* This bound is not depend on VCD, d+1 (Feature Dimension)

Theorem 4.7:

Let μ be a distribution on $\mathbb{R}^d \times \{-1, +1\}$ such that for $(\mathbf{X}, Y) \sim \mu$ we have almost surely $\|\mathbf{X}\| \le \rho < \infty$. Then for

$$\mathbf{w}_s := \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\, \lambda\|\mathbf{w}\|^2 + \mathcal{R}_s^{\text{hinge}}(f_{\mathbf{w},0}),$$

we have

corresponds to ε_{app}

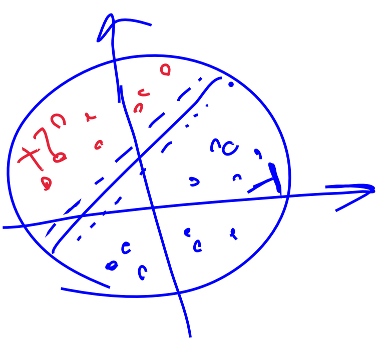
tends to 0

$$\mathbb{E}_{\mu^m}\left[\mathcal{R}_\mu^{\text{hinge}}(f_{\mathbf{w}_S,0})\right] \le \min_{\mathbf{v}\in\mathbb{R}^d} \left(\underline{\mathcal{R}_\mu^{\text{hinge}}(f_{\mathbf{v},0})} + \lambda\|\mathbf{v}\|^2\right) + \frac{2\rho^2}{\lambda m}.$$

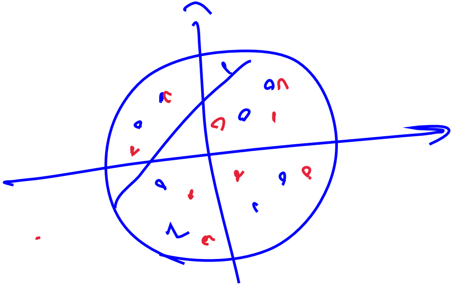
→ similar & corresponds ε_{est}

* True expected risk

- The term $\frac{2\rho^2}{\lambda m}$ bounds the (mean) estimation error $\mathbb{E}_{\mu^m}[\varepsilon_{\text{est}}(S)]$ and the green highlighted text the approximation error.
- Again, the bound for the generalization error does not depend on the feature dimension $d = \mathrm{VCD}(\mathcal{L}_d^0)$. This has some advantages in practice, e.g., in text classification where $d \gg 10^4$ but $\|\mathbf{x}\| \le 1 = \rho$.



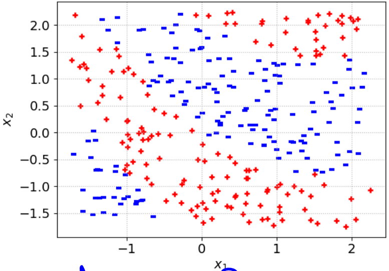
Hard svm



Soft svm

4.3 Kernel SVM

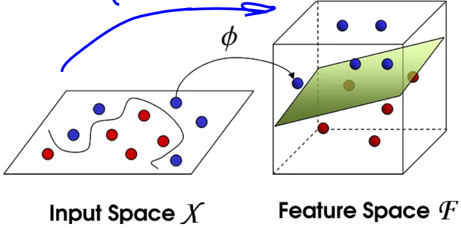
- The expressive power of linear hypotheses on $\mathcal{X}$ is of course quite restricted.
- For example, classification patterns like the one on the right cannot be described by linear hypotheses in $\mathbf{x}$.
- We introduce an approach to solve more complicated classification tasks with still using linear methods.



idea of KSVM

- **The idea:** Map the original inputs $\mathbf{x} \in \mathcal{X}$ into a (much) larger feature space $\mathcal{F}$, and apply linear hypotheses $h_{\mathcal{F}}\colon \mathcal{F} \to \{-1, +1\}$ in $\mathcal{F}$ for classification.
- The corresponding embedding $\psi\colon \mathcal{X} \to \mathcal{F}$ is called feature map and the resulting nonlinear hypotheses on $\mathcal{X}$ are then $h_{\mathcal{F}} \circ \psi$.

کمشی کن سای ر (لای)

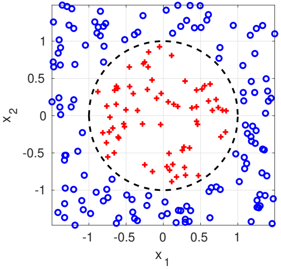


Source: towardsdatascience.com

Example

- We consider classification of in $\mathcal{X} = \mathbb{R}^2$ with true hypothesis

$$h^\dagger(\mathbf{x}) = \begin{cases} +1, & \|\mathbf{x}\| \leq 1 \\ -1, & \text{else.} \end{cases}$$



- A corresponding sample s with random $\mathbf{x}_i \in \mathcal{X}$ is shown on the right.

- This is obviously not reasonably explainable by linear hypotheses.

- But if we use the feature map

Feature space

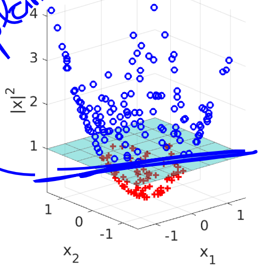
main Feature

n+1

seprated via a hyperplane

$$\psi(\mathbf{x}) := (\mathbf{x}, \|\mathbf{x}\|^2) \in \mathcal{F} = \mathbb{R}^3$$

so the embedded sample $\psi(s)$ is easily linearly separable and $h^\dagger = h_{\mathcal{F}} \circ \psi$.



Futher example

- One can easily extend the previous example naturally by considering the following feature mapping:

$$\psi(x) := (1, x, x^2, \dots, x^n) \in \mathbb{R}^{n+1} =: \mathcal{F}$$

$$h(x) = \mathrm{sgn}(w\psi(x) + b)$$

- With this, one can now easily solve any polynomial hypothesis

$$h_p(x) = \mathrm{sgn}\,(p(x))\,, \qquad p(x) = \sum_{i=0}^{n} w_i x^i,$$

$$h: \mathcal{X} \to \pm 1$$

simply represent as a linear hypothesis in the feature space $\mathcal{F}$:

$$h_p(x) = \mathrm{sgn}\,(\mathbf{w} \cdot \psi(x)) = h_{\mathbf{w}}(\psi(x)),$$

where $\mathbf{w} = (w_0, w_1, \dots, w_n) \in \mathbb{R}^{n+1}$ and $h_{\mathbf{w}} \in \mathcal{L}_{n+1}$

- This can be extended to multivariate polynomials in $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$.

Learning in $\mathcal{F}$

For high- or even infinite-dimensional inner product feature spaces $\mathcal{F}$ learning a linear hypothesis $h \in \mathcal{L}_{\mathcal{F}}$,

$$\mathcal{L}_{\mathcal{F}} = \{h \colon \mathcal{F} \to \{\pm 1\} \mid h(\psi) = \operatorname{sgn}(\mathbf{w} \cdot \psi + b),\ \mathbf{w} \in \mathcal{F},\ b \in \mathbb{R}\},$$

using the embedded sample $\psi(s) = \{(\psi(\mathbf{x}_1), y_1), \ldots, (\psi(\mathbf{x}_m), y_m)\}$ is demaing for two reasons:

1. The VC dimension $\mathrm{VCD}(\mathcal{L}_{\mathcal{F}})$ grows linearly with the dimension of $\mathcal{F}$ – for a small estimation error we need (infinitely) many training data. So VCD of this is ∞ (very bad)
2. The computational cost to determine a weight vector $\mathbf{w} \in \mathcal{F}$ also increases with the dimension of $\mathcal{F}$ – simply because of the representation or discretization of $\mathbf{w}$.

For the first problem we have already learned a solution: giving

1. The hard and soft SVM rule have, given certain conditions on μ, a dimension-independent sample complexity and the estimation error for $\mathrm{SVM}_{\text{hard}}$ and $\mathrm{SVM}_{\text{soft}}$ can be independent of $\mathrm{VCD}(\mathcal{L}_{\mathcal{F}})$.

SVM-rules in $\mathcal{F}$

From now on we assume that $\mathcal{F}$ is an inner product space with inner product also denoted by "$\cdot$"

Hard SVM-rule in $\mathcal{F}$

Compute: $h_s(\mathbf{x}) := \operatorname{sgn}\left(\mathbf{w}_s \cdot \psi(\mathbf{x}) + b_s\right)$ by

$$(\mathbf{w}_s, b_s) = \underset{\mathbf{w}\in\mathcal{F}, b\in\mathbb{R}}{\operatorname{argmin}} \|\mathbf{w}\|^2 \qquad \text{subject to:} \qquad y_i\left(\mathbf{w}\cdot\psi(\mathbf{x}_i) + b\right) \geq 1 \qquad \forall i.$$

Soft SVM-rule in $\mathcal{F}$

Compute: $h_s(\mathbf{x}) := \operatorname{sgn}\left(\mathbf{w}_s \cdot \psi(\mathbf{x}) + b_s\right)$ by

$$(\mathbf{w}_s, b_s) \in \underset{\mathbf{w}\in\mathcal{F}, b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i[\mathbf{w}\cdot\psi(\mathbf{x}_i) + b]\} + \lambda\|\mathbf{w}\|^2$$

very high weights so what to do?

z.B chatGPT uses 1M.

For svm it is called support of maxing the min of margin

For actually computing $\mathbf{w}_s \in \mathcal{F}$ and resolving the second challenge we need a new result:

Theorem 4.8: Representer theorem

✱✱✱ will come in exam

The outcome $\mathbf{w}_s \in \mathcal{F}$ of the hard and soft SVM rule in $\mathcal{F}$ as well as any

$$(\mathbf{w}_s, b_s) \in \underset{\mathbf{w}\in\mathcal{F}, b\in\mathbb{R}}{\operatorname{argmin}} f\left(\mathbf{w}\cdot\psi(\mathbf{x}_1)+b,\ldots,\mathbf{w}\cdot\psi(\mathbf{x}_m)+b\right)+R(\|\mathbf{w}\|)$$

for arbitrary $f\colon \mathbb{R}^m \to \mathbb{R}\cup\underline{\{+\infty\}}$ and strictly increasing $R\colon [0,\infty)\to\mathbb{R}$ can be represented as

for infinite w_s

$$\underline{\mathbf{w}_s = \sum_{i=1}^m \alpha_{s,i}\,\boxed{\psi(\mathbf{x}_i)},} \qquad \alpha_{s,i}\in\mathbb{R}.$$

$$w_s = \sum_{i=1}^m \alpha_s \cdot \boxed{w(x)}$$ finite works

Consequence: Instead of learning a high and maybe infinite-dimensional $\mathbf{w}_s$ we can "simply" learn the finitely many coefficients α_i

$$\boldsymbol{\alpha}_s = (\alpha_{s,1},\ldots,\alpha_{s,m}).$$

which can be computed over Kernel !. (what is it)

Proof Th 4,8:

let (w_s, b_s) Solves the minimization problem, Since feature space(f) is an inner product space, we can consider the orthogonal projection of w_s onto span of $\{\psi(x_1), \dots \varphi(x_m)\}$:

$$w_s = W + \underline{\underline{v}}, \quad W = \sum_{i=1}^{m} \alpha_i \varphi(x_i), \quad v \perp \varphi(x_i) \ \forall i$$

we wanna show the v is $\underline{\underline{0}}$! thus we have:

$\|w_s\|^2 = \|W\|^2 + \|v\|^2$, i.e $\|W\| \leq \|w_s\|$, Since R is assumed to be strictly increasing, we get $R(\|W\|) \leq R(\|w_s\|)$, & particularly if $v \neq 0$, then $R(\|W\|) < R(\|w_s\|)$.

Since $\boxed{v \perp}\ \underline{\varphi(x_i)\ \forall i}$, we have also $f(w\varphi(x_1 + b_s, \dots, w.\varphi(x_m)+b_s)$

$\longrightarrow v.\varphi(x_i) = 0$

$= f(w_s.\varphi(x_1)+b_s, \dots, w_s.\varphi(x_m)+b_s)$

Thus if $v \neq 0$, then objective function $(f+R)$ would be smaller at w then at w_s – a contradiction since $v = 0$ #

The Kernel trick

- The coefficients $\boldsymbol{\alpha}_s \in \mathbb{R}^m$ can be very efficiently computed using a so called kernel.
- To motivate that, we notice that by the representer theorem we have

$$\mathbf{w}_s \cdot \psi(\mathbf{x}_i) = \sum_{j=1}^{m} \alpha_{s,j} \; [\psi(\mathbf{x}_j) \cdot \psi(\mathbf{x}_i)] , \qquad \|\mathbf{w}_s\|^2 = \sum_{i,j=1}^{m} \alpha_{s,j}\alpha_{s,i} \; [\psi(\mathbf{x}_j) \cdot \psi(\mathbf{x}_i)] .$$

corespond to suppor vector

orignal feature

Definition 4.9:

Given a feature map $\psi\colon \mathcal{X} \to \mathcal{F}$ we define a kernel $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$K(\mathbf{x}, \mathbf{y}) := \psi(\mathbf{x}) \cdot \psi(\mathbf{y})$$

- Thus, we can express everything by K, e.g.,

$$\mathbf{w}_s \cdot \psi(\mathbf{x}_i) = \sum_{j=1}^{m} \alpha_{s,j} K(\mathbf{x}_i, \mathbf{x}_j), \qquad h_s(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_{s,i} K(\mathbf{x}_i, \mathbf{x}) + b_s \right)$$

Hence, by introducing the symmetric and positive (semi-)definite (Gram) matrix

b cuz Kernel Function is symmetric

like this

$$\mathbf{K} := \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \dots & K(\mathbf{x}_1, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_m, \mathbf{x}_1) & \dots & K(\mathbf{x}_m, \mathbf{x}_m) \end{pmatrix} \in \mathbb{R}^{m \times m}$$

we get with $\mathbf{K}_{i\bullet}$ denoting the i-th row of $\mathbf{K}$

Hard Kernel SVM-rule

Compute: $h_s(\mathbf{x}) := \operatorname{sgn}\left(\sum_{i=1}^m \alpha_{s,i} K(\mathbf{x}_i, \mathbf{x}) + b_s\right)$ by

$$(\boldsymbol{\alpha}_s, b_s) = \underset{(\boldsymbol{\alpha}, b) \in \mathbb{R}^{m+1}}{\operatorname{argmin}} \ \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \qquad \text{subject to:} \qquad y_i \left(\mathbf{K}_{i\bullet} \boldsymbol{\alpha} + b\right) \geq 1 \qquad \forall i.$$

✓ (and mostly) ⇝ more seprable (linearly seprable)

Soft Kernel SVM-rule

Compute: $h_s(\mathbf{x}) := \operatorname{sgn}\left(\sum_{i=1}^m \alpha_{s,i} K(\mathbf{x}_i, \mathbf{x}) + b_s\right)$ by

ith (a row)

$$(\boldsymbol{\alpha}_s, b_s) \in \underset{(\boldsymbol{\alpha}, b) \in \mathbb{R}^{m+1}}{\operatorname{argmin}} \ \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} + \frac{1}{m} \sum_{i=1}^m \max\left\{0, 1 - y_i \left(\mathbf{K}_{i\bullet} \boldsymbol{\alpha} + b\right)\right\}$$

αs

for bias

⇒ Eest → good down

Eopt → Max min

Eapp ⇒ 0

Notes

- Due to the kernel trick we can simply choose a kernel function K on $\mathbb{R}^d \times \mathbb{R}^d$ and apply the kernel SVM-rule without working with ψ of $\mathcal{F}$ explicitly!

- Inparticular, we can define again support vectors $\mathbf{x}_i \in \mathcal{X}$: for the hard kernel SVM-rule we have

$$h_s(\mathbf{x}) = \operatorname{sgn}\left(\sum_{j\in J} \alpha_{s,j} K(\mathbf{x}_j, \mathbf{x}) + b_s\right), \qquad J := \{i \colon y_i\left(\mathbf{K}_{i\bullet}\boldsymbol{\alpha}_s + b_s\right) = 1\}\,.$$

for the support vector class

- An analogous representation holds for the soft kernel SVM rule with $J := \{i \colon y_i\left(\mathbf{K}_{i\bullet}\boldsymbol{\alpha}_s + b_s\right) \leq 1\}$.

- The kernel K or $K(\mathbf{x}, \mathbf{y})$ can be thought of a measure of similarity between the original features $\mathbf{x}, \mathbf{y}$

Polynomial kernel

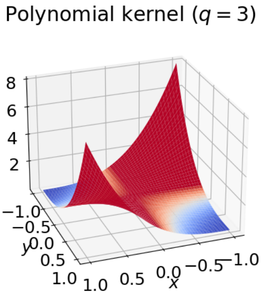
→ for Finitly Dimensional

- Let $\mathcal{X} = \mathbb{R}^d$ and for a $q \in \mathbb{N}$ set $K(\mathbf{x}, \mathbf{y}) := (1 + \mathbf{x} \cdot \mathbf{y})^q$.

- As a feature mapping $\psi \colon \mathbb{R}^d \to \mathbb{R}^{(1+d)^q}$ we then define

$$\psi(\mathbf{x}) := \left(\prod_{i=1}^{q} x_{j_i} : \mathbf{j} = (j_1, \dots, j_q) \in \{0, \dots, d\}^q \right),$$

  i.e., ψ collects the values of all multivariate monomials of $\mathbf{x}$ of total degree q



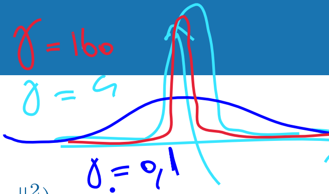
- It then holds with the Euclidean inner product in $\mathbb{R}^d$ or $\mathcal{F} = \mathbb{R}^{(1+d)^q}$

$$K(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^q = \sum_{\mathbf{j} \in \{0, \dots, d\}^q} \prod_{i=1}^{q} x_{j_i} y_{j_i} = \psi(\mathbf{x}) \cdot \psi(\mathbf{y}).$$

- This kernel thus allows us to learn hypotheses $h(\mathbf{x}) = \operatorname{sgn}(p(\mathbf{x}))$ with multivariate polynomials $p \colon \mathbb{R}^d \to \mathbb{R}$ of total degree q.

Gaussian kernel

RBF: Kernel



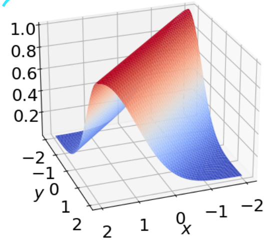
- A very popular kernel on $\mathcal{X} = \mathbb{R}^d$ is

$$K(\mathbf{x}, \mathbf{y}) := \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$$

density of mutinomial

with scaling parameter $\gamma > 0$.

Gaussian kernel ($\gamma = 0.5$)



- The Gaussian kernel evaluates the similarity of the features with respect to their distance $\|\mathbf{x} - \mathbf{y}\|$, where the distance enters nonlinearly.

→ infinite Dimensional

- The feature map contains the product of $e^{-\gamma\|\mathbf{x}\|^2}$ and any multivariate monomials of $\mathbf{x}$.
- The Gaussian kernel is also called the RBF kernel, where RBF stands for radial basis function.
- The choice of the scaling parameter $\gamma > 0$ can have a large influence on the result of kernel SVM rules: Too small $\gamma \ll 1$ may not allow a good fit to the data and too large $\gamma \gg 1$ will lead to overfitting

([illegible]) it's concentrated more

Example: Synthetic dataset (black line) — True decsion boundry

- We consider the example from the beginning and learn a hypothesis by the soft kernel SVM rule.

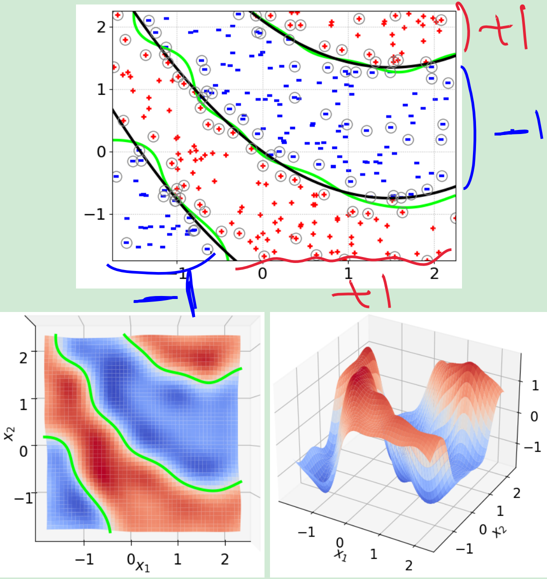
- We use a Gaussian kernel with $\gamma = 4, \quad \lambda = \frac{1}{2m}$.

- A very good fit is obtained, see comparison of the partition lines of the domains with $h_s(\mathbf{x}) = \pm 1$ and $h^\dagger(\mathbf{x}) = \operatorname{sgn}(\sin(\frac{1}{2}x_1^2 - \frac{3}{2}x_2 - \frac{3}{2}x_1))$

- The learned function $f_s(\mathbf{x})$, $h_s = \operatorname{sgn} \circ f_s$, is shown on the righthand side with support vectors $\mathbf{x}_j$ circled in grey above:

$$f_s(\mathbf{x}) = \sum_{j=1}^{J} \alpha_{s,j} K(\mathbf{x}, \mathbf{x}_j) + b_s$$

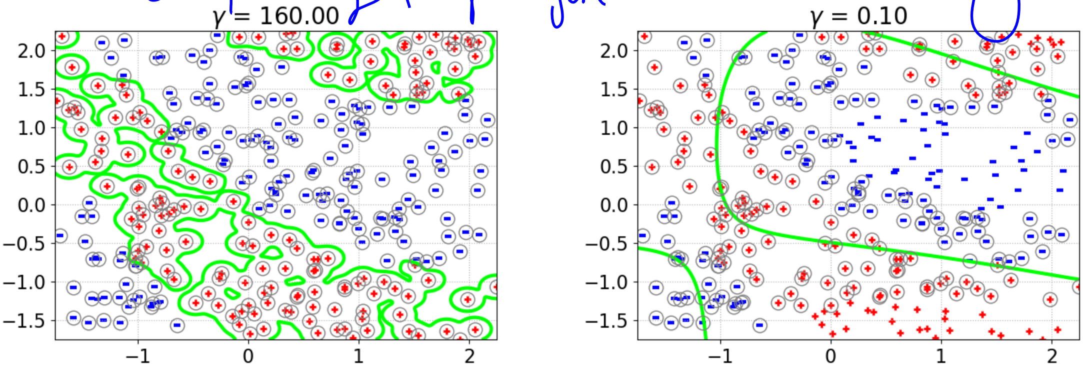
- The green lines correspond to $\{\mathbf{x} \colon f_s(\mathbf{x}) = 0\}$.



Influence of scale parameter γ

Overfitting + high E_{gen}

underfitting

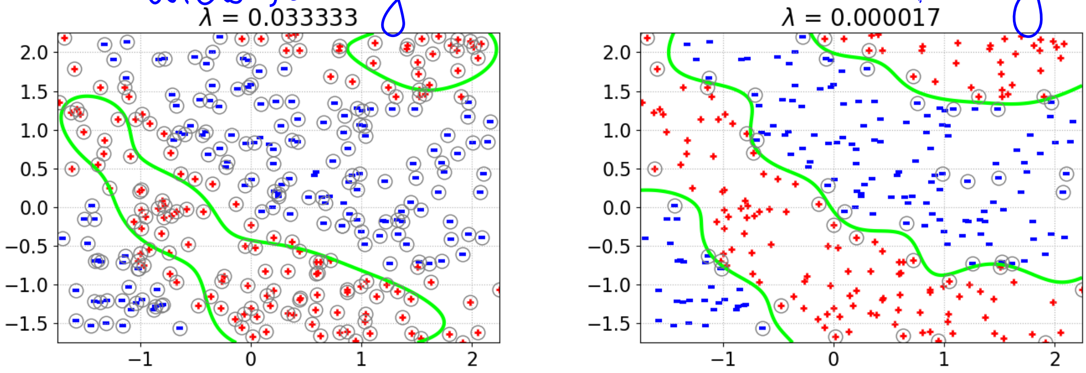


- The green lines display the decision boundaries of $h_s(\mathbf{x}) = \pm 1$, the circled data points are again the support vectors.
- The choice of γ has great influence on the learned hypothesis h_s.
- An adaptive choice or estimation of γ is generally advisable.

Influence of regularization parameter λ

underfitting

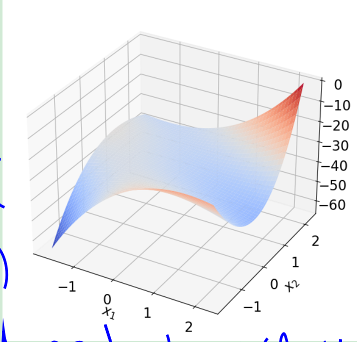
overfitting



- The green lines display the decision boundaries of $h_s(\mathbf{x}) = \pm 1$, the circled data points are again the support vectors.
- As in Section 3.4: Too large λ leads to underfitting and very small λ leads to overfitting.
- The *default value* in `scikit-learn` is set to $\lambda = \frac{1}{2m}$.

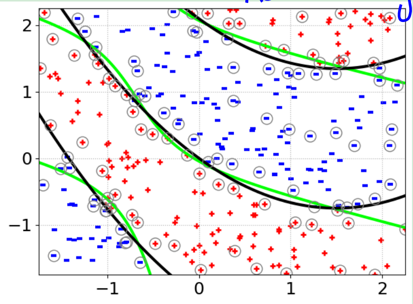
Example: Polynomial kernel for synthetic dataset → a bit restricted in shapes not as good as RBF

- We choose now $K(\mathbf{x},\mathbf{y}) = (1 + \mathbf{x}\cdot\mathbf{y})^q$ and obtain:



(تابع یعنی این تابع

$\sin(K(x,y)))$



→ strange not good shape (limited)

- Left: the learned function f_s; right: data, support vectors, and the decision boundaries $f_s(\mathbf{x}) = 0$
- scikit-learn chooses $q = 3$ and $\lambda = \frac{1}{2m}$ as default and allows for scaling parameters by

default value $K(\mathbf{x},\mathbf{y}) = (c_0 + \gamma\mathbf{x}\cdot\mathbf{y})^q$

it cannot have a good about approximation than RBF

Excurse: Reproducing kernel Hilbert spaces (RKHS) (not come very hard in exam)

- We want to learn more about the underlying hypothesis class of kernel methods

$$\mathcal{H}_K = \{h(\mathbf{x}) = \operatorname{sgn}(f(\mathbf{x}) + b) \mid f \in \mathcal{F}_K, b \in \mathbb{R}\},$$

  i.e., what is the corresponding function space $\mathcal{F}_K$?

- Recall, that the output of kernel SVM-rules is $h_s(\mathbf{x}) = \operatorname{sgn}(f_s(\mathbf{x}))$ where for a chosen kernel K

$$f_s(\mathbf{x}) = \sum_{j=1}^{J} \alpha_{s,j} K(\mathbf{x}, \mathbf{x}_j) + b$$

- Here J can become arbitrarily large. Thus, $\mathcal{F}_K$ should include the limit for $J \to \infty$!
- What are the properties of the resulting function space $\mathcal{F}_K$?
- In particular, which kind of functions $f^\dagger$ can be approximated well by $f \in \mathcal{F}_K$?

Assumption

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and let $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite function.

The latter means that for any $n \in \mathbb{N}$ and any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ the corresponding (Gram) matrix

$$\mathbf{K} := \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \ldots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & \ldots & K(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

is symmetric and positive semidefinite.

- Given a positive semi-definite $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ we define a vector space F_K of functions $f\colon \mathcal{X} \to \mathbb{R}$ by

$$F_K := \left\{ f(\mathbf{x}) := \underline{\sum_{i=1}^{n} a_i K(\mathbf{x}, \mathbf{x}_i)}\colon n \in \mathbb{N},\ a_i \in \mathbb{R},\ \mathbf{x}_i \in \mathcal{X} \right\}$$

finite linear combination → since it aint inerproduct it aint compelete

- We can equip the vector space F_K

$$F_K := \left\{ f(\mathbf{x}) := \sum_{i=1}^{n} a_i K(\mathbf{x}, \mathbf{x}_i) \colon n \in \mathbb{N},\ a_i \in \mathbb{R},\ \mathbf{x}_i \in \mathcal{X} \right\}$$

with the following inner product:

$$\langle f, g \rangle_K := \sum_{i=1}^{m} \sum_{j=1}^{n} a_i\ b_j\ K(\mathbf{x}_i, \mathbf{y}_j), \quad \text{where} \quad f(\cdot) = \sum_{i=1}^{m} a_i K(\cdot, \mathbf{x}_i), \quad g(\cdot) = \sum_{j=1}^{n} b_j K(\cdot, \mathbf{y}_j).$$

⇒ with this you take compeletion for the incompelet kernel method via

Definition 4.10:

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive semi-definite. Then, the reproducing kernel Hilbert space $\mathcal{F}_K$ is the completion of F_K w.r.t. $\langle \cdot, \cdot \rangle_K$.

Properties of RKHS

why reproducing → any f is recovered if you take via inner product with kernel

Proposition 4.11: Reproducing property

For any $f \in \mathcal{F}_K$ we have $f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x})\rangle_K, \quad \mathbf{x} \in \mathcal{X}$.

Theorem 4.12:

Any positive definite function $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^d$ is a kernel, i.e., there exists a Hilbert space $\mathcal{F}_K$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}_K}$ and a feature map $\psi_K\colon \mathcal{X} \to \mathcal{F}_K$ such that

$$K(\mathbf{x}, \mathbf{y}) = \langle \psi_K(\mathbf{x}), \psi_K(\mathbf{y})\rangle_{\mathcal{F}_K}.$$

- This theorem tells us again that working with kernels K and working with feature maps ψ is equivalent.
- The existence of $\mathcal{F}_K$ is clear ($\mathcal{F}_K$ being the RKHS of K) and the feature map $\psi_K\colon \mathcal{X} \to \mathcal{F}_K$ is then $\psi_K(\mathbf{x}) := K(\cdot, \mathbf{x})$ since

$$\langle \psi_K(\mathbf{x}), \psi_K(\mathbf{y})\rangle_{\mathcal{F}_K} = \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{y})\rangle_{\mathcal{F}_K} = K(\mathbf{x}, \mathbf{y})$$

Proof of proposition: 4.11:

we have $K(\underset{\substack{\downarrow \\ \text{argument for the function } x}}{\cdot}, x) : X \longrightarrow \mathbb{R}$ belongs to $\mathcal{F}_K$, since:

$$K(\cdot, x) = 1 \cdot K(\cdot, x) = \sum_{i=n}^{1} 1 \cdot K(\cdot, \underset{x}{\underset{\shortparallel}{x_1}}) .$$

assume $f = \sum_{i=1}^{m} a_i K(\cdot, x_i)$, then $\langle f, K(\cdot, x) \rangle_K$

$$= \sum_{i=1}^{m} a_i \cdot 1 \cdot \underbrace{K(x_i, x)}_{\text{has to be symmetric}} = \sum_{i=1}^{m} a_i K(x, x_i) = f(x) \quad \#$$

- An RKHS is no “exotic” space:

Theorem 4.13:

Let $\mathcal{F}$ be a Hilbert space of functions $f \colon \mathcal{X} \to \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Then, there exists a positiv semidefinite $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\mathcal{F} = \mathcal{F}_K$ if and only if

$$\mathcal{F} \ni f \mapsto f(\mathbf{x}) \in \mathbb{R}$$

is continuous for every $\mathbf{x} \in \mathcal{X}$.

Proof: See, e.g., Chapter 4 in I. Steinwart and A. Christmann: "Support Vector Machines" (2009)

- Hence, for suitable function spaces $\mathcal{F}$ such as Sobolev spaces of sufficient high regularity we can find the correct kernel K such that the kernel SVM rule yields a hypothesis from

  $$\mathcal{H} = \operatorname{sgn} \circ \mathcal{F}.$$

  see, e.g., Reproducing kernels of Sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability (2017) for more details

Universal Approximation

- There are many approaches to study the approximation error of given hypothesis classes $\mathcal{H}$

$$\varepsilon_{\mathsf{app}}(\mathcal{H}) = \inf_{h \in \mathcal{H}} \mathcal{R}_\mu(h).$$

- We focus on a concept or property of $\mathcal{H}$ which yield for suitable μ that

$$\varepsilon_{\mathsf{app}}(\mathcal{H}) = 0.$$

Definition 4.14:

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact. We say a class $\mathcal{F}$ of real-valued functions $f\colon \mathcal{X} \to \mathbb{R}$ is a universal approximator or satisfies universal approximation if $\mathcal{F}$ is *dense* in the space $\mathcal{C}(\mathcal{X})$ of all continuous functions $g\colon \mathcal{X} \to \mathbb{R}$. This means, for any $\underline{g \in \mathcal{C}(\mathcal{X})}$ and any $\epsilon > 0$ there exists a $f_\epsilon \in \mathcal{F}$ such that

Can be chosen arbitary (continuous function)

$$\|g - f_\epsilon\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} |g(\mathbf{x}) - f_\epsilon(\mathbf{x})| \underline{\underline{\leq \epsilon}}.$$

- **Weierstrass theorem:** The class of polynomials

$$\mathcal{F} = \left\{ f(x) = \sum_{k=0}^{n} a_k x^k : n \in \mathbb{N},\ a_k \in \mathbb{R} \right\}$$

is a universal approximator on $\mathcal{X} = [a, b] \subset \mathbb{R}$.

K. Weierstrass
(1815–1897)

- It would be beneficial if the RKHS underlying a kernel SVM is a universal approximator.

Definition 4.15:

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel. We call K universal if the associated RKHS $\mathcal{F}_K$ is a universal approximator, i.e., dense in $\mathcal{C}(\mathcal{X})$.

Theorem 4.16:

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact and let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ satisfy universal approximation. not continuous

Then we have for any data distribution μ given by $(X, h^\dagger(X))$ with (measurable) hypothesis $h^\dagger \colon \mathcal{X} \to \mathcal{Y}$

1. in case of **classification**, i.e., $|\mathcal{Y}| = 2$, w.r.t. 0-1 loss and

$$\mathcal{H} = \operatorname{sgn} \circ \mathcal{F} = \{h(\mathbf{x}) = \operatorname{sgn}(f(\mathbf{x})) \colon f \in \mathcal{F}\} \quad \xrightarrow{\varepsilon_{\text{app}}} 0$$

2. or in case of **regression**, i.e., $\mathcal{Y} = \mathbb{R}$, and $\mathcal{H} = \mathcal{F}$ w.r.t. bounded squared loss

$$\ell(h, (\mathbf{x}, y)) = \max\{c, |y - h(\mathbf{x})|^2\}$$

that

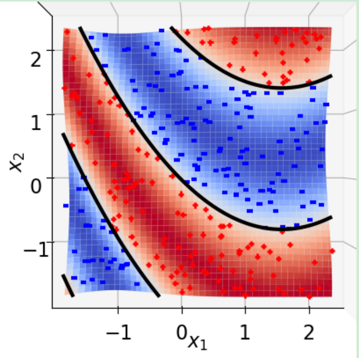
$$\varepsilon_{\text{app}}(\mathcal{H}) = 0.$$

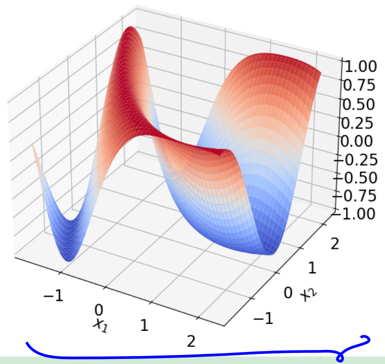
I.e., if we choose a universal kernel K, we have $\varepsilon_{\text{app}}(\mathcal{H}_K)$ for $\mathcal{H}_K = \operatorname{sgn} \circ \mathcal{F}_K$.

Example 4.17:

The training data set from the beginning was drawn according to (X, Y) where $X \sim \mathrm{U}(\mathcal{X})$ with $\mathcal{X} = [-1.75, 2.25]^2$ and $Y = h^\dagger(X)$ with

$$h^\dagger(\mathbf{x}) = \operatorname{sgn}\left(\sin\left(\frac{1}{2}x_1^2 - \frac{3}{2}x_2 - \frac{3}{2}x_1\right)\right)$$





how to approxim
this with a
powerful
kernel

I.e.., we are in the setting of the previous proposition. But were the employed kernels universal?

RBF

- Well-known universal approximation theorem from analysis: Stone–Weierstrass theorem.

- which tells us that we can approximate any continuous $f\colon \mathcal{X} \to \mathbb{R}$ on compact $\mathcal{X} \subset \mathbb{R}^d$ arbitrarily well by polynomials (of arbitrary degree!)

- The following can be seen as an analogue in terms of kernel functions:

Theorem 4.18:

Consider a kernel $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined on $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d \colon \|\mathbf{x}\|^2 < r\}$, $r > 0$. If K is given by

$$K(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} \cdot \mathbf{y}) \qquad \text{or} \qquad K(\mathbf{x}, \mathbf{y}) = \frac{k(\mathbf{x} \cdot \mathbf{y})}{\sqrt{k(\mathbf{x}, \mathbf{x})\ k(\mathbf{y}, \mathbf{y})}}$$

for an analytic function $k\colon (-r, r) \to \mathbb{R}$ with

$$k(x) = \sum_{n=1}^{\infty} k_n x^n \qquad \text{where} \qquad \underline{k_n > 0}\ \ \forall n \in \mathbb{N},$$

then K is universal.

Proof: See, e.g., "On the Influence of the Kernel on the Consistency of Support Vector Machines" (2001)

Example 4.19: Universal Gaussian kernel

***

The Gaussian kernel satisfies the assumptions of the previous theorem due to

RBF kernel

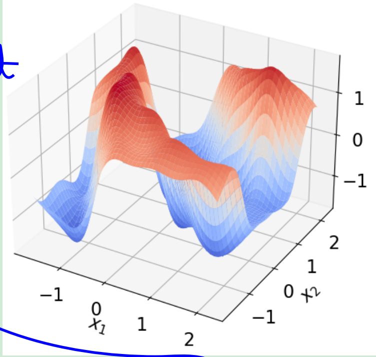
$$K(\mathbf{x},\mathbf{y}) \;=\; \underline{\exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)} \;=\; \underline{\exp(-\gamma\|\mathbf{x}\|^2)\exp(-\gamma\|\mathbf{y}\|^2)\exp(2\gamma(\mathbf{x}\cdot\mathbf{y}))}$$

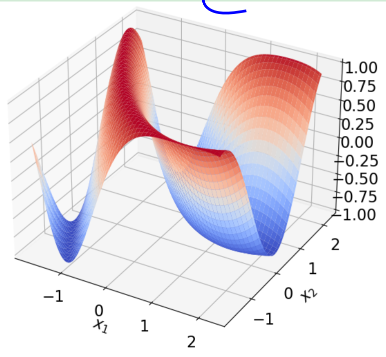
and, thus, by using $k(x) = \exp(2\gamma\, x) = \underline{\sum_{n=0}^{\infty} \frac{(2\gamma)^n}{n!} x^n}$.

اسى دمج

This will not end to 0

Then





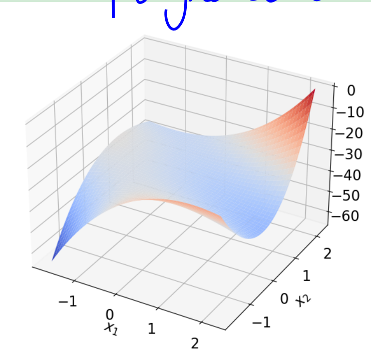
Thus, the Gaussian kernel is universal on compact $\mathcal{X}$ and we have $\varepsilon_{\text{app}}(\mathcal{H}_K) = 0$ here!

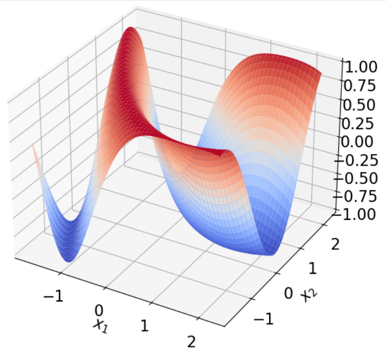
Example 4.20: Non-Universal polynomial kernel

The polynomial kernel does not satisfy the assumptions of the previous theorem due to

$$K(\mathbf{x},\mathbf{y}) \;=\; (1+\mathbf{x}\cdot\mathbf{y})^q \;=\; k(\mathbf{x}\cdot\mathbf{y}), \qquad k(x) = \sum_{n=0}^{q} \binom{q}{n} x^n$$

True relations polynomial
will be
here





Although, the theorem states only a sufficient condition, the polynomial kernel is indeed not universal!

Kernel SVM: Summary

The soft kernel SVM rule using universal kernels such as the Gaussian kernel

- has an approximation error of size zero $\varepsilon_{\text{app}} = 0$ if an underlying true continuous hypothesis $h^\dagger = \operatorname{sgn}(f^\dagger)$ exists,
- has a estimation error which can be controlled independently of feature dimension $\dim(\mathcal{F})$ or $\mathrm{VCD}(\mathcal{L}_\mathcal{F})$ for suitable μ (cf. Theorem 4.7)
- has an optimization error which is easy to control since h_s can be computed efficiently via convex optimization (i.e., no local minima!)

This explains the success of kernel SVMs in the 1990's and early 2000's.

Further fun facts about kernels

- In statistics in kernel density estimation of probability density functions we use kernels located at data points $x_i \in \mathbb{R}$. Here, an optimal kernel is the Epanechnikov kernel

$$K(x,y) = \begin{cases} \frac{3}{4}(1-(x-y)^2), & \text{if } |x-y| \leq 1 \\ 0 & \text{else.} \end{cases}$$

- In geostatistics any covariance function of spatial processes or random fields Z is a kernel, e.g., Matérn covariances

$$c(\mathbf{x},\mathbf{y}) = \sigma^2 \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{\rho}\right) \quad = \quad \mathbb{C}\mathbf{ov}[Z(\mathbf{x}), Z(\mathbf{y})]$$

- For approximating functions $f \colon \mathcal{X} \to \mathbb{R}$ one can also study the radial basis function interpolant

$$\hat{f}_m(\mathbf{x}) = \sum_{i=1}^{m} a_i K(\mathbf{x}, \mathbf{x}_i) \qquad \text{such that} \qquad \hat{f}_m(\mathbf{x}_i) = f(\mathbf{x}_i) \quad \forall i = 1, \dots, m.$$

Based on the smoothness of f one can then derive convergence rates for $\hat{f}_m \to f$ in suitble norms

Take home table of Chapter 4

Evaluate the four linear methods we have learned so far, i.e., complete the table below by inserting in each cell either

$$1 \text{ (best)}, \quad 2 \text{ (medium)}, \quad 3 \text{ (worst)}$$

for the performance regarding the corresponding error:

Method	ε_{app}	ε_{est}	ε_{opt}
Perceptron			
Logistic regression			
Hard / soft SVM rule			
Kernel SVM rules			