Machine Learning

Support Vector Machines

Fabio Vandin

November 27^{th}, 2023

Some Kernels

The following are the most commonly used kernels

- linear kernel: $\psi(\mathbf{x}) = \mathbf{x}$ (standard SVM)

Some Kernels

The following are the most commonly used kernels

hyperbolic tangent

- linear kernel: $\psi(\mathbf{x}) = \mathbf{x}$
- sigmoid: $K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma\langle \mathbf{x}, \mathbf{x}' \rangle + \zeta)$ (for γ, $\zeta > 0$)

Some Kernels

The following are the most commonly used kernels

- linear kernel: $\psi(\mathbf{x}) = \mathbf{x}$
- sigmoid: $K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma\langle\mathbf{x}, \mathbf{x}'\rangle + \zeta)$ (for γ, $\zeta > 0$)
- degree-Q polynomial kernel
- Gaussian-radial basis function (RBF) kernel

Degree-Q polynomial kernel

Definition

For given constants $\gamma > 0, \zeta > 0$ and for $Q \in \mathbb{N}$, the *degree-Q polynomial kernel* is

$$K(\mathbf{x}, \mathbf{x}') = (\zeta + \gamma \langle \mathbf{x}, \mathbf{x}' \rangle)^Q$$

time? $\cancel{\Theta(d)}$
$\cancel{\Theta(d+Q)}$
$\Theta(d + \log Q)$

Example

For $Q = 2$:

$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$

$$\psi(\mathbf{x}) = [\zeta, \sqrt{2\zeta\gamma}x_1, \sqrt{2\zeta\gamma}x_2, \ldots, \sqrt{2\zeta\gamma}x_d, \gamma x_1 x_1, \gamma x_1 x_2, \ldots, \gamma x_d x_d]^T \in \mathbb{R}^{1+d+d^2}$$

In general: $\psi(\vec{x}) \in \mathbb{R}^{\Theta(d^Q)}$

Gaussian-RBF Kernel

Definition

For a given constant $\gamma > 0$ the *Gaussian-RBF kernel* is

$$K(\mathbf{x}, \mathbf{x}') = e^{-\gamma ||\mathbf{x}-\mathbf{x}'||^2}$$

What is $\psi(\mathbf{x})$? Assume $\gamma = 1$ and $\mathbf{x} = x \in \mathbb{R}$ for simplicity, then

$$\begin{aligned}
K(x, x') &= e^{-||x-x'||^2} \\
&= e^{-x^2} e^{2xx'} e^{-(x')^2} \\
&= e^{-x^2} \underbrace{\left(\sum_{k=0}^{+\infty} \frac{2^k (x)^k (x')^k}{k!} \right)}_{\text{Taylor's expansion of}} e^{-(x')^2}
\end{aligned}$$

$$e^{2xx'} = \sum_{k=0}^{+\infty} \frac{(2xx')^k}{k!}$$

Gaussian-RBF Kernel

Definition

For a given constant $\gamma > 0$ the *Gaussian-RBF kernel* is

$$K(\mathbf{x}, \mathbf{x}') = e^{-\gamma ||\mathbf{x} - \mathbf{x}'||^2}$$

What is $\psi(\mathbf{x})$? Assume $\gamma = 1$ and $\mathbf{x} = x \in \mathbb{R}$ for simplicity, then

$$\begin{aligned} K(x, x') &= e^{-||x - x'||^2} \\ &= e^{-x^2} e^{2xx'} e^{-(x')^2} \\ &= e^{-x^2} \left(\sum_{k=0}^{+\infty} \frac{2^k (x)^k (x')^k}{k!} \right) e^{-(x')^2} \end{aligned}$$

$$\Rightarrow \psi(x) = e^{-x^2} \left(1, \sqrt{\tfrac{2}{1!}} x, \sqrt{\tfrac{2^2}{2!}} x^2, \sqrt{\tfrac{2^3}{3!}} x^3, \ldots \right)^T$$

Gaussian-RBF Kernel

Definition

For a given constant $\gamma > 0$ the *Gaussian-RBF kernel* is

$$K(\mathbf{x}, \mathbf{x}') = e^{-\gamma||\mathbf{x}-\mathbf{x}'||^2}$$

What is $\psi(\mathbf{x})$? Assume $\gamma = 1$ and $\mathbf{x} = x \in \mathbb{R}$ for simplicity, then

$$\begin{aligned} K(x, x') &= e^{-||x-x'||^2} \\ &= e^{-x^2} e^{2xx'} e^{-(x')^2} \\ &= e^{-x^2} \left(\sum_{k=0}^{+\infty} \frac{2^k (x)^k (x')^k}{k!} \right) e^{-(x')^2} \end{aligned}$$

$$\Rightarrow \psi(x) = e^{-x^2} \left(1, \sqrt{\tfrac{2}{1!}}x, \sqrt{\tfrac{2^2}{2!}}x^2, \sqrt{\tfrac{2^3}{3!}}x^3, \dots \right)^T$$

$\Rightarrow \psi(x)$ has infinite number of dimensions!

Choice of Kernel

Notes

- polynomial kernel: usually used with $Q \leq 10$
- Gaussian-RBF kernel: usually $\gamma \in [0, 1]$
- many other choices are possible!

Mercer's condition

$K(\mathbf{x}, \mathbf{x}')$ is a valid kernel function if and only if the kernel matrix

$$K = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) \dots & K(\mathbf{x}_1, \mathbf{x}_m) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) \dots & K(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & \vdots \\ K(\mathbf{x}_m, \mathbf{x}_1) & K(\mathbf{x}_m, \mathbf{x}_2) \dots & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

is always symmetric positive semi-definite for any given $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$.

Support Vector Machines for Regression

$Y = \mathbb{R}$

SVMs can be also used for regression. The function to be minimized will be

loss function

observed value

$$\ell_2\text{-regularization} \leftarrow \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_{i=1}^{m} V_\varepsilon(y_i \underbrace{- \langle \mathbf{x}_i, \mathbf{w} \rangle - b}_{-\text{ prediction for } \vec{x}_i \text{ (by model } \vec{w}, b)})$$

where

$$V_\varepsilon(r) = \begin{cases} 0 & \text{if } |r| < \varepsilon \\ |r| - \varepsilon & \text{otherwise} \end{cases}$$

One can prove that the solution has the form:

$$\mathbf{w} = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\mathbf{x}_i$$

in the training set

and that the final model produced in output is

if we use a transformation $\psi \Rightarrow \langle \psi(\vec{x}), \psi(\vec{x}') \rangle = K_\psi(\vec{x}, \vec{x}')$

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

where $\alpha_i^*, \alpha_i \geq 0$ and are the solution to a suitable QP.

Definition

Support vector: $\mathbf{x}_i$ such that $\alpha_i^* - \alpha_i \neq 0$

One can define kernels, similarly to SVM for classification.

Exercise 4

Assuming we have the following dataset ($x_i \in \mathbb{R}^2$) and by solving the SVM for classification we get the corresponding optimal dual variables:

i	x_i^T	y_i	α_i^*
1	[0.2 -1.4]	-1	0
2	[-2.1 1.7]	1	0
3	[0.9 1]	1	0.5
4	[-1 -3.1]	-1	0
5	[-0.2 -1]	-1	0.25
6	[-0.2 1.3]	1	0
7	[2.0 -1]	-1	0.25
8	[0.5 2.1]	1	0

Answer to the following:

(A) What are the support vectors?

(B) Draw a schematic picture reporting the data points (approximately) and the optimal separating hyperplane, and mark the support vectors. Would it be possible, by moving only two data points, to obtain the SAME separating hyperplane with only 2 support vectors? If so, draw the modified configuration (approximately).

Bibliography [UML]

SVM: Chapter 15

- no sections 15.1.2, 15.2.1, 15.2.2,15.2.3,

Kernels: Chapter 16

- no section 16.3