Assignment 2: Probability Theory and Support Vector Machines Machine Learning

Deadline: Friday 02 Nov 2018, 21:00

Introduction

In this assignment, you will further deepen your understanding of Support Vector Machines (SVMs). Please provide a latex based report in the PDF format.

Your report must be archived in a file named ”firstname.lastname” and uploaded to the iCorsi website before the deadline expires. Your report should also contain your name. Late submissions will be subject to a point penalty according to USI regulations.

Where to get help

We encourage you to use the tutorials to ask questions or to discuss exercises with other students. However, do not look at any report written by others or share your report with others. Violation of that rule will result in 0 points for all students involved.

Grading

The assignment consists of five parts totalling at 100 points. Bonus points are not necessary to achieve the maximal grade. You will be awarded 5 bonus points for a well-presented report (clear writing, annotated equations). Further bonus points are elaborated in the text below.

Exercise 1: Theoretical Probability (5 points)

Let X be a random variable with density function given by
$p(X = x) = [0.3e^{-x} + ke^{-2x}]1_{(0,+\infty)}(x)$,
where $1_{(0,+\infty)}(x)$ denotes the indicator function of the set $(0, +\infty)$:

1. (2 points) Compute the value of k such that it is a density
2. (3 points) Compute E[X]
   Hint: Remember the integration by parts formula:
   $\int_0^{+\infty} g(x)f'(x)\,dx = f(x)g(x)|_0^{+\infty} - \int_0^{+\infty} f(x)g'(x)\,dx$

Exercise 2: Theory of SVMs (50 points)

We consider an input vector $\mathbf{x} \in \mathbb{R}^N$ and a classifier $sign(f(\mathbf{x}))$ assigning a label of -1 or $+1$ to $\mathbf{x}$ based on the function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$, where $\mathbf{w}^T\phi(\mathbf{x})$ denotes the inner product of vectors $\mathbf{w}$ and $\phi(\mathbf{x})$ and $\phi(\mathbf{x})$ is a fixed feature-space transformation.

1. (10 points) Explain in your own words how an SVM roughly works. Make sure to add key characteristics, advantages, and disadvantages.
2. (10 points) Explain the objective functions of hard-margin and soft-margin support vector machine training as well as the constraints of the corresponding optimization problems.
3. (10 points) What is the Kernel trick? When is a kernel valid? Provide necessary and/or sufficient conditions.
4. (5 points) Consider the soft-margin optimization problem in point 2: write the equivalent dual problem in which the Kernel function appears.
5. (5 points) After solving the optimization problem, how would you classify a new point $\mathbf{x}$? Express the formula in terms of the kernel function.
6. (10 points) Explain which are the hyperparameters of an SVM and what are their effects

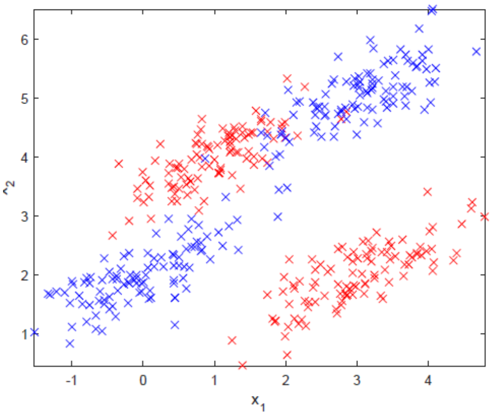
Exercise 3: Kernels I (15 points)

Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, d \in \mathbb{N}$. Explain in detail why each of the following functions is or is not a valid kernel:

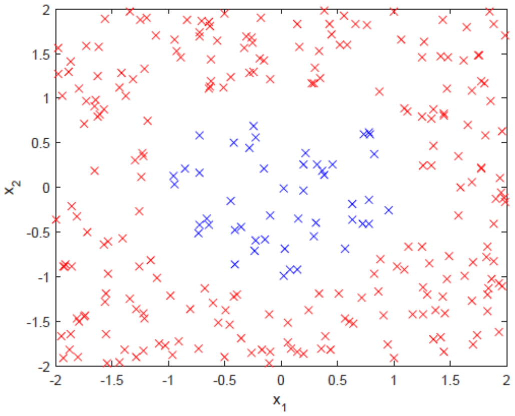
- $K(x, y) = x^T y + (x^T y)^2$
- $K(x, y) = x^2 e^{-y}$, d = 1
- $K(x, y) = ck_1(x, y) + k_2(x, y)$, where $k_1(x, y), k_2(x, y)$ are valid kernels in $\mathbb{R}^d$

Exercise 4: Kernels II (16 points)

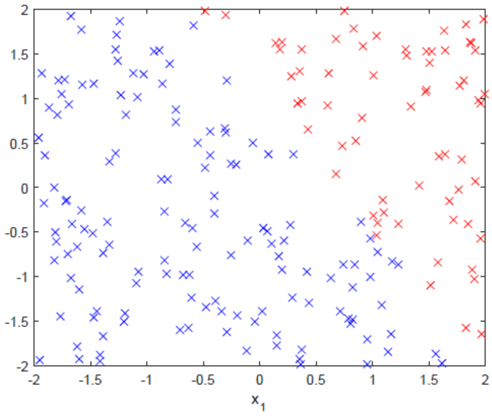
You are provided the following 4 dataset for solving a binary classification problem. You want to preprocess your data before applying a linear solver. Explain in each case if you would apply the kernel trick to represent your data. Which kind of kernel would you use? Why? If relying on kernels is not a good choice, which other transformation would you use?



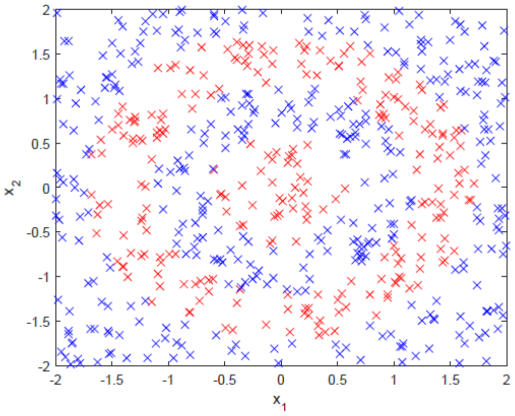
(a)



(b)



(c)



(d)

Exercise 5: SVMs (14 points)

Answer the following questions:

- (2 points) Suppose you have 2D examples: is the decision boundary of an SVM with linear kernel a straight line?
- (3 points) Suppose that the input data are linearly separable. Will an SVM with linear kernel return the same parameters w regardless of the chosen regularization value C?
- (3 points) Suppose you have 3D input examples ($\mathbf{x_i} \in \mathbf{R}^3$). What is the dimension of the decision boundary of the SVM with linear kernel?
- (3 points) Is the computational effort for solving a kernel SVM increasing as the dimension of the basis functions increases? Why?

- (3 points) Suppose that after our computer works for an hour to fit an SVM on a large data set, we notice that x_4, the feature vector for the fourth example, was recorded incorrectly, i.e., we use $\tilde{x}_4$ instead of x_4 to train our model. However, your coworker notices that the pair $(\tilde{x}_4,\ y_4)$ did not turn out to be a support point in the original fit. He says there is no need to train again the SVM on the corrected data set, because changing the value of a non-support point can't possibly change the fit. Is this true or false?