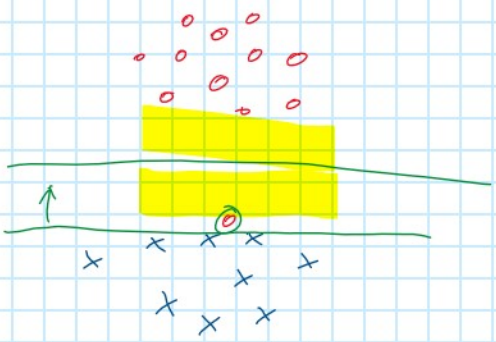
7. HW Soft-Margin SVM & Kernels

Mittwoch, 12. Dezember 2018 15:57

1 Soft-margin SVM

Problem 1: Assume that we have a linearly separable dataset $\mathcal{D}$, on which a soft-margin SVM is fitted. Is it guaranteed that all training samples in $\mathcal{D}$ will be assigned the correct label by the fitted model? Explain your answer.

No

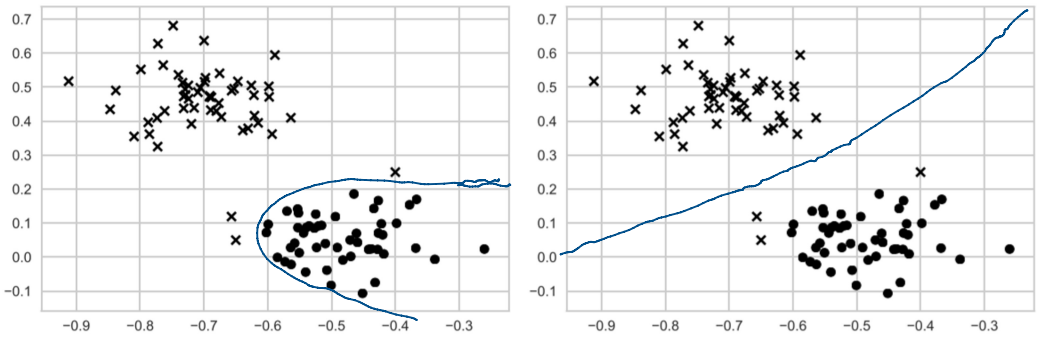


Problem 2: Why do we need to ensure that $C > 0$ in the slack variable formulation of soft-margin SVM? What would happen if this was not the case?

$C = 0$: $\min f_0 = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{N}\xi_i$, $C = 0$, $\xi_i \geq 1 - y_i(\vec{w}^T\vec{x}_i + b)$

$C < 0$: $\min f_0 = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{N}\xi_i$, $C < 0$, $\xi_i \to \infty$

Problem 3: Sketch the decision boundary of an SVM with a quadratic kernel (polynomial with degree 2) for the data in the figure below, for two specified values of the penalty parameter C. (The two classes are denoted as •'s and ×'s.)



(a) $C = 10^{10}$ (b) $C = 10^{-10}$

Explain the reasoning behind your sketch of the decision boundary for both cases (one sentence for each plot).

2 Kernels

Problem 4: Show that for $N \in \mathbb{N}$ and $a_i \geq 0$, with $i \in [0, N]$ the function

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{N} a_i \left(\boldsymbol{x}_1^T \boldsymbol{x}_2\right)^i + a_0$$

is a valid kernel.

a_0: $\phi(\vec{x}) = \sqrt{a_0}$

$\phi(\vec{x}_1) \cdot \phi(\vec{x}_2) = a_0 \Rightarrow$ kernel

Σ of kernels is kernel

if $a_i > 0$: $a_i \cdot k(\vec{x}_1, \vec{x}_2)$ is kernel

$a_i = 0$: $\phi(\vec{x}) = 0$, $\phi(\vec{x}_1)\phi(\vec{x}_2) = 0$ is kernel

$k(\vec{x}_1, \vec{x}_2)^i$, $i \in \mathbb{N}$

Proof by induction: Step: $k(\vec{x}_1, \vec{x}_2)^i = k(\vec{x}_1, \vec{x}_2)^{i-1} \cdot k(\vec{x}_1, \vec{x}_2)$

kernel ✓ (ind. hyp.)

product of kernels $\Rightarrow$ kernel

$k(\vec{x}_1, \vec{x}_2)^1 = k(\vec{x}_1, \vec{x}_2)$ is kernel ✓

$\vec{x}_1^T \vec{x}_2$: $\phi(\vec{x}_1) = \vec{x}_1$, scalar product: kernel ✓

Problem 5: Find the feature transformation $\phi(x)$ corresponding to the kernel

$$k(x_1, x_2) = \frac{1}{1 - x_1 x_2},$$

with $x_1, x_2 \in (0, 1)$.

Hint: Consider an infinite-dimensional feature space.

$$k(\vec{x}_1, \vec{x}_2) = \phi(\vec{x}_1)^T \phi(\vec{x}_2) = \underbrace{\sum_{i=0}^{\infty} \phi_i(\vec{x}_1) \cdot \phi_i(\vec{x}_2)}_{\text{Series}}$$

(think hard) $\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i$ if $x \in (0,1)$

$$\phi(x) = (1, x^1, x^2, x^3, \dots)^T$$

Problem 6: Consider the following algorithm.

Algorithm 1: Counting something

input : Character string x of length m (one based indexing)
input : Character string y of length n (one based indexing)
output: A number $s \in \mathbb{R}$

```
s ← 0;
for i ← 1 to m do
    for j ← 1 to n do
        if x[i] == y[j] then
            s ← s + 1;
```

a) Explain, in no more than two sentences, what the above algorithm is doing.

b) Let $\mathcal{S}$ denote the set of strings over a finite alphabet of size v. Define a function $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ as the output of running algorithm 1 on a pair of strings x, y. Show that $k(x, y)$ is a valid kernel.

a) Sum of how many times each character c from string x appears in string y

Example: $x = ABBA$ $\#A=2$ $\#B=2$

$y = ACDC$ $\#A=1, \#C=2, \#D=1$

result: $2 \cdot 1 + 2 \cdot 0 + 0 \cdot 2 + 0 \cdot 1 = 2$

b) Count occurrences of each letter in x, save in $\vec{v}_x$

" " in y, save in $\vec{v}_y$

"Counting occurrences" $\hat{=} \phi \Rightarrow \phi(x) = \vec{v}_x$

$$k(\vec{x},\vec{y}) = \phi(x)^T \cdot \phi(y) = \sum_{i=1}^{N} v_{x,i} \cdot v_{y,i}$$

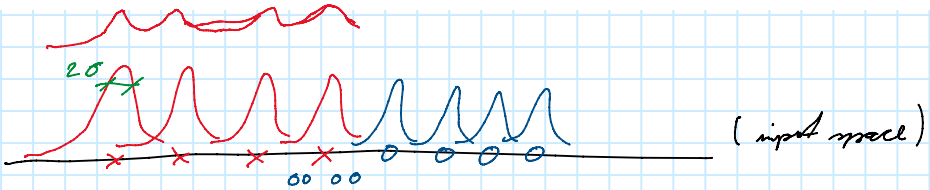
Scalar prod. of feature maps <=> kernel ✓

3 Gaussian kernel

Problem 7: Can any *finite* set of points be linearly separated in the feature space of the Gaussian kernel

$$k_G(x_1, x_2) = \exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right),$$

if σ can be chosen freely?



$\sigma \to 0$ $\quad k(x_1,x_2) = \begin{cases} 1 & \text{if } x_1 = x_2 \\ 0 & \text{else} \end{cases}$

Kernel matrix: I

Correct classification:

$$y_i\left(\vec{w}^T \phi_G(x_i) + b\right) > 0 \quad \forall i$$

$$\vec{w} = \sum_j y_j \alpha_j \phi_G(x_j)$$

$$y_i\left(\sum_j y_j \alpha_j \underbrace{\phi_G(x_i)^T \phi_G(x_j)}_{k(x_i,x_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{else} \end{cases}} + b\right) > 0$$

$$y_i\left(y_i \alpha_i + b\right) = y_i^2 \alpha_i + y_i b > 0$$

Choose $b=0$: $\quad y_i^2 \alpha_i > 0$

$y_i \in \{-1, 1\} \Rightarrow y_i^2 = 1$

$\Rightarrow$ Choose $\alpha_i > 0$

$\Rightarrow$ always fulfilled

$\Rightarrow$ σ small enough $\Rightarrow$ always lin. separable

BUT: Overfitting!