Assignment 2: SVM

Luca Brena

October 2019

1 Exercise 1: Kernels I

For solving theese questions I'm using the rules defined during the tutorial about constructing a new valid kernel from two given valid kernels.

- Valid kernel. The first term of the addition is a valid kernel because it's a linear kernel (we know that the linear kernel is a valid kernel). After this, the second term of the addition is a valid kernel because it's is a product between 2 valid kernels (in particular two linear kernels). It's easy to see the previous statement if we decompose the term in this way:

  $$(x^T y)^2 = (x^T y)(x^T y)$$

  Finally, we have the sum of two valid kernels which is a valid kernel.

- This is not a valid kernel. The reason is that a necessary condition for a kernel to be valid is simmetry. The function must be symmetric. In this case the kernel function isn't symmetric, because $K(x, y) \neq K(y, x)$. It's easy to prove this with a simple example (we don't have problems about dimension because we are working in a monodimensional space).

  $$K(1, -1) = 1 * e^1 = e$$

  $$K(-1, 1) = 1 * e^{-1} = \frac{1}{e}$$

- Valid kernel only if $c > 0$. We have a sum in which the second term is a valid kernel. Now we have to check if the first term is a valid kernel. We have a valid kernel times a costant. So if the constant is less then or equal to zero, this product is not a valid kernel. In the case the costant is greater than 0, the product is a valid kernel and so the sum of two valid kernels is a valid kernel.

2 Exercise 2: Kernels II

- This data are very noisy. A possible approach is a Gaussian kernel which maps in infinit dimension. There it's possible to find a hyperplane but with possible bad generalization of the model caused by overfitting.

- In this case we can follow a kernel approach or not. If we want to use a kernel we could use a 2 dregree polinomial kernel. If we don't want use a kernel we can use a simple mapping like $\phi(x) : [x_1, x_2, ||x||^2]$ and use a linear classifier. This second approach works because the blue data are centered in the origin of the two axes.
- We use a linear kernel because the 2 classes ar linearly separable.
- For this kind of distribution we can use a high degree polinomial kernel (like degree 6), or a Gaussian kernel that works well on this concentric data.

3 Exercise 3: SVMs

- Yes because we are using a linear kernel in a 2 dimensional space so the feature space has 2 dimensions too (in this case we implicitly use a identity mapping). So the hyperplane in the feature space is a straight line. In this case the original space and the feature space have the same dimensions so if we project the feature space hyperplane that is a line in the original space we obtain a line.
- No. The parameter C is used to balance the contribution of the misclassified points and the point within the pargin (more precisely, the distance between the points and the right margin boundary) in the margin width maximization problem. A high C means thin margin and few missclassification; diminuishing C we allow to more missclassifications and maybe we can find a better way to separate the points, leading to a weights change.
- It's a two dimensonal plane. This follows from the observation that if we consider 1-dimensional inputs the boundary is a point that has no dimension; if we consider 2-dimension inputs the hyperplane has one dimension, in other words a line.
- No because when we use the kernel trick we do not apply directly the basis function, but we operate in the original space, so the computational effort doesn't depend on the dimension of the feature space but on the dimension of the original space.
- The maximim value of the Gaussian Kernel is 1.

  $$K(x, x') = e^{-\frac{||x-x'||^2}{\sigma^2}}$$

  We can rewrite as follow because the exponent is a negative quantity

  $$K(x, x') = \frac{1}{e^{\frac{||x-x'||^2}{\sigma^2}}}$$

  We are maximizing, in order to do so we have to minimize the denominator. More precisely, the exponent of e is a positive quantity so the range of

the denominator goes from 1 to plus infinite. So we in order to maximize the fraction we want the denominator to be as small as possible. This is shown by this limit, considering fixed the variance.

$$\lim_{x' \to x} e^{\frac{||x-x'||^2}{\sigma^2}} = 1$$

So we can say

$$K(x, x') = \frac{1}{e^0} = 1$$

- In case we want just to find a hyperplane that separates the data and we don't care about the performance of the model, the statement is true. We can map our data in a very high dimensional space and find a hyperplane that separates them, but in this way the model is gonna perform pretty bad on new data because of the overfitting. In the case of noisy data we can use soft margin to allow some miscalssification but getting a model able to better generalize.

4 Exercise 4: SVMs

- x_1 is classified as positive following this procedure:

$$sgn(x_1^T w + b) = sgn(\begin{bmatrix} 3 \\ -1 \end{bmatrix} \begin{bmatrix} 3 & 2 \end{bmatrix} + 2) = sgn(9) = 1$$

- We notice that:

$$(x_2^T w + b) = 0$$

$$(\begin{bmatrix} 3 \\ -1 \end{bmatrix} \begin{bmatrix} 3 & 2 \end{bmatrix} + 2) = 0$$

$$0 = 0$$

This means that x_2 is on the separate boundary. In this case we have to train again the model.

- No, because we don't always need to train the model again. If the point is well classified (even if it sits exacly on the right margin) we don't need to retrain because in the optimization problem this sample is not considered. In case of missclasssifications and for points within the margin we need to train again the model.