Name: Salim Fathy Selim Sherif
Intake2 AI Mansoura

DISCUSS THE ORIGINS OF KERNELS IN MACHINE LEARNING

1. Who proposed the idea with SVM? History of SVM?

- Started develop at AT&T Bell Labs by Vladimir Vapnik with colleagues.
  SVM proposed by Vladimir Vapnik and Alexey Chervonenkis.
  SVC proposed by Bernhard Boser, Isabelle Guyon, and Vapnik.
  SVR proposed by Chris Burges, Linda Kaufman, and Alex Smola.

- **History and Background:**
  Vladimir Vapnik and Alexey Chervonenkis developed the Vapnik-Chervonenkis theory (also known as VC theory) during 1960–1990. VC theory is a computational learning theory related to the statistical learning theory from distribution free data. This was the basis and starting point for the support vector machine (SVM), a supervised learning technique for classification and regression. Although Vapnik (Vapnik 1995, 1998) introduced the subjects of linear classifiers and optimal separating hyperplanes in the 1960s, it did not receive much attention until the nineties.

  Bernhard Boser, Isabelle Guyon, and Vapnik introduced a way to create nonlinear classifiers by applying the kernel trick to maximum margin hyperplanes. The kernel trick is a method for converting a linear classifier algorithm into a nonlinear one by using a nonlinear function to map the original observations into a higher-dimensional space, originally proposed by Aizerman. SVMs have been successfully applied in various classification and machine learning applications.

  Chris Burges, Linda Kaufman, and Alex Smola proposed a version of SVM for utilization in the regression tasks (Drucker et al. 1996). This is popularly known as support vector regression (SVR). SVM is an expanding field in the machine learning field. SVM incorporates different ideas like the VC-theory, statistical learning, maximum margin optimal hyperplane, kernel tricks, and so on.

Name: Salim Fathy Selim Sherif
Intake2 AI Mansoura

2. What is the motivation?

The goal of SVM, like any classification algorithm, is to find the best decision boundary that splits the data into two classes. So can many hyperplanes might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So, we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum-margin classifier or equivalently, the perceptron of optimal stability.

3. The usage of kernels in SVM?

A kernel is a function used in SVM for helping to solve problems whether it is linear, not, or something else. They provide shortcuts to avoid complex calculations. The amazing thing about the kernel is that we can go to higher dimensions and perform smooth calculations with the help of it. We can go up to an infinite number of dimensions using kernels. Sometimes, we cannot have a hyperplane for certain problems. This problem arises when we go up to higher dimensions and try to form a hyperplane.

4. Extending SVC for multi-class classification?

The idea stands to map data points to high dimensional space and used to separate a multiclass classification problem into multiple binary classification datasets and train a binary classification model for each. called the 'One-to-One (OVO)' technique, which splits the multiclass problem into multiple binary classification problems. A binary classifier per each pair of classes. Another technique one can use is 'One-to-Rest (OVA)'. In that technique, the breakdown is set to a binary classifier per class.

Name: Salim Fathy Selim Sherif
Intake2 AI Mansoura

5. What rules that should be fulfilled to implement a kernel function?

These rules are the deciding factors of what kernel should be implemented for classification. One such rule is the **moving window classifier** or we can also call it the **window function**.

This function is shown as:

- **fn(x) = 1, if ∑ ? (|| x – xi|| <= h) ? (yi=1) > ∑ ? (|| x – xi|| <= h)? (yi=0)**
- **fn(x) = 0**, otherwise.

Here, the summations are from i=1 to n. 'h' is the width of the window. This rule assigns weights to the points at a fixed distance from 'x'.
'xi' are the points nearby 'x'.

It is essential that the weights should be distributed in the direction of xi. This ensures the smooth working of the weight functions. These weight functions are the kernel functions. The kernel function is represented as **K: Rd-> R**.
The functions are generally positive and monotonically decreasing. There are some forms of these functions:

- Uniform kernels: **K(x) = ?(||x||<=1)**
- Gaussian kernels: **K(x) = exp(-||x||2)**
- Epanechnikov kernels: **?(1-||x||2)?(||x||<=1)**

6. Comparison between kernel types:

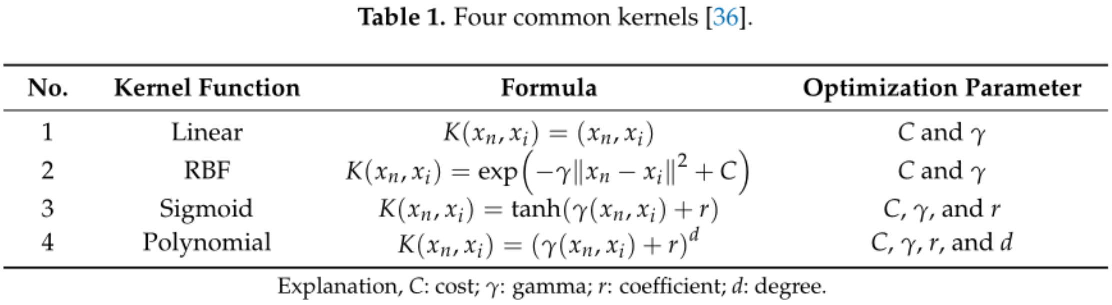

Table 1. Four common kernels [36].

No.	Kernel Function	Formula	Optimization Parameter
1	Linear	$K(x_n, x_i) = (x_n, x_i)$	C and γ
2	RBF	$K(x_n, x_i) = \exp\left(-\gamma\|x_n - x_i\|^2 + C\right)$	C and γ
3	Sigmoid	$K(x_n, x_i) = \tanh(\gamma(x_n, x_i) + r)$	C, γ, and r
4	Polynomial	$K(x_n, x_i) = (\gamma(x_n, x_i) + r)^d$	C, γ, r, and d

Explanation, C: cost; γ: gamma; r: coefficient; d: degree.

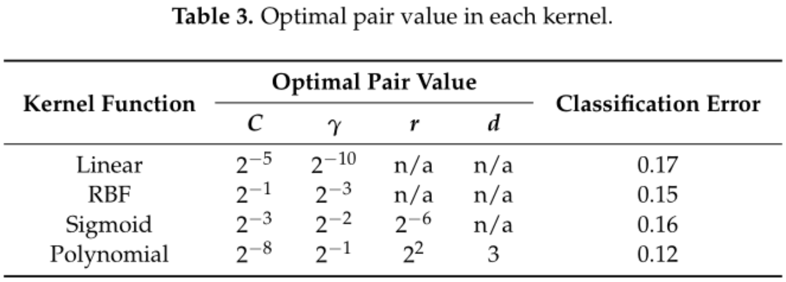

Table 3. Optimal pair value in each kernel.

Kernel Function	Optimal Pair Value				Classification Error
	C	γ	r	d	
Linear	2^{-5}	2^{-10}	n/a	n/a	0.17
RBF	2^{-1}	2^{-3}	n/a	n/a	0.15
Sigmoid	2^{-3}	2^{-2}	2^{-6}	n/a	0.16
Polynomial	2^{-8}	2^{-1}	2^{2}	3	0.12

The overall summary of the various kernels is listed in Table 3. It can be seen that the best hyperplane model is owned by the polynomial kernel function. The reason is that this kernel has the lowest classification error of its competitors (linear, RBF and sigmoid)

Name: Salim Fathy Selim Sherif
Intake2 AI Mansoura

7. How to choose correct Kernel for an ML problem?

Mostly we analyze the data and see if it is linear, we can use a linear kernel or RBF kernel and if it is non-linear, we can use a polynomial kernel with the change in any hyperparameter selection for SVMs done via cross-validation in combination with Grid Search.

8. The Concept of VC dimension

Vapnik–Chervonenkis (VC) dimension of SVM Statistical Learning of {f(α)} is the maximum number of training points that can be shattered by {f(α)} and not directly related to number of parameters. Vapnik (1995) has an example with 1 parameter and infinite VC dimension. VC dimension is a property of a set of functions {f(α)} and can be defined for various classes of function f.

VC dimension gives concreteness to the notion of the capacity of a given set of functions.

9. The Curse of Dimensionality?

Curse of Dimensionality refers to a set of issues that occur when working with high-dimensional data. The dimension of a dataset corresponds to the number of features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data.

For example, SVM are well known for their effectiveness in high dimensional spaces, where the number of features is greater than the number of observations. The model complexity is of O(m^2 samples * n-features) so it's perfect for working with data where the number of features is bigger than the number of samples. The SVMs create hyper-planes (could be more than one) of n-dimensional space and the goal is to separate the hyperplanes.

Some Models as KNN not working well with Curse of Dimensionality can use some techniques to help to best result Feature selection Techniques (Multicollinearity,...) and Feature Extraction Techniques (PCA,..)