Machine Learning Worksheet Solution 8

Julius Jankowski

1 Soft-margin SVM

Problem 1:

Yes, it is still guaranteed that the resulting hyperplane will seperate two linearly seperable sets. When looking at the hinge loss formulation, we can see that we will minimize the sum over the slack variables $C \cdot \sum_{i=1}^{N} \xi$ which is directly coupled to the classification error. As long as C is positive and larger than 0, this will lead to a seperating hyperplane. Since the gradient of the regularization term $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ is orthogonal to the other term, this does not change the argumentation.

Problem 2:

If C is smaller than zero, this would mean that we want to maximize the sum over the slack variables. On one hand, this would lead to a switched classification because it tries to classify all samples wrong. On the other hand, the norm of $\mathbf{w}$ would tend to infinity as it would increase the slack. However this would balance with the regularitzation term.

2 Kernels

Problem 3:

Since we can reformulate the kernel function as a multiplication of multiple kernel functions, we just have to prove that the multiplied kernel function k is valid:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + c)^d = k^d \tag{2.1}$$

$$k = \Phi(\mathbf{x})^T\Phi(\mathbf{y}) = \begin{pmatrix} \mathbf{x} \\ \sqrt{c} \end{pmatrix}^T \begin{pmatrix} \mathbf{y} \\ \sqrt{c} \end{pmatrix} \tag{2.2}$$

$$\implies \Phi(\mathbf{x}) = \begin{pmatrix} \mathbf{x} \\ \sqrt{c} \end{pmatrix} \tag{2.3}$$

The fact that the kernel function k can be expressed by an inner product of a feature transformation Φ is a satisfying proof for a valid kernel.

3 Gaussian kernel

Problem 4:

We can not directly apply the infinite feature transform since we are interested in a solution in finite time.

Problem 5:

When we put the feature transform into the kernel function we get the following sum:

$$K(x,y) = \sum_{i=0}^{\infty} \Phi(x)_i \Phi(y)_i = e^{-\frac{x^2+y^2}{2\sigma^2}} \cdot \sum_{i=0}^{\infty} (\frac{xy}{\sigma^2})^i \cdot \frac{1}{i!} \tag{3.1}$$

From evaluating the Taylor series for e^x, we know that $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$ which yields:

$$K(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}} \cdot e^{\frac{xy}{\sigma^2}} = e^{-\frac{(x-y)^2}{2\sigma^2}} \tag{3.2}$$

This is basically the gaussian function applied to the distance between the two data points. The resulting kernel function shows that the overfitting characteristics of the feature transform depends on the hyperparameter σ, because for large values the function is smooth and leads to high generalization while the function tends to a dirac-function as σ tends to zero. This leads to overfitting.

Problem 6:

Following the same arguments as in problem 5, the kernel function becomes a dirac-function as σ tends to zero. This means that no sample in the training data has any impact on the decision for another sample. The result is that every finite set is seperable.

4 Kernelized k-nearest neighbors

Problem 7:

In order to reformulate the classic distance measure for k-nearest neighbours by a valid kernel function, we can incorporate the findings we made before. Since we can generalize the kernel found in problem 5 to arbitrary dimensional data:

$$K(\mathbf{x}, \mathbf{y}) = e^{-\frac{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}{2\sigma^2}} \tag{4.1}$$

We can use this as a kernel for k-nearest neighbours because the kernel is monotonic in $(\mathbf{x} - \mathbf{y})^T(\mathbf{x}-\mathbf{y})$. The search for the nearest neighbours corresponds then to the search for the highest kernel response.