Machine Learning Worksheet 7

Kernels

1 Infinite Feature Spaces

Lets define a new (infinitely) more complicated feature transformation $\phi_n : \mathbb{R} \to \mathbb{R}^n$ as follows:

$$\phi_n(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}}, \ldots, \frac{e^{-x^2/2}x^n}{\sqrt{n!}} \right\}$$

Suppose we let $n \to \infty$ and define a new feature transformation:

$$\phi_\infty(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}}, \ldots \right\}$$

You can think of this feature transformation as taking some finite feature vector and producing an infinite dimensional feature vector.

Problem 1. Can we directly apply this feature transformation to data? Is there a *finite* set of points that cannot be linearly seperated in this feature space? Explain why or why not!

Problem 2. From the lecture, we know that we can express a linear classifier using only inner products of input vectors in the transformed feature space. It would be great if we could some how use the feature space obtained by the feature transformation ϕ_∞. However, to do this, we must be able to compute the inner product of samples in this infinite vector space. We define the inner product between two *infinite* vectors a and b as the infinite sum given in the following equation:

$$k(a, b) = \sum_{i=1}^{\infty} a_i b_i$$

Now, for the above definition for ϕ_∞, what is the explicit form of $k(a, b)$? (Hint: Think of the Taylor series of e^x.) With such a high dimensional feature space, should we be concerned about overfitting?

2 Constructing kernels

Problem 3. One of the nice things about kernels is that new kernels can be constructed out of already given ones. Assume that $k_1(\boldsymbol{x}, \boldsymbol{x}')$ and $k_2(\boldsymbol{x}, \boldsymbol{x}')$ are valid kernels (i.e. they correspond to inner products of some feature vectors). Show that

- $k(\boldsymbol{x}, \boldsymbol{x}') = ck_1(\boldsymbol{x}, \boldsymbol{x}')$ for any $c > 0$,
- $k(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')$ for any real valued function $f(\boldsymbol{x})$,

- $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') + k_2(\boldsymbol{x}, \boldsymbol{x}')$,
- $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}')k_2(\boldsymbol{x}, \boldsymbol{x}')$

are all valid kernels.

3 Perceptron kernel

Problem 4. In this exercise, we develop a dual formulation of the perceptron algorithm. Using the perceptron learning rule you learned in the lecture, show that the learned weight vector can be written as a linear combination of the vectors $t_n\boldsymbol{\phi}(\boldsymbol{x}_n)$ where $t_n \in \{-1, +1\}$. Denote the coefficients of this linear combination by α_n and derive a formulation of the perceptron learning algorithm, and the predictive function for the perceptron in terms of the α_n. Show that the feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ enters only in the form of the kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x})$.

4 Support Vector Machines

You are given a data set with data from a single feature x_1 in $\mathbb{R}^1$ and corresponding labels $y \in \{+1, -1\}$. Data points for $+1$ are at $-3, -2, 3$ and data points for -1 are at $-1, 0, 1$.

Problem 5. Can this data set in its current feature space be separated using a linear separator? Why/why not?

Let's define a simple feature map $\phi(u) = (u, u^2)$ that transforms points in $\mathbb{R}^1$ to points in $\mathbb{R}^2$.

Problem 6. After applying ϕ to the data, can it now be separated using a linear separator? Why/why not (plotting the data may help you with your answer ...)?

Problem 7. Draw (approximately) a maximum-margin separating hyperplane (i.e. you do not need to solve a quadratic program). Clearly mark the support vectors. Also draw the resulting decision boundary in the original feature space. Is it possible to add another point to the training set in such a way, that the hyperplane *does not* change? Why/why not?