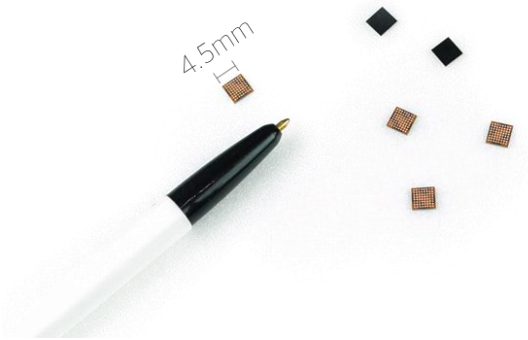
Как работают нейрочипы

Пересказ документации на микросхему CM1K от NeuroMem. Без нейронных сетей, но с распознаванием образов.

Предыстория

- 2015 год 100$ + 100$ пересылка у спец. фирмы



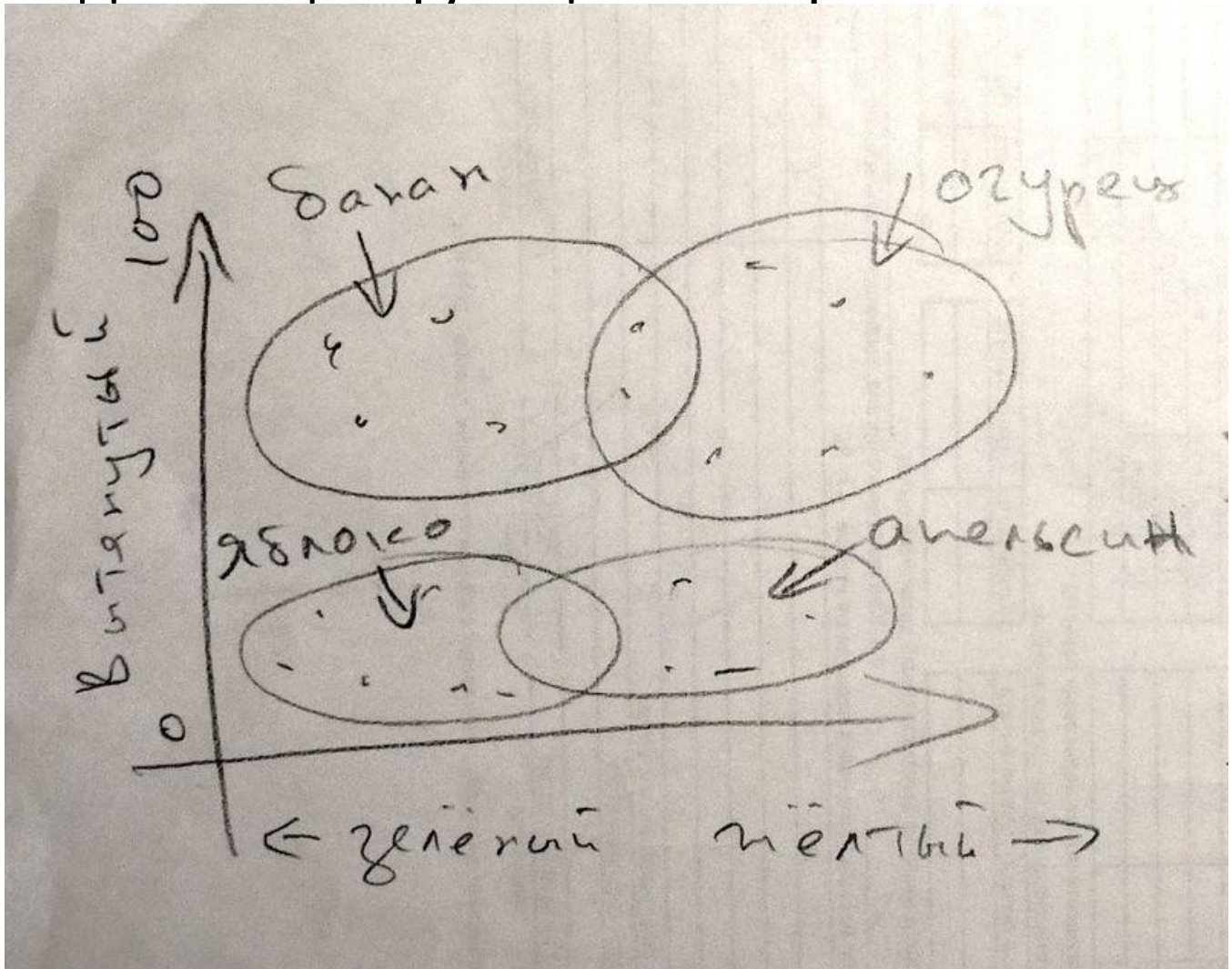
- 2020 год 50$ + 25$ пересылка с Amazon

Сайты производителей

- Чип NM500 – General Vision (США?) https://www.general-vision.com/hardware/nm500/
- Платы – nepes AI (Корея) http://www.theneuromorphic.com/
- Программный эмулятор https://github.com/kebwi/CM1K_emulator

Распознавание образов

- Распознавание образов = построение разделяющих функций-поверхностей

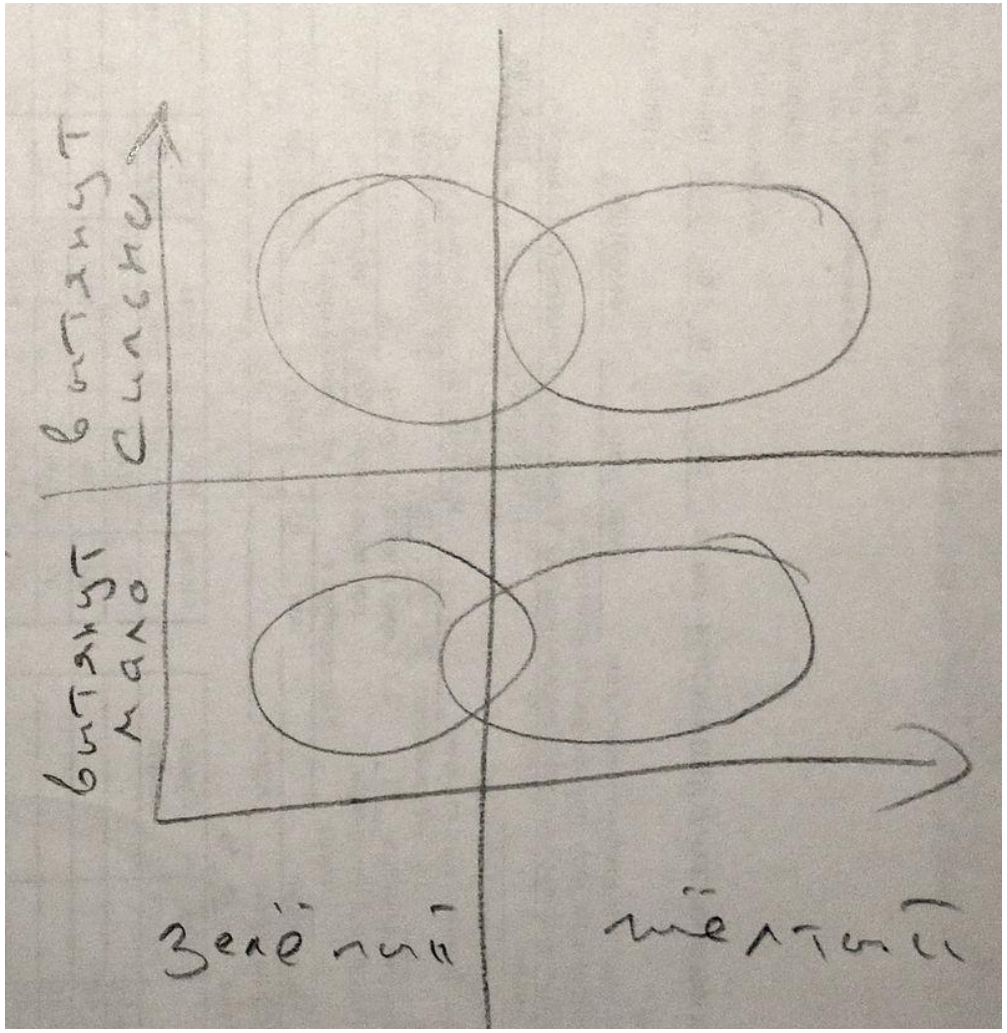


Обучение

- Обучение = приближение к разделяющей функции на основе неполных данных
- Если доступны все классифицируемые объекты, то можно построить идеальное разделение
- На практике приближение должно быть достаточно близким, чтобы правильно классифицировать новые объекты
- Увеличение размерности признаков увеличивает точность (?), но требует памяти

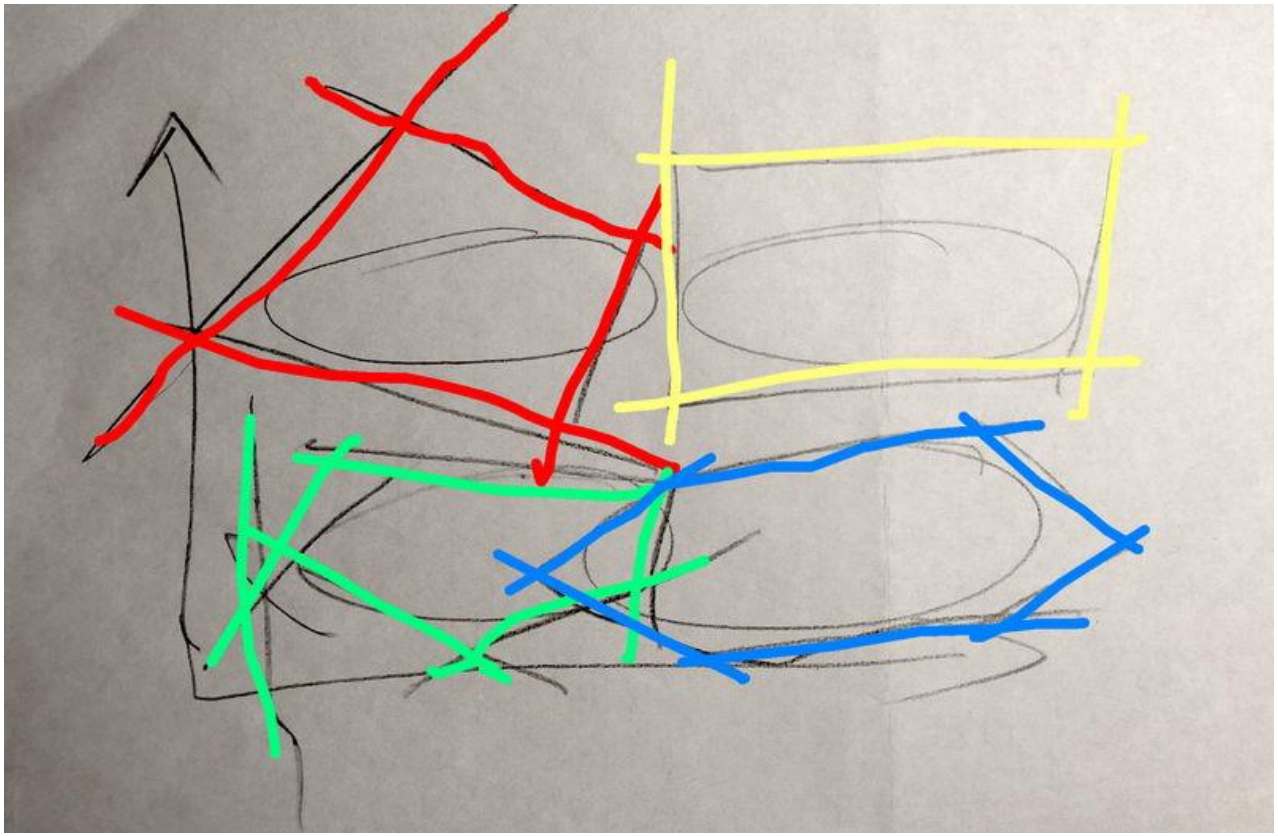
Разные разделяющие функции

- Дерево принятия решений – кусочнолинейное – полуплоскости



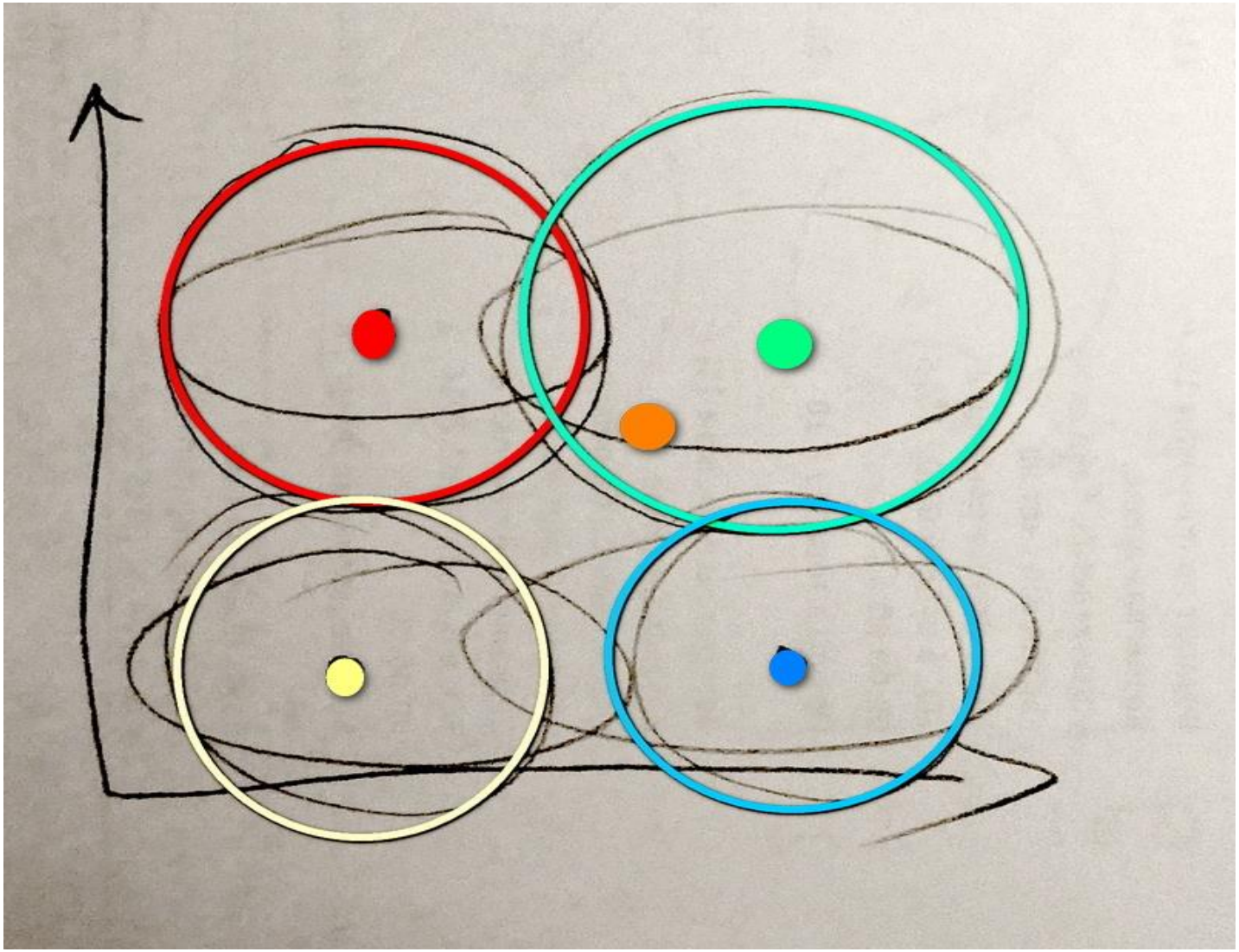
Разные разделяющие функции

- Линейное программирование – кусочнолинейное – системы неравенств



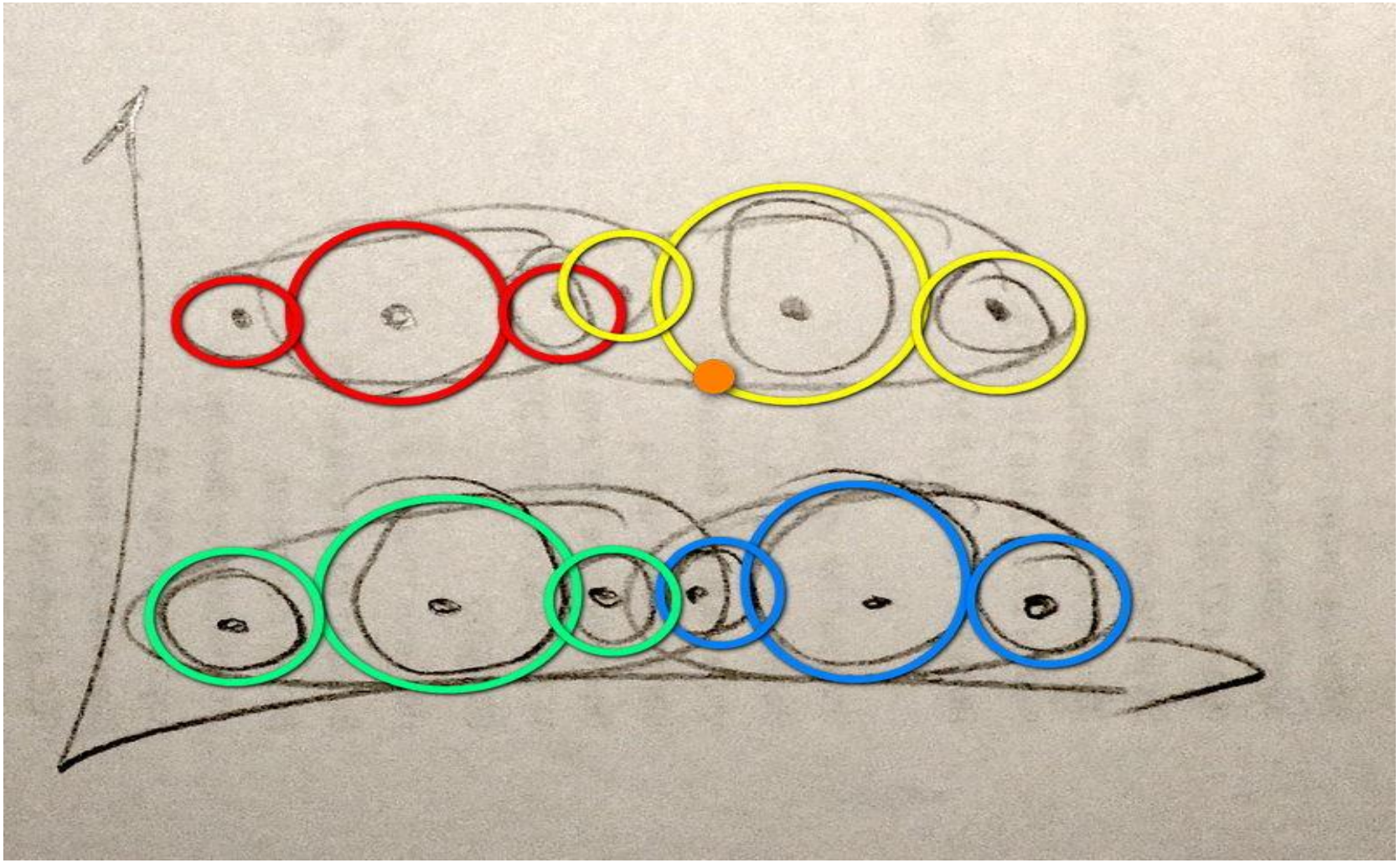
Радиальные функции (RBF)
Ограниченная кулоновская энергия (RCE)

- http://www.igce.comcor.ru/AI_mag/NN/RadNets/RadNets.html



RCE

- Если надо точнее



Обучение RCE

- Обучение включает два этапа:

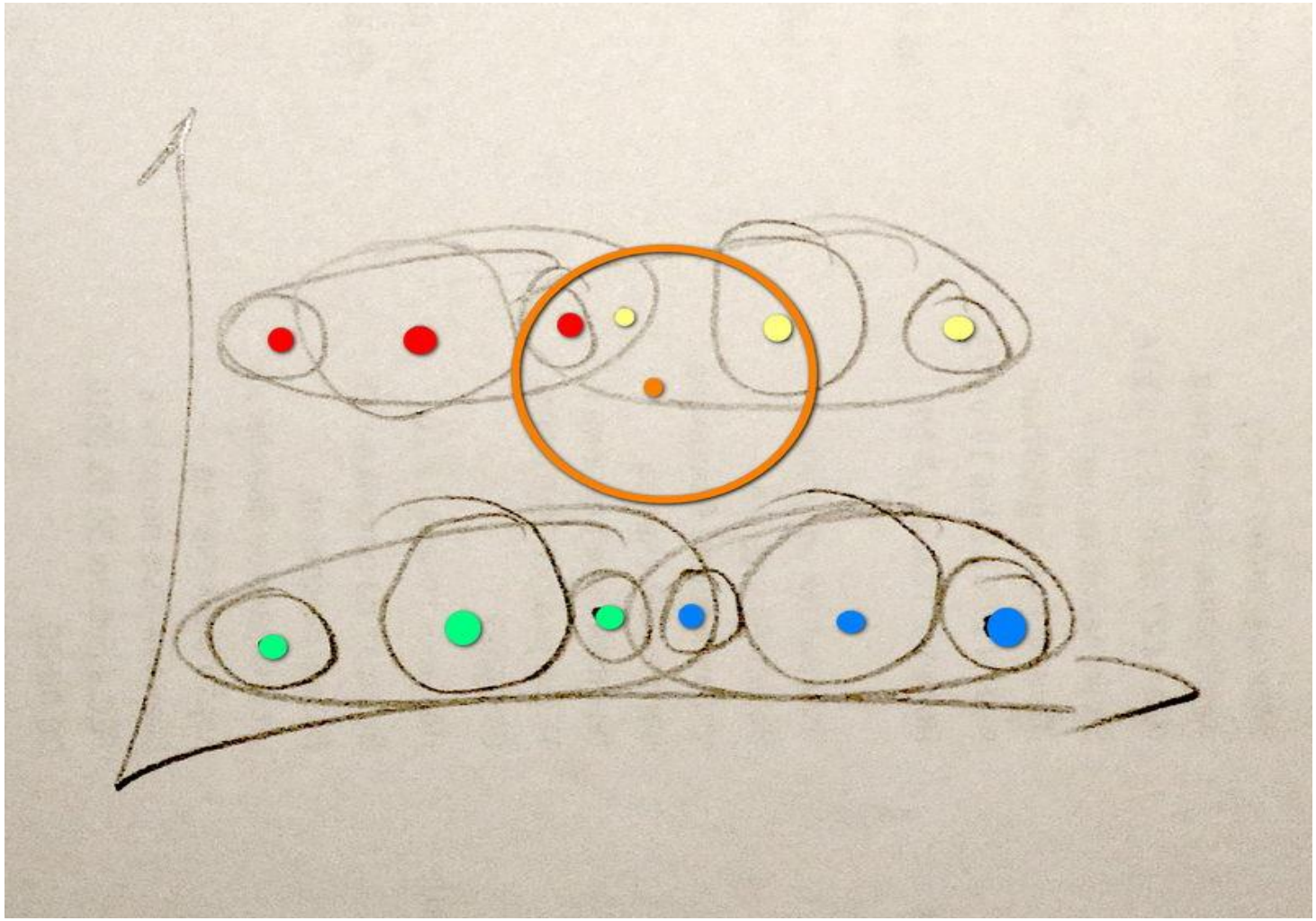
1. Выбор образцов
2. Назначение весов

- Требования к образцам

1. полнота покрытия области определения;
2. равномерность распределения.

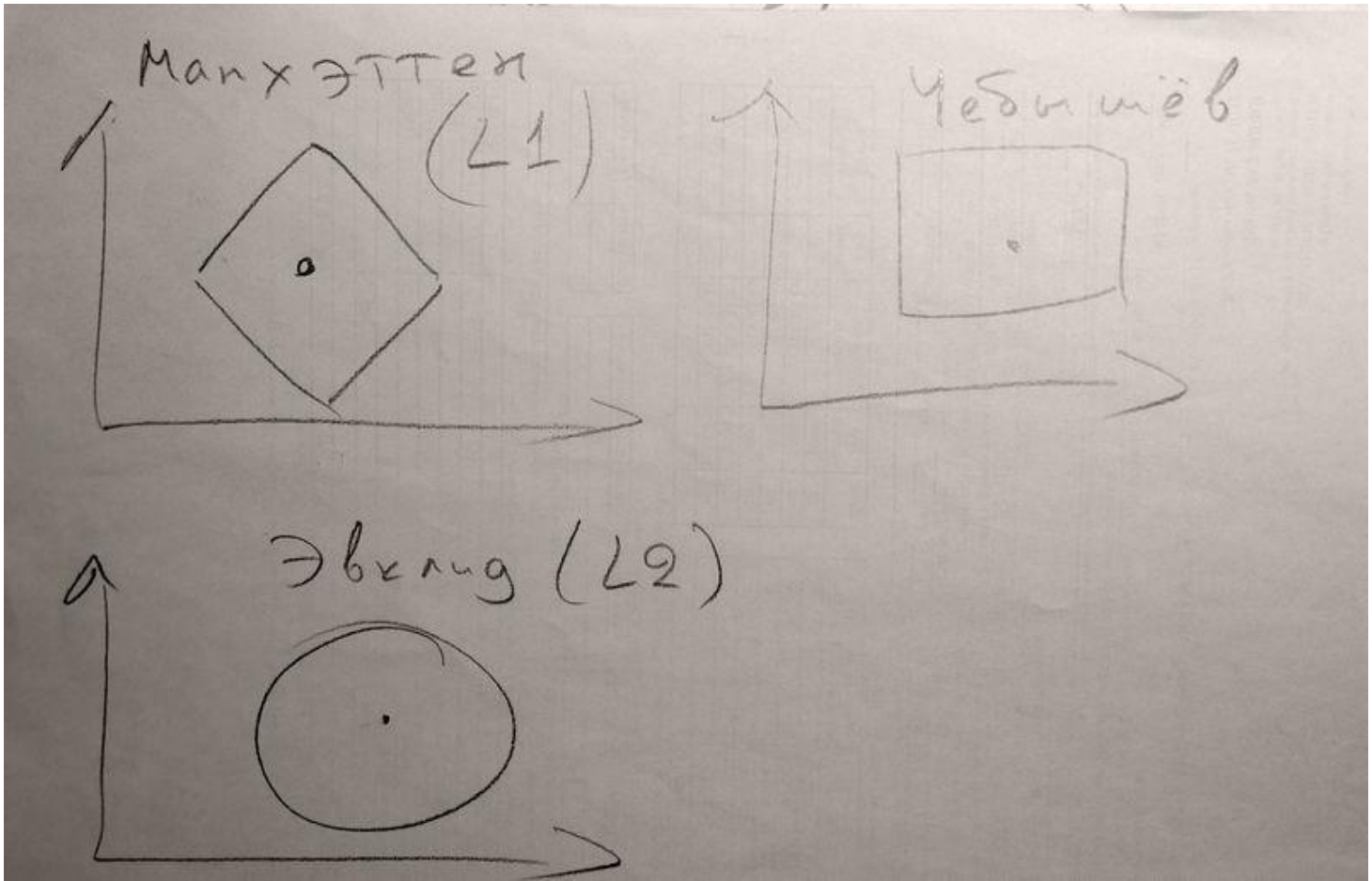
К ближайших соседей KNN

- Обратная сторона RBF. Нет весов. Есть расстояния.



Расстояния между объектами могут вычисляться различными способами:

- манхэттенское расстояние (L1) Dist(A,B)=|Ax - Bx| + |Ay - By|
  - квадрат стоит на угле
- расстояние Чебышёва Dist(A,B)=MAX(|Ax - Bx|, |Ay - By|)
  - квадрат лежит на стороне
- эвклидово расстояние (L2) Dist(A,B)=SQRT(Ax * Bx + Ay * By)
  - круг



Качественный результат классификации

- Неизвестно: объект не притягивается ни к одному из образцов
- Идентифицировано: объект притягивается к одному или нескольким образцам, которые относятся к одной категории
- Неопределенно: объект притягивается к нескольким образцам, которые относятся к разным категориям

Что делать при неопределённости

- засчитать категорию образца с лучшим соответствием (то есть ближайшего)
- выбрать N ближайших образцов и применить некоторые статистические или вероятностные правила для вывода единого глобального ответа
- отменить ответ и воспользоваться функцией, которая обучена на других образцах

Общие слова

- Мы можем считать, что каждый из признаков-координат является вещественным числом, но в практических реализациях достаточно часто используются целые числа определённой разрядности.

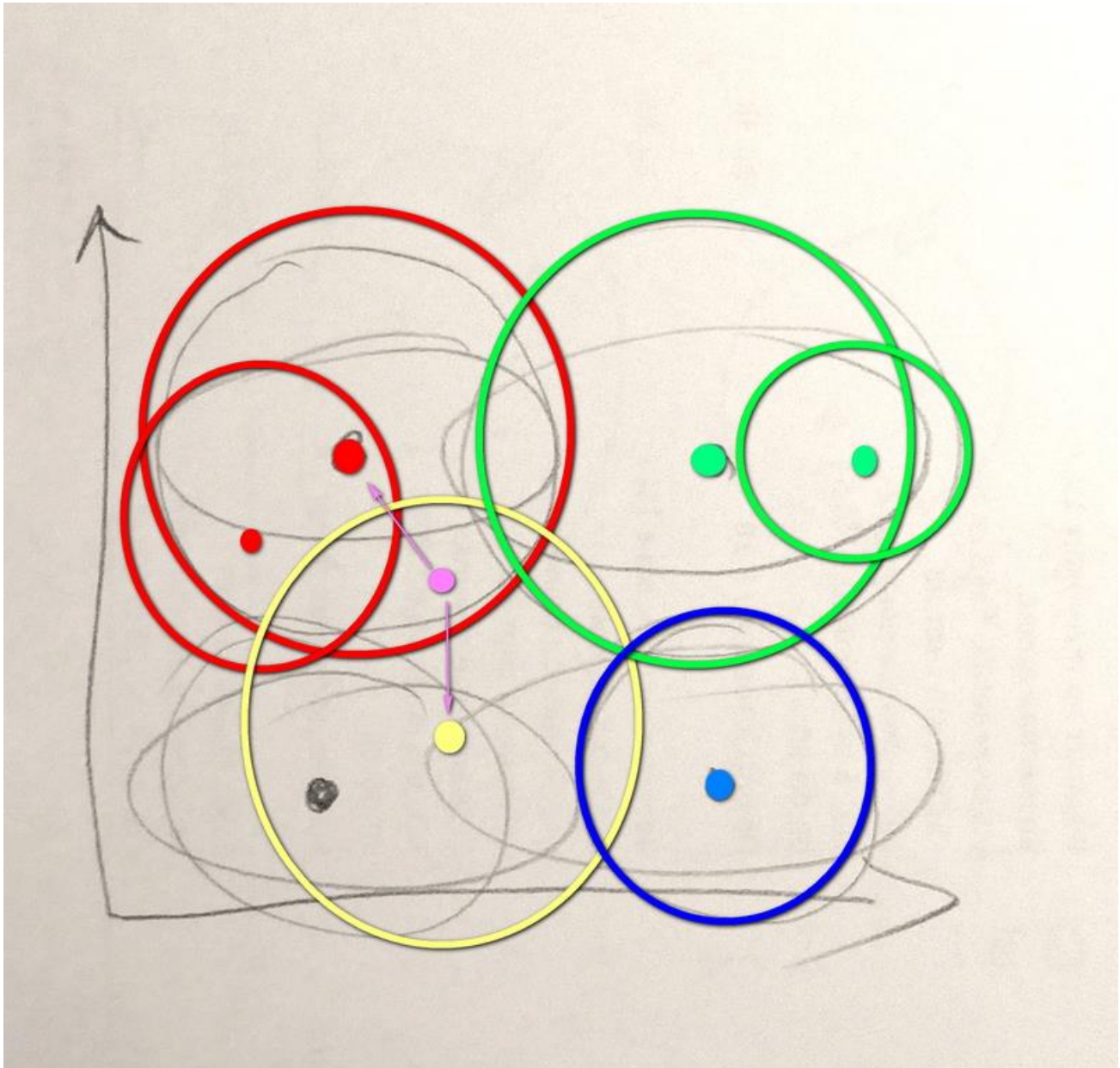
Чипы NeuroMem

- Чипы CM1K и NM500 предназначены для ускорения классификации с помощью RCE и KNN
- Чипы хранят N векторов-образцов размерностью до 256 и разрядностью 8 бит
- По мере запоминания образцов чипы автоматически вычисляют их веса
- Чипы умеют за фиксированное время, не зависящее от N, посчитать N расстояний от испытуемого вектора до образцов
- Чипы умеют за один такт найти среди вычисленных расстояний наименьшее, причем в соответствии с весами образцов

Пример вычислений

- N = 6 размерность 2.

  За 2 такта будет введена исследуемая точка.
  За 1 такт найден ближайший образец (красный)

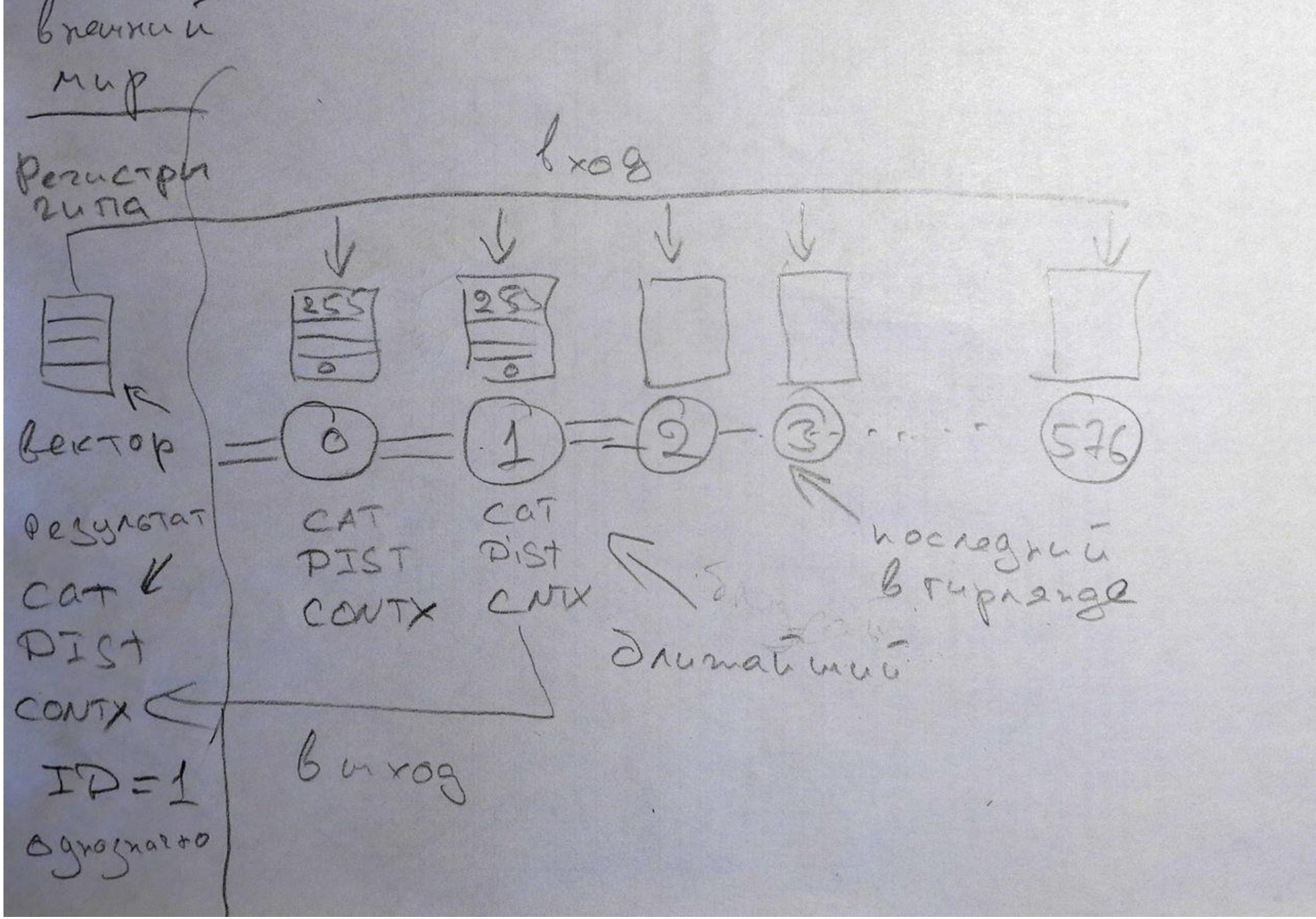


Устройство чипов NeuroMem

- Чип состоит из множества (500-1000) примитивных процессоров-нейронов. Один нейрон соответствует одному образцу
- Нейроны содержат по 256 байт памяти, что соответствует пространству с 256 измерениями и целочисленным координатам от 0 до 255.
- В регистрах нейронов хранятся категория, вес-энергия, минимально допустимый вес, контекст (аналог VLAN)
- Нейрон умеет считать дистанцию от себя до классифицируемого вектора

Шина чипа

- Чипы объединены широковещательной шиной передачи данных, которая доставляет поступающие на вход чипа векторы одновременно всем процессорам
- Шина ответа позволяет считать память и регистры самого ближнего образца + качественный ответ (0, 1, много категорий)
- Дополнительная линия объединяет чипы в гирлянду от первого, до последнего обученного


внешний мир
Регистры чипа
вход
255
0
255
0
вектор
0
1
2
3
576
CAT DIST CONTX
CAT DIST CNTX
последний в гирлянде
элиматический
результат CAT DIST CONTX
выход
IP=1
однозначно

Состояния нейрона

- Начальное - за пределами гирлянды; категория равна нулю; память пуста; нейрон не участвует ни в обучении ни в распознавании
- Готов к обучению - в хвосте гирлянды; память пуста; нейрон участвует только в обучении
- Обучен - в гирлянде; память хранит точку в пространстве признаков; нейрон участвует в обучении и распознавании
- Обучен, но Деградировал - память хранит точку в пространстве признаков; вес минимален; нейрон участвует только в распознавании

Процесс распознавания

- Вектор координата за координатой передается нейронам
- После приёма очередной координаты все нейроны пересчитываются расстояния
- Шина ответа подключает регистры чипа к регистрам нейрона, ближайшего в пространстве признаков и выставляет качественную оценку результата

Считывание результата

- В регистре NSR (Network Status Register) содержится качественный результат распознавания
- Одновременно с качественным результатом становится доступным данные с ближайшего к вектору нейрона. Можно прочитать *DIST, CAT , NID , NCR* - контекст нейрона
- Считывание сбрасывает флажок срабатывания ближайшего нейрона и шина ответа переключается на следующий по расстоянию

Процесс обучения

- Если после отправки вектора записать категорию в регистр CAT то вектор используется для обучения
- Если сочетание вектор – категория уникальны, то к гирлянде присоединяется новый нейрон, который заполняется этими данными.
- Если распознавание категории вектора было неоднозначным, то все сработавшие «чужие» нейроны уменьшают свой вес на вычисленное расстояние.

Управление весами

- Регистр максимального веса используется для установки начального веса нейрона
- Регистр минимального допустимого веса запоминается в нейроне и используется при контроле при снижении веса.
- Если вес снизился ниже минимального, то он устанавливается в минимальный, а нейрон помечается как малоинформативный (слишком часто срабатывает на чужие категории)

Работа в режиме KNN

- Веса игнорируются
- Обучение не требуется. Достаточно просто записать координаты образцов в нейроны
- Всегда выстреливают все нейроны
- По желанию (и при наличии времени) можно прочитать содержимое ближайшего нейрона или К ближайших

Режим сохранения-восстановления

- Особый режим внутренней сети чипа позволяет извлечь состояние нейронов, обнулить их и залить в них данные из внешнего источника.
- Эта методика позволяет тренировать сеть на одном лабораторном чипе, а потом заливать ее в серийные изделия или продавать, как интеллектуальную собственность.
- Смешивать нейроны без обнуления сети не рекомендуется за исключением режима KNN

Регистры чипа

NSR	Результат распознавания + выбор режима RCE/KNN
GCR	Control Register
MINIF	Минимальный возможный вес
MAXIF	Максимальный возможный вес
NCR	Контекст нейрона + старший байт ID
COMP	Ячейка памяти. В нормальном режиме служит для рассылки компонентов вектора
INDEXCOMP	Номер текущей ячейки памяти нейрона (0-255). Автоматически увеличивается при каждой операции чтения-записи COMP
DIST	Вычисленное расстояние ближайшего нейрона. Каждое последующее считывание возвращает следующее лучшее расстояние. Расстояние 0xFFFF означает, что всё считано
CAT	Категория нейрона. Считывается после DIST. Категория 0xFFFF означает, что всё считано
AIF	Текущий вес нейрона
NID	Идентификатор нейрона (2 младших байта). Автоматически назначается при обучении нового нейрона.
FORGET	Команда очисткки нейронов. Сбрасывает их категорию в 0.
NCOUNT	Число обученных нейронов (младшие 2 байта) 0xFFFF - сеть заполнена
RESETCHAIN	Команда перехода к первому нейрону при восстановлении сети

Быстродействие

Operation	Clock cycles	@35 Mhz (single chip) L=256, N=576, K=3
Broadcast a vector of Length L	L+3	7.3 µs
Learn a vector of length L	L+3 + 18	7.9 µs
Status of a vector of length L	L+3+1	7.4 µs
Best match of a vector of length L	L+3+37	8.3 µs
Get the K top match of a vector of length L	L+3+ (K*37)	9.1 µs
Save N neurons	4+ (N*260)	4.27 ms
Restore N neurons	4+ (N*260)	4.27 ms

КОНЕЦ