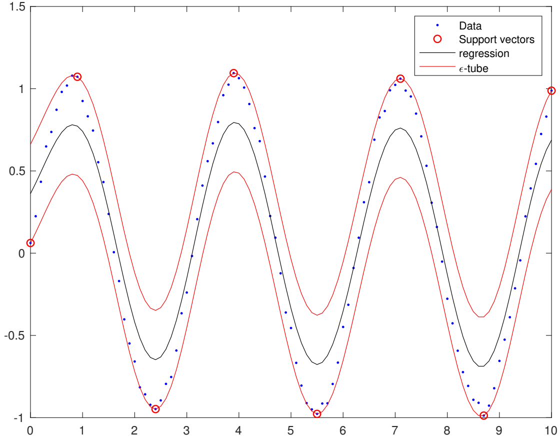
Nonlinear SVM

Exercise 5.5. Consider the training data in the file 5-1.txt. Find the nonlinear regression function given by the nonlinear SVM using a polynomial kernel with degree $p = 3$ and parameters $\varepsilon = 10$, $C = 10$. Moreover, find the support vectors.



Nonlinear SVM

Exercise 5.6. Consider the training data in the file 5-6.txt. Find the nonlinear regression function given by the nonlinear SVM using a Gaussian kernel with $\gamma = 1$ and parameters $\varepsilon = 0.3$, $C = 10$. Moreover, find the support vectors.



If you set $\varepsilon = 0.1$, how do support vectors change?