Lab 6: 23rd April 2012

Exercises on Support Vector Machines

1. What is the goal of the SVM algorithm? When can be it successfully applied?

Solution
SVMs are linear classifiers that find a hyperplane to separate two classes of data, positive and negative. They are successfully applicable when the two classes of data in the training set are linearly separable. They are suitable especially for high dimensional data. The attributes of the training data need to be real numbers.

2. What linear function is used by a SVM for classification? How is an input vector $\mathbf{x_i}$ (instance) assigned to the positive or negative class?

Solution
SVMs find a linear function of the form $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$ where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is a vector of weights, and b is the *bias*. An input vector $\mathbf{x_i}$ is assigned to the positive (1) or negative class (-1) as follows

$$y_i = \begin{cases} 1 & if \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 0 \\ -1 & if \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b < 0 \end{cases}$$

3. If the training examples are linearly separable, how many decision boundaries can separate positive from negative data points? Which decision boundary does the SVM algorithm calculate? Why?

Solution
There are infinite decision boundaries that separate positive from negative data points. The SVM algorithm found the boundary (hyperplan) with maximum margin. This boundary minimizes the upper bound of the classification error.

4. What is the margin? Which are the equations of the two margin hyperplans H_+ and H_-?

Solution
The margin is the distance between the two margin hyperplanes H_+ and H_-.
Their equations are:

H_+ : $<\mathbf{w} \bullet \mathbf{x}^+>+b=1$ H_- : $<\mathbf{w} \bullet \mathbf{x}^->+b=-1$

where $\mathbf{x}^+$ and $\mathbf{x}^-$ are the data points that are closest to the hyperplane $<\mathbf{w} \cdot \mathbf{x}> + b = 0$

5. Derive the formula that calculates the margin given a set of linearly separable training examples.

Solution
See slides titled “Compute the margin” at pages 4-5 of lecture 8.

6. Illustrate the constrained minimization problem that defines the SVM learning given a set of linearly separable training examples. What is the outcome of solving the problem?

Solution
The constrained minimization problem is illustrated in the slides titled "A optimization problem!" at page 5 of lecture 8. By solving the problem the SVM algorithm founds the decision boundary with maximum margin.

7. Explain why the constrained minimization problem is transformed into the Wolfe dual maximization problem. Which input vectors (training instances) are used to calculate the solution?

Solution
The transformation is performed to easy the finding of the optimal solution (i.e., the decision boundary with maximum margin). Only the support vectors (those data points on the margin hyperplanes H_+ and H_-) are used to calculate the solution. This considerably reduces the complexity of the SVM algorithm.

8. Is the linear separable SVM always applicable in case of noise in the training data? If no, how can be SVM be modified in order to deal with noisy training data?

Solution
With noisy data, the constraints of the optimization problem solved by the SVM algorithm may not be satisfied. Therefore, finding optimal decision boundary might not be possible. Noisy data is taken into consideration in SVM by relaxing the constraints and introducing *slack variables* in them. Details in slides from page 9 to page 14 of lecture 8.

9. For many real life data sets the decision boundaries are not linear. How is this non-linearity dealt with by SVMs?

Solution
The idea is to transform the non linearly separable input data into another (usually higher dimensional) space. A linear decision boundary can separate positive and negative examples in the transformed space. The transformed space is called the feature space. The original data space is called the input space. Details in slides from page 14 to page 15 of lecture 8.

10. What is a kernel function? What is the kernel trick?

Solution
A kernel function K is of the form $K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}) \bullet \phi(\mathbf{z}) \rangle$ where $\mathbf{x}$ and $\mathbf{z}$ are two input vectors, ϕ is the transformation function from the input space to the features space, and $\bullet$ denotes the dot product.
By kernel trick we mean that if we apply a kernel function to 2 input vectors we obtain the dot product of their transformation. The important thing is that we don't need to know the transformation function and the feature vectors obtained by the transformation in order to calculate the dot product. The kernel trick allows the SVM algorithm to be computationally tractable. Details in slides from page 17 to page 18 of lecture 8.

11. Make an example of a commonly used kernel.

Solution
A commonly used kernel is the polynomial kernel: $K(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x} \bullet \mathbf{z} + \theta \rangle^d$ where θ is a real number and d is a natural number.

12. Summarize the main advantages and limitations of SVM.

Solution

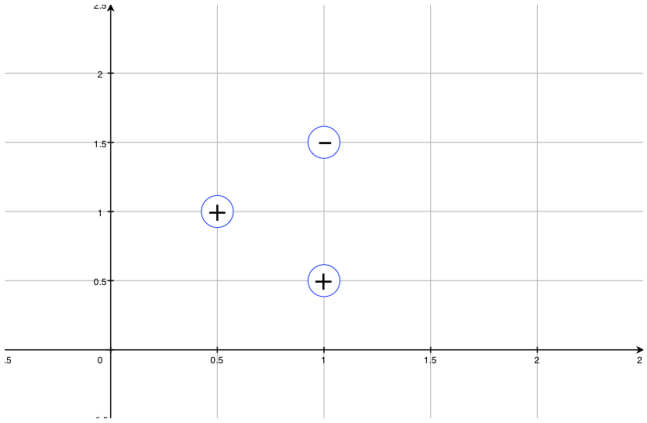
Main advantages of SVMs:

- Has rigorous theoretical foundation
- Performs classification more accurately than most other methods in applications, especially for high dimensional data
- They can be applied also to non linear classification problems by using kernel functions. The "kernel trick" allows that the these problems are computationally tractable
- Different kernels can be plugged into the same learning machinery and studied independently of it.

Main limitations of SVM:

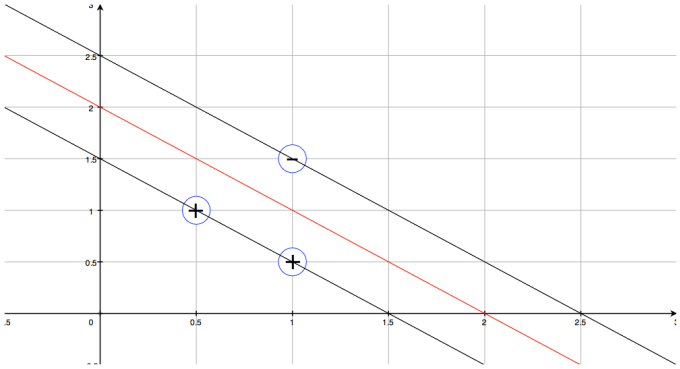
- Works only in a real-valued space
  - For a categorical attribute, we need to convert its categorical values to numeric values
- Does only two-class classification
  - For multi-class problems, some strategies can be applied, e.g., one-against-rest
- The hyperplane produced by SVM is hard to understand by human users. The matter is made worse by kernels
  - SVM is commonly used in applications that do not required human understanding

13. Consider the three linearly separable two-dimensional input vectors in the following figure. Find the linear SVM that optimally separates the classes by maximizing the margin.



Solution

All three data points are support vectors. The margin hyperplan H_+ is the line passing through the two positive points. The margin hyperplan H_- is the line passing through the negative point that is parallel to H_+. The decision boundary is the red line "half way" between H_+ and H_-. The equation of the decision boundary is –x+2=0. The following picture illustrates the solution:



14. Show for the polynomial kernel function

$$K(\mathbf{x},\mathbf{z})=(<\mathbf{x}\bullet\mathbf{z}> + \vartheta)^d,\ d=2,\ \vartheta=1,\ \mathbf{x}=(x_1,x_2),\ \mathbf{z}=(z_1,z_2)$$

that

$$K(x,z)=<\Phi(\mathbf{x})\bullet\ \Phi(\mathbf{z})> \text{ for } \Phi(\mathbf{y})=(1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1y_2)$$

Solution

$K(\mathbf{x},\mathbf{z})=(<\mathbf{x}\bullet\mathbf{z}> + 1)^2=(x_1z_1+x_2z_2+1)^2=x_1^2z_1^2+x_2^2z_2^2+1+2x_1z_1x_2z_2+2x_1z_1+2x_2z_2$

$<\Phi(\mathbf{x})\bullet\ \Phi(\mathbf{z})>=<(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)\bullet(1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2)>=$
$= 1+\sqrt{2}x_1\sqrt{2}z_1+\sqrt{2}x_2\sqrt{2}z_2+ x_1^2 z_1^2+ x_2^2 z_2^2+\sqrt{2}x_1x_2\sqrt{2}z_1z_2=$
$=1+ 2x_1 z_1+2x_2z_2+ x_1^2 z_1^2+ x_2^2 z_2^2+ 2x_1 z_1x_2z_2$
QED