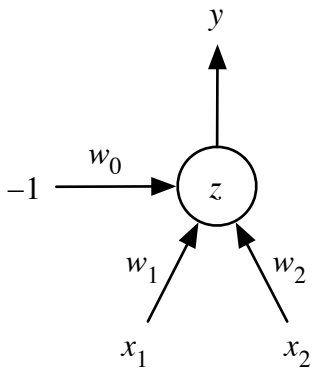
1 Classification

For this problem, we will consider the simple 1-layer neural network shown below, but rather than using a sigmoid function, as we've seen in class, it will use a Gaussian. That is, the output of the final unit, y, is computed as follows:

$$y = g(z)$$
$$g(z) = e^{-z^2}$$
$$z = -w_0 + w_1 x_1 + w_2 x_2$$

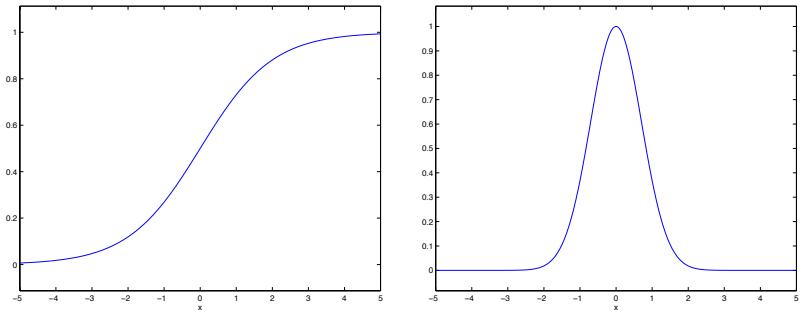


We can use this network to classify our inputs by assigning any input for which the output is greater than or equal to 1/2 to class 1 and for which the output is less than 1/2 to class 2. Using a weight vector of [0 1 1], compute the class for each of these input vectors:

input	class
[-1 -1]	$z = -2$, $y = e^{-4} = 0.02$, class 2
[-1 1]	$z = 0$, $y = e^{0} = 1$, class 1
[1 -1]	$z = 0$, $y = e^{0} = 1$, class 1
[1 1]	$z = 2$, $y = e^{-4} = 0.02$, class 2

2 Separator

1. Notice that in the last problem, we successfully created a separator for XOR, which we were unable to do with the sigmoidal transfer function. In this section we will see why this is the case. On the left, below, is a plot of the sigmoid as a function of its input. On the right, draw the corresponding plot for the Gaussian transfer function we used above.



2. Which of the 3D plots in Figure 1 corresponds to this Gaussian function as a function of a two-dimensional input?

The "Gaussian stripe" (c) is the right answer

3. So how can we describe the separator that we get with a simple 1-layer network using this transfer function? Can we describe it with (an) equation(s)?

   Solution:

   It is a Gaussian stripe over a line. The width of the stripe corresponding to a certain class is determined by the threshold (1/2 in our case). Thus, we can get equations for the two lines defining the region assigned to class 1 by setting $e^{-z^2} = 1/2$. Taking the log of both sides and then finding the square root, we get the following:

$$z = \pm\sqrt{\ln(2)} = \pm 0.83$$

$$w_0 + w_1 x_1 + w_2 x_2 = 0.83$$

$$w_0 + w_1 x_1 + w_2 x_2 = -0.83$$

3 Different Separators

Figure 2 shows three separators for the same set of data points. Each attempted separator uses a different kernel function.

(a) What function is being plotted in Figure 2? Where is it positive? Negative? What is the decision boundary in each diagram? Where are the support vectors?

Solution:

The figure maps the value of the signed distance function of SVM, that is, $h'(u) = \sum_i \alpha_i y^i K(x^i, u) + b$ for each u to a color. The support vectors are circles on the gray region, indicating $|h'(u)| = 1$; this is true since we required the margin of the closest points to the separator to be 1. The decision boundary is a line passing through the white region with equal distance from the locus of $|h'(u)| = 1$. The region to the positive support vectors' side of the decision boundary (blue side) correspond to positive values, and the other side (red side) correspond to negative values.

(b) Fill in the table below, indicating which kernel functions might have been used:

Solution:

Kernel	Attempted Separator
Linear	**B**
Polynomial (n=2)	**C**
Radial Basis Function	**A**

4 Post-Training Calculations

Assume that we are using an SVM with a **polynomial kernel of degree 2**, i.e. $k(x^i, x^j) = (1 + x^i \cdot x^j)^2$. Say that training produces the following support vectors for each of which the α value is equal to 0.05.

x_1	x_2	y
-1	2	+1
1	2	-1

(a) What is the value of b?

Solution:

Let Φ be a transformation to a feature space whose kernel is polynomial with degree 2. Given the alphas, we can use the optimality condition for w and the equality constraint on each support vectors x^j,

$$w^* = \sum_i \alpha_i y^i \Phi(x^i) \qquad\qquad y^j(w^* \cdot x^j + b) - 1 = 0$$

to compute the offset b:

$$b \;=\; 1/y^j - w^* \cdot \Phi(x^j) = 1/y^j - \sum_i \alpha_i y^i \Phi(x^i) \cdot \Phi(x^j) = 1/y^j - \sum_i \alpha_i y^i K(x^i, x^j)$$

Without loss of generality, we'll use the first support vector $j = 1$ so $(y^j, x_1^j, x_2^j) = (+1, -1, 2)$.

$$b \;=\; 1 - 0.05(1)K(x^1, x^1) - 0.05(-1)K(x^2, x^1) = 1 - 0.05(36 - 16) = 0$$

(b) What value does this SVM compute for the input point $(1, 3)$?

$$\begin{aligned} h'((1,3)) &= \sum_i \alpha_i y^i K(x^i, (1,3)) = 0.05(1 + (-1,2) \cdot (1,3))^2 - 0.05(1 + (1,2) \cdot (1,3))^2 \\ &= 0.05(36 - 64) = -1.4 \end{aligned}$$

5 Hand Training

What are the values for the α_i and the offset b that would give the maximal margin linear classifier for the two data points shown below. You should be able to find the answer without deriving it from the dual Lagrangian.

i	x^i	y^i
1	0	+1
2	4	-1

Solution:
We know that $w = \sum_i \alpha_i y^i x^i$. Thus:

$$\begin{aligned} w &= \alpha_1 y^1 x^1 + \alpha_2 y^2 x^2 \\ &= \alpha_1(1)(0) + \alpha_2(-1)(4) \\ &= -4\alpha_2 \end{aligned}$$

We know further that $\sum_i \alpha_i y^i = 0$, so that the alphas must be equal. Lastly, we know that the margin for the support vectors is 1, so $w \cdot x_1 + b = 1$, which tells us that $b = 1$, and $w \cdot x_2 + b = -1$, which tells us that $w = -0.5$. Thus we know that $\alpha_1 = \alpha_2 = 1/8$.

6 RBF Complexity

Consider the one-dimensional classification problem defined by the following data points. Imagine attacking this problem with an SVM using a radial basis function kernel.

i	x^i	y^i
1	1	+1
2	2	-1
3	3	-1
4	4	+1
5	5	+1

(a) Which of the above data points will be support vectors?

Solution:

$i = 1, 2, 3, 4$

(b) Assume that we want the classifier to return a positive output for the +1 points and a negative output for the -1 points. Draw a plausible classifier output for every feature value in the interval $[0, 6]$. Do this twice. Once, assuming that the standard deviation σ is very small relative to the distance between adjacent training points. And again, assuming that the standard deviation σ is approximately the distance between adjacent training points. Note that a Gaussian kernel is close to zero for values farther than three standard deviations from its center.

Solution: Get the diagrams from the past quizzes

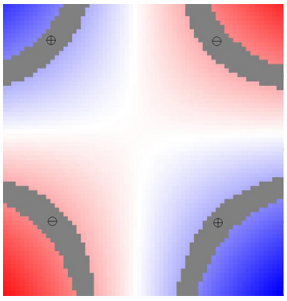
(c) Can you conclude anything from these sketches about the complexity of the RBF kernel as a function of its standard deviation σ?

Solution: Complexity increases as σ decreases.

7 XOR Using an RBF

Given the following four data points, we find a separator using a radial basis function (RBF) with $\sigma = 1$ that produces the following plot:

i	x^i	y^i
1	[-1 -1]	-1
2	[-1 +1]	+1
3	[+1 -1]	+1
4	[+1 +1]	-1



(a) Recall that the equation that determines the distance from the separator is:

$$h'(u) = \sum_{i=1}^{k} \alpha_i y^i K(x^i, u) + b$$

$$K(x^i, u) = e^{-(||x^i - u||)^2/(2\sigma^2)}$$

t	e^t
-1	0.3679
-5	0.0067
-9	0.0001

The α values are all 1.33 in this example and the offset is 0. Assume you are given a **test** point [+2 +2]. What is the distance from the separator (use the e^t values from the table above)?

$$\begin{aligned} h'(u) &= 1.33(-K(x^1, u) + K(x^2, u) + K(x^3, u) - K(x^4, u)) \\ &= 1.33(-e^{-9} + e^{-5} + e^{-5} - e^{-1}) \\ &= -0.4716 \end{aligned}$$

(b) Consider what we needed to do to compute this separator. We first needed to optimize the α values. To do this, we needed to compute $K(x^i, x^j), \forall i, j$. Compute the kernel outputs using the same radial basis function ($\sigma = 1$) for the points below.

K(i, j)	kernel value
K(1, 1)	e^0
K(1, 2)	e^{-2}
K(1, 3)	e^{-2}
K(1, 4)	e^{-4}

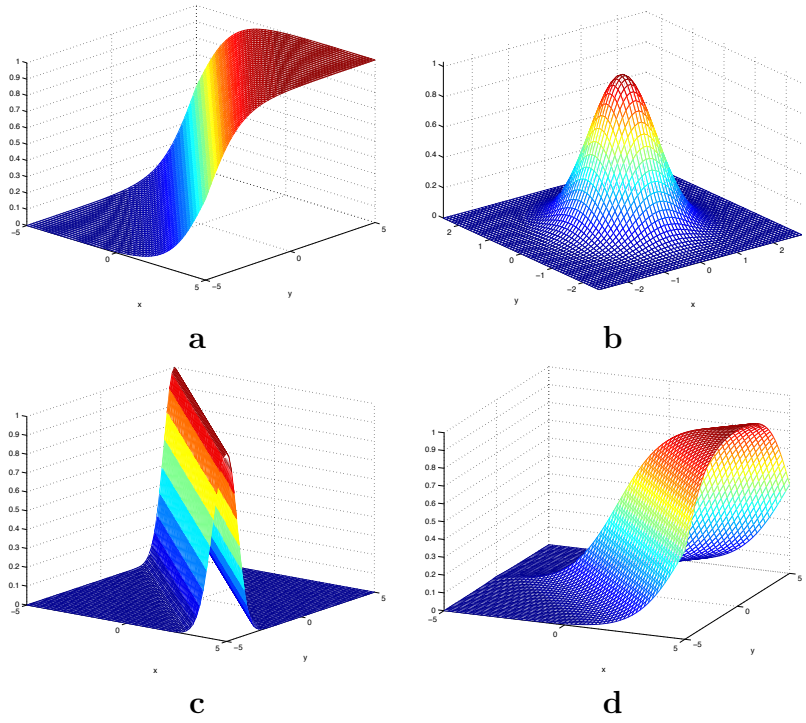


Figure 1: Problem 2 (Separator)

(c) Sketch a picture of how the plot would change if we had been given a fifth **training** point $(y^5, x^5) = (-1, [+2\ +2])$? What if the fifth training point had been $(y^5, x^5) = (-1, [+\frac{1}{2}\ +\frac{1}{2}])$ instead?

Solution:

In the first case, the the plot will not change at all, since the point $(y^5, x^5) = (-1, [+2\ +2])$ will not be a support vector.

In the second case, the decision boundary must change, since our original $h'(u)$ (plotted above) will assign the new point $(y^5, x^5) = (-1, [+\frac{1}{2}\ +\frac{1}{2}])$ a negative value greater than -1. It's clear that the new data point will be a support vector, since it is closer to both the +1 data points. Therefore, the locus of margin equal to 1 on the $x_1 > 0, x_2 > 0$ plane will move closer to the origin and so will other loci with margin = 1 and the decision boundary change accordingly.

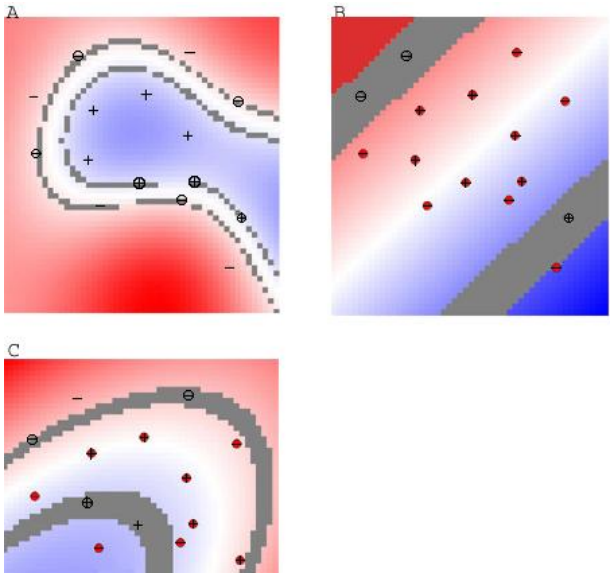


Figure 2: Problem 3 (Different Separators)