Soft-margin SVM and Kernels

Machine Learning - Prof. Dr. Stephan Günnemann

Leonardo Freiherr von Lerchenfeld

December 16, 2017

Contents

1 Soft-margin SVM

1.1 Problem 1

For a linearly separable dataset, a hard-margin SVM can be applied. Soft-margin SVM might find a large margin, because it also optimizes the margin. If the datapoints of *class* - and *class* + have in general a large distance to each other, but there is a single datapoint from *class* - very close to the datapoints from *class* +, then it is very likely that this particular point **will not be correctly labeled.**

1.2 Problem 2

From $\alpha_i = C - \mu_i$ and dual feasibility $\alpha_i \geq 0$, we get

$$0 \leq \alpha_i \leq C$$

Therefore, we need to ensure that $C > 0$
If $C = 0$, then $\forall i \ \alpha_i = 0$
If $C < 0$, then dual feasibility does not hold anymore.

2 Kernels

2.1 Problem 3

To show that for $c \geq 0$ and $d \in \mathbf{N}^+$ the function $K(x, y) = (x^T y + c)^d$ is a valid kernel, we use **Techniques for constructing new kernels**: Given valid kernels K_1, K_2 and any positive constant $\alpha \geq 0$, the following new kernels will also be valid:

$$\begin{aligned} K(x, y) &= K_1(x, y)K_2(x, y) \qquad (1) \\ K(x, y) &= K_1(x, y) + \alpha \qquad (2) \end{aligned}$$

While (1) is given (from the Practical Session), (2) has to be proven. Let Φ_1 denote a feature map of K_1. Then, using the feature map $\Phi : x \mapsto [\Phi_1(x), \sqrt{\alpha}]^T$, we have

$$\langle \Phi(x), \Phi(y) \rangle = \langle \Phi_1(x), \Phi_1(y) \rangle + \alpha = K_1(x, y) + \alpha = K(x, y)$$

Let‘s start the show

$$\begin{aligned} K(x, y) &= (x^T y + c)^d \\ &= \prod_{i=1}^{d} x^T y + c \end{aligned}$$

Using Rule (1) we have to prove that $x^T y + c$ is a valid kernel. We know that $x^T y$ is the linear kernel and a valid kernel as discussed in the lecture. Rule (2) states that adding a positive constant is valid. Hence, $K(x, y) = (x^T y + c)^d$ is a valid kernel.

3 Gaussian kernel

3.1 Problem 4

We cannot directly apply $\Phi_\infty(x)$ to data, because an infinite feature space requires infinite storage space.

3.2 Problem 5

Taylor series of $e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!}$

$$\begin{aligned}
\Phi_\infty(x) &= exp\left(-\frac{x^2}{2\sigma^2}\right)\left\{1, \frac{x}{\sigma}, \frac{1}{\sqrt{2}}(\frac{x}{\sigma})^2, \ldots, \frac{1}{\sqrt{i!}}(\frac{x}{\sigma})^i, \ldots\right\} \\
K(x,y) &= \Phi_\infty(x)^T\Phi_\infty(y) \\
&= \langle exp\left(-\frac{x^2}{2\sigma^2}\right)\left\{1, \frac{x}{\sigma}, \frac{1}{\sqrt{2}}(\frac{x}{\sigma})^2, \ldots, \frac{1}{\sqrt{i!}}(\frac{x}{\sigma})^i, \ldots\right\}, \\
& \quad exp\left(-\frac{y^2}{2\sigma^2}\right)\left\{1, \frac{y}{\sigma}, \frac{1}{\sqrt{2}}(\frac{y}{\sigma})^2, \ldots, \frac{1}{\sqrt{i!}}(\frac{y}{\sigma})^i, \ldots\right\}\rangle \\
&= exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)\sum_{i=0}^{\infty}\frac{1}{i!}\left(\frac{xy}{\sigma^2}\right)^i \\
&= exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) exp\left(\frac{xy}{\sigma^2}\right) \\
&= exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)
\end{aligned}$$

As the chapter suggests, we have a Gaussian kernel, where overfitting occurs with a too little variance σ.

3.3 Problem 6

Yes, any finite set can be linearly separated in the feature space. However, when when we have outliers, then we must set σ very small, which comes at the cost of generality.

4 Kernelized k-nearest neighbors

4.1 Problem 7

The distance to sample in feature space is given by

$$d(x,y) = ||\Phi(x) - \Phi(y)||_2$$

As we are not interested in the absolute values of the distance, but on the relative values, we can take the squared distance

$$\begin{aligned}
d(x, x^{s_i})^2 &= \left(\Phi(x) - \Phi(x^{s_i})\right)^T\left(\Phi(x) - \Phi(x^{s_i})\right) \\
&= \Phi(x)^T\Phi(x) - 2\left(\Phi(x) - \Phi(x^{s_i})\right) + \Phi(x^{s_i})^T\Phi(x^{s_i}) \\
&= K(x,x) - 2K(x, x^{s_i}) + K(x^{s_i}, x^{s_i})
\end{aligned}$$