

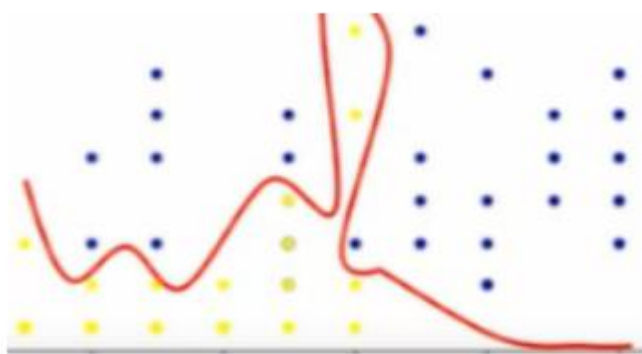
## **SVM (Support Vector Machines) / машины опорных векторов**

Support Vector Machine - это контролируемый алгоритм - обучение с учителем (Supervised learning), который может классифицировать случаи путем поиска разделителя.

SVM также называется большим классификатором поля. SVM функционирует как большой классификатор полей, который пытается отделить положительные и отрицательные примеры с максимально большим запасом.

SVM работает путем первого сопоставления данных с пространством пространственных объектов с высокой размерностью, так что точки данных могут быть классифицированы, даже если данные не могут быть линейно разделены. Затем вычисляется разделитель для данных.

Две категории могут быть разделены кривой, но не линией.



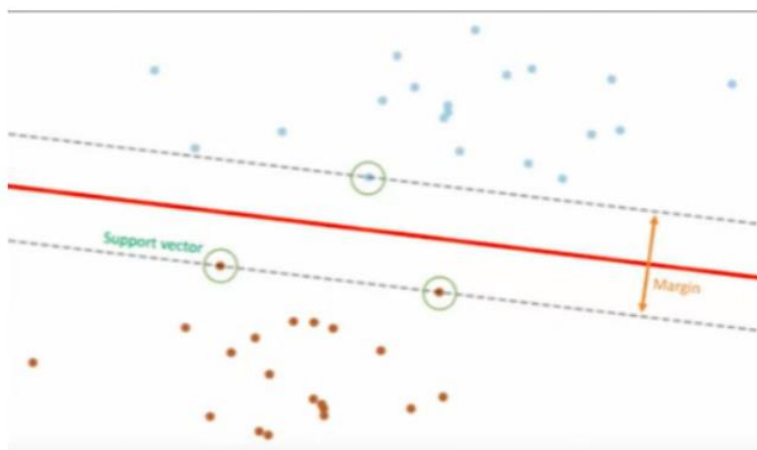
Чтобы добиться возможности разделить данные с помощью линейного объекта (линия, плоскость, гиперплоскость) необходимо преобразовать данные в более высокую размерность. Отображение данных в более высоком пространственном пространстве называется, кернелингом.

Математическая функция, используемая для преобразования, известна как функция ядра и может быть различных типов, таких как линейная, полиномиальная, радиальная базисная функция ( RBF) и сигмоид.

Каждая из этих функций имеет свои характеристики, свои плюсы и минусы, а также свое уравнение ( большинство из них уже реализованы в библиотеках языков программирования данных). Кроме того, поскольку нет простого способа узнать, какая функция лучше всего работает с любым набором данных, обычно выбирают разные функции и сравнивают результаты.

Один разумный выбор в качестве лучшей гиперплоскости является тот, который представляет наибольшее разделение или поле между двумя классами. Таким образом, цель состоит в том, чтобы выбрать гиперплоскость с как можно большим запасом. Данные, ближайшими к гиперплоскости, являются векторы поддержки.

Для достижения цели важны только векторы поддержки. И нужно найти гиперплоскость таким образом, чтобы она имела максимальное расстояние для поддержки векторов. Гиперплоскость извлекается из обучающих данных с помощью процедуры оптимизации, которая максимизирует маржу. Данную оптимизационную задачу можно решить с помощью градиентного спуска.



Используя [датасет](#) из Лабораторной работы №3 и результат лабораторной работы №6 (.csv файл с признаками и разметкой) выполнить следующие задания:

1. Разбить данные на train (60% от датасета), val(20%) и test(20%)
2. С помощью метода SVM построить классификатор, распознающий пол говорящего на аудиозаписи
3. Рассчитать accuracy обученной модели
4. Рассчитать precision и recall обученной модели
5. Рассчитать F1 score обученной модели

6. Найти аудиозаписи на которых модель ошибается; прослушайте их, визуализируйте и сравните признаки; сформулируйте предположение почему модель ошибается на этих аудиозаписях