



Обзор методов машинного обучения

Малых Валентин,
Отдел качества поиска
malih@corp.sputnik.ru

Ноябрь 2014

Зачем нужно?

- Асессоры, люди, которые могут оценить, насколько хорош или релевантен документ, но они не могут объяснить, как они это делают.
- Мы приближаемся к оценке людьми нашей выдачи по их запросам.

Где применяется у нас?

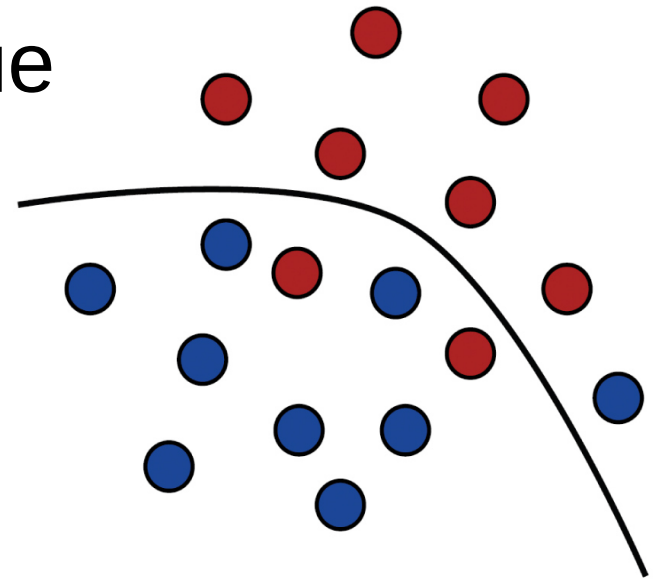
- Ранжирование
 - Самое важное
- Тематизация запросов
- Тематизация документов
- Детекция спама
- Детекция синонимов
 - и, вероятно, многое другое

Методы машинного обучения

1. Что такое машинное обучение?
2. Классификация
3. Примеры

Что такое машинное обучение?

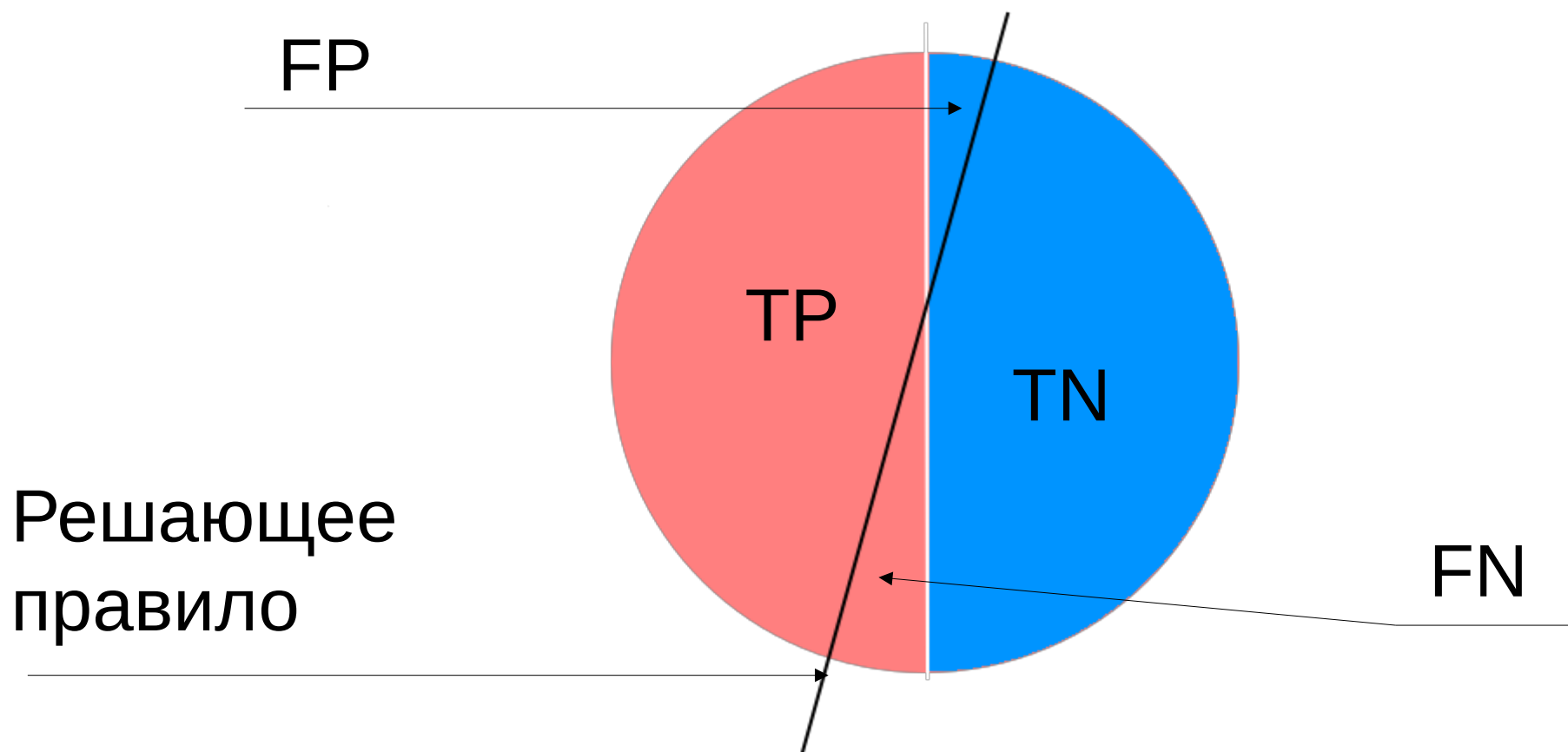
- Сбор данных
- Выделение особенностей (features)
- Построение модели данных
- Переобучение / недообучение



Признаки

- Числовые (целые и рациональные)
- Бинарные
- Заданные перечислением
- Нормализация признаков

Оценка качества



Оценка качества

- Точность (Precision)

- $$\text{Precision} = \frac{TP}{TP + FP}$$

- Полнота (Recall)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Оценка качества

Мера ван Рийсбергена
(F-мера)

$$F = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \right)}$$

Классификация

- С учителем / без учителя (supervised / unsupervised)
 - Еще можно отметить reinforcement learning – обучение с подкреплением
- По механизму (by approach)
 - Следующий слайд

Различные механизмы машинного обучения

- Статистическая классификация – Байес
- Классификация на основе сходства – kNN
- Классификация на основе разделимости - SVM
- Нейронные сети – перцептрон
- Индукция правил – решающие деревья
- Кластеризация – k-means
- Регрессия – линейная регрессия

Различные механизмы машинного обучение (продолжение)

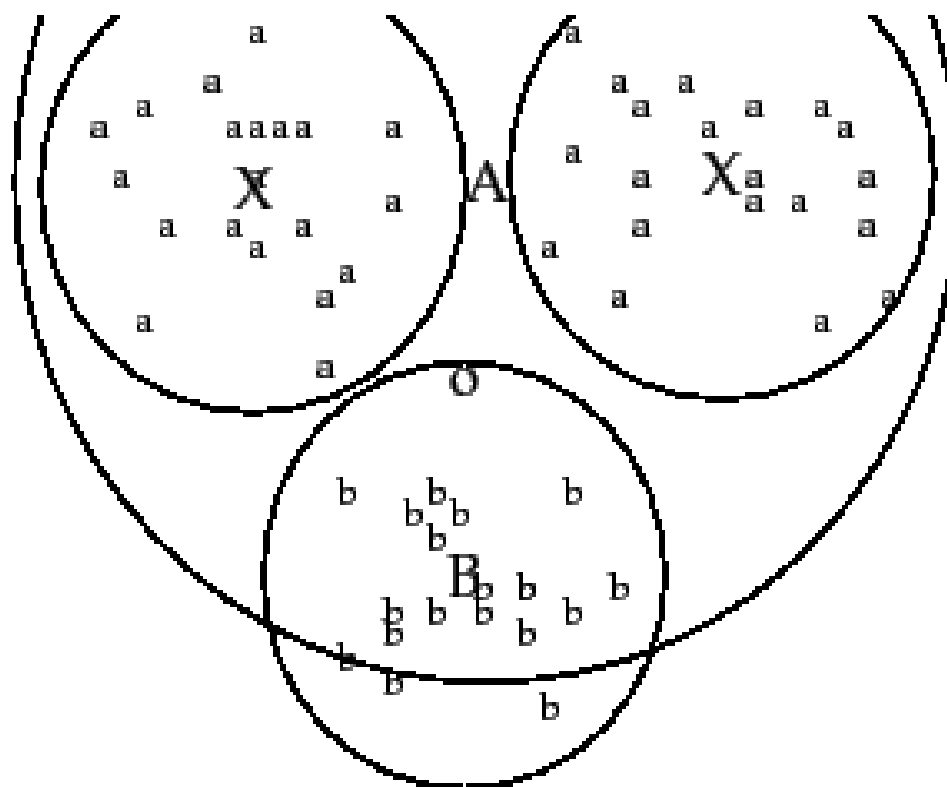
- Алгоритмические композиции – бустинг
- Сокращение размерности - метод главных компонент

Классификатор Роше

- Вычисляются центроиды для всех классов в обучающей выборке
- Результат получается, как близость исследуемого объекта к классу

$$\bar{g}_c = \frac{1}{|R_c|} \sum_{d \in R_c} \bar{d} - \gamma \frac{1}{|R_{c,k}|} \sum_{d \in R_{c,k}} \bar{d}$$

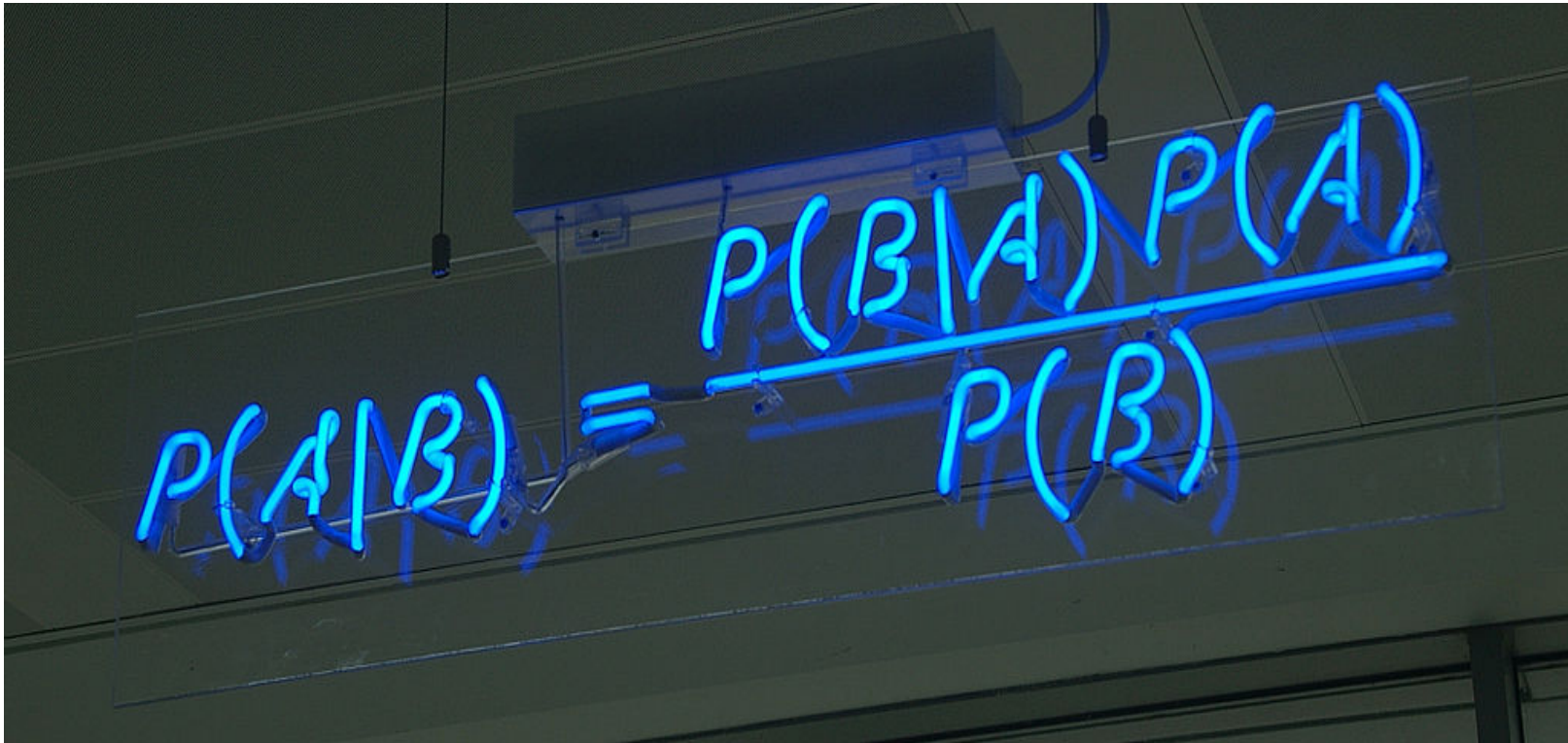
Классификатор Роше



► **Figure 14.2** The multimodal class "a" consists of two different clusters (small upper circles centered on X's). Rocchio classification will misclassify "o" as "a" because it is closer to the centroid A of the "a" class than to the centroid B of the "b" class.

Байесовский метод

- Формула Байеса
 - Да, она такая важная.

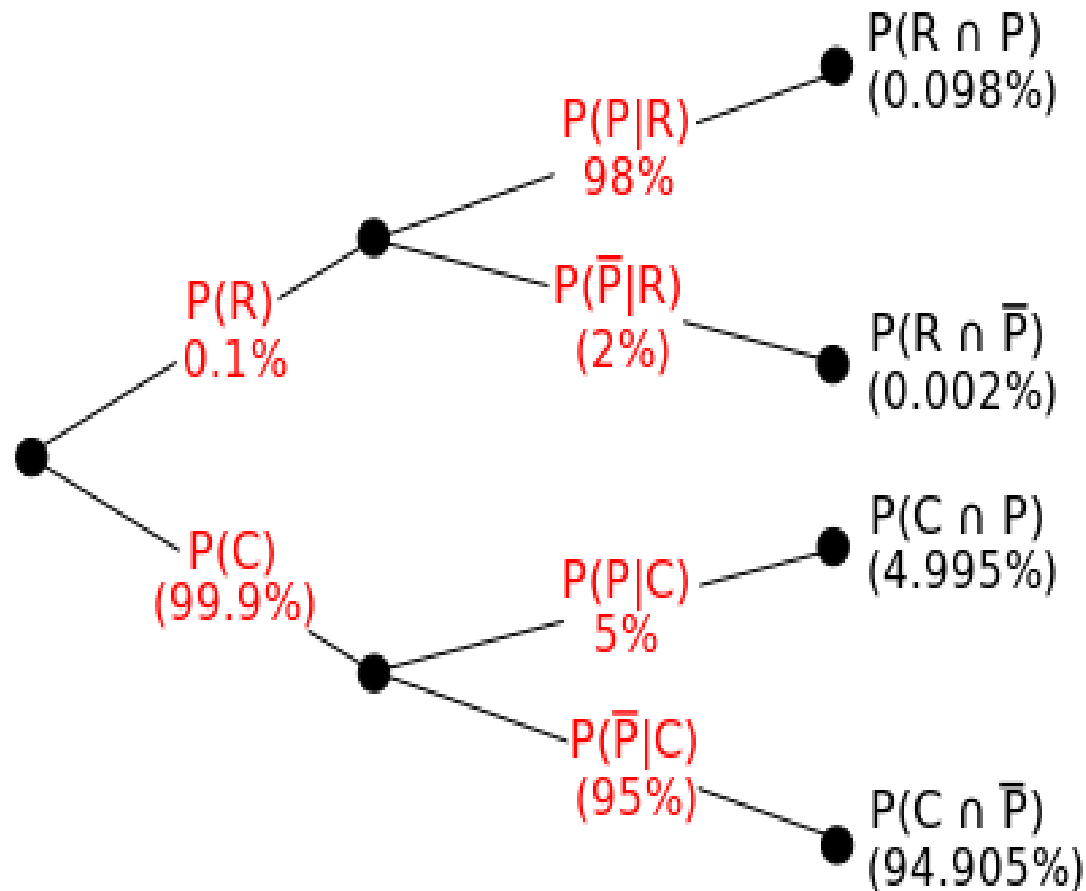


A blue neon sign displaying the formula for Bayes' theorem. The sign is illuminated and mounted on a dark background. The formula is written in a stylized, glowing blue neon font. It reads: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The sign is slightly tilted and has a soft glow around it.

Байесовский метод

- Признак Р и его отсутствие

- Классы: R и C



Линейная регрессия

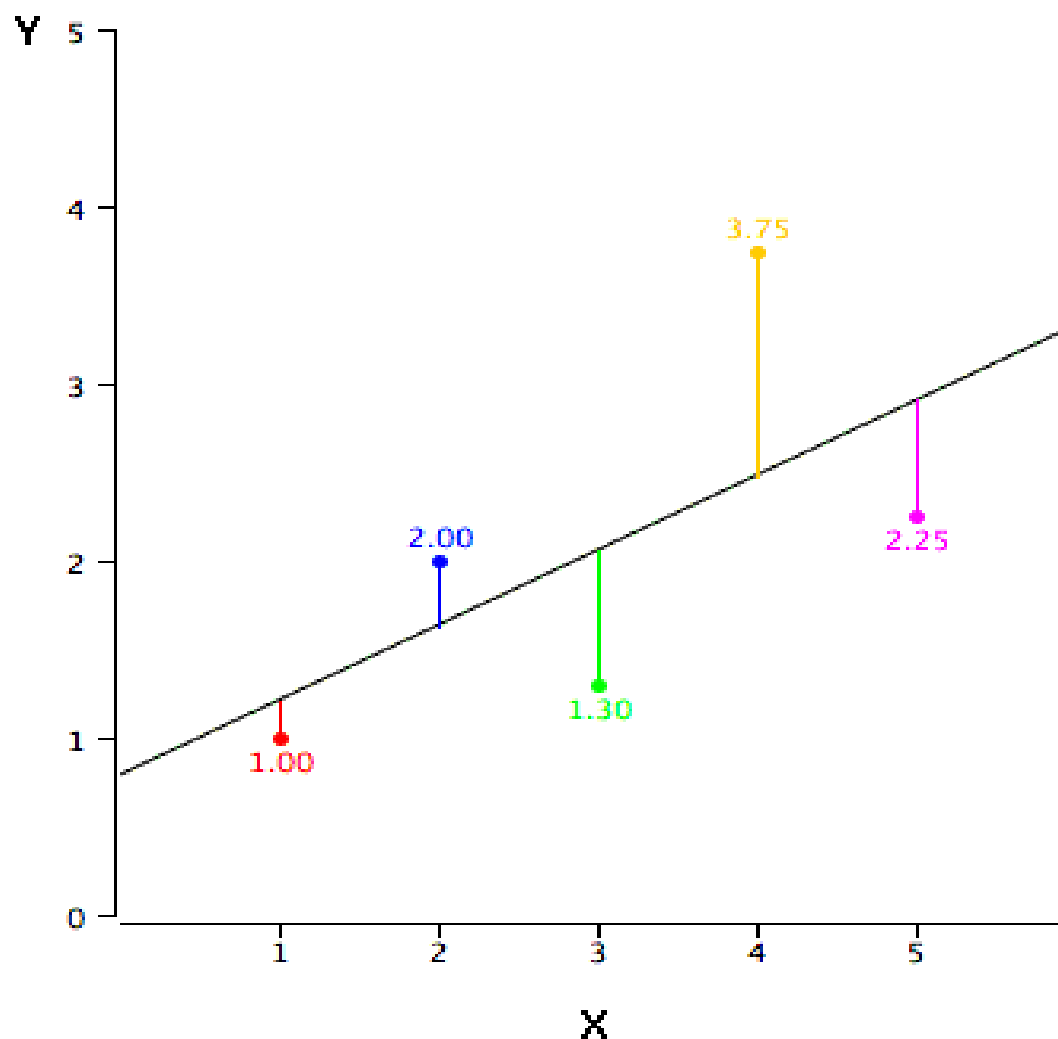
- Линейная функция для приближения

$$Y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n = b_0 + (\vec{b}\vec{x})$$

$$err = \sum_{j=1}^M (\vec{b}\vec{x}^j + b_0 - Y^j)^2$$

- Для оценки – метод наименьших квадратов

Линейная регрессия



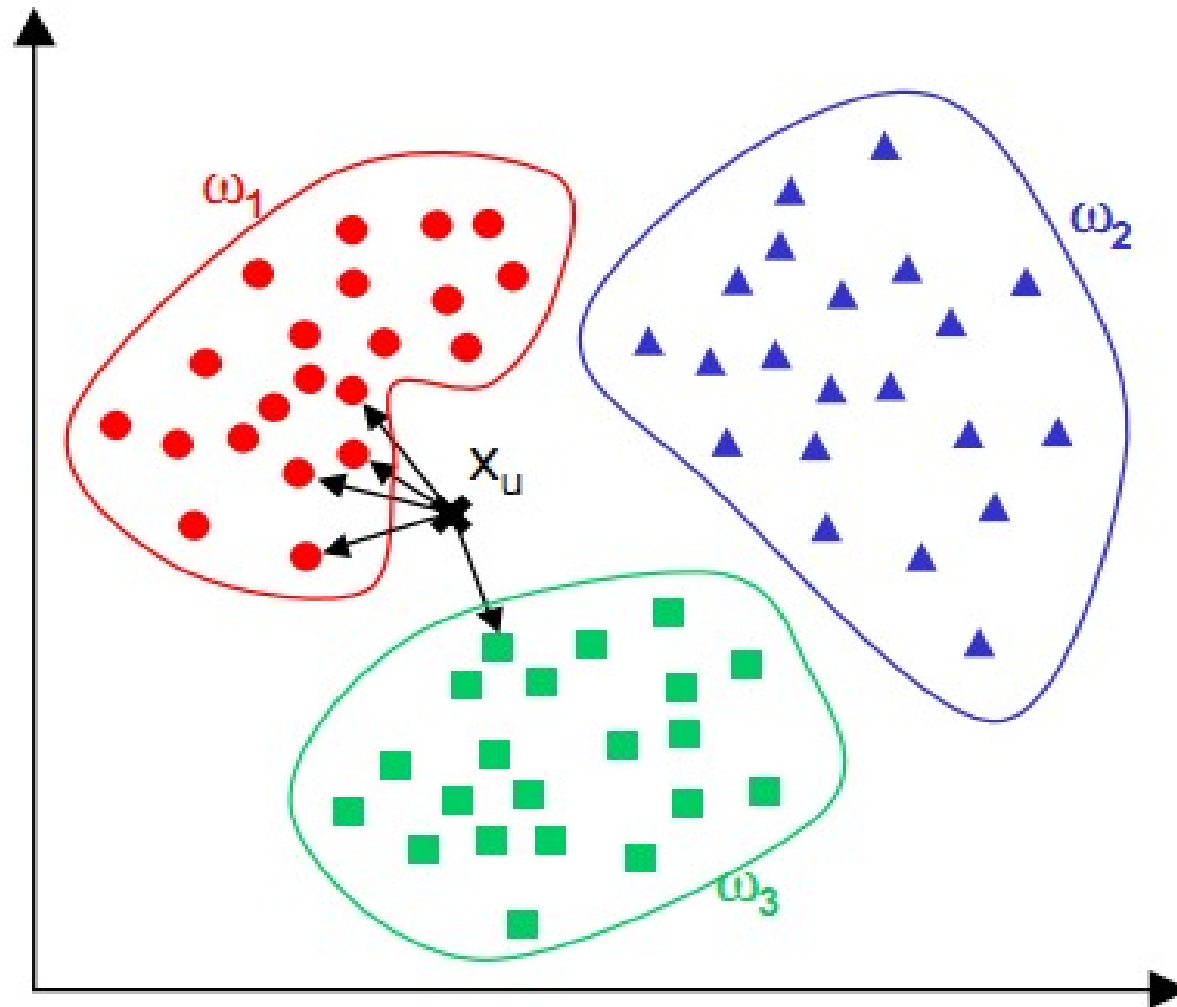
Метод k ближайших соседей (kNN)

- В общем случае:

$$a(u) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^m [x_{i,u} = y] w(i, u),$$

- Где $w(i, u)$ задает специфику подхода

Метод k ближайших соседей (kNN)



Метод k средних (k-means)

- Выбираем k
- Выбираем начальные средние
- Приписываем каждую точку к ближайшему кластеру
- Строим центроиды для существующих кластеров – это новые средние

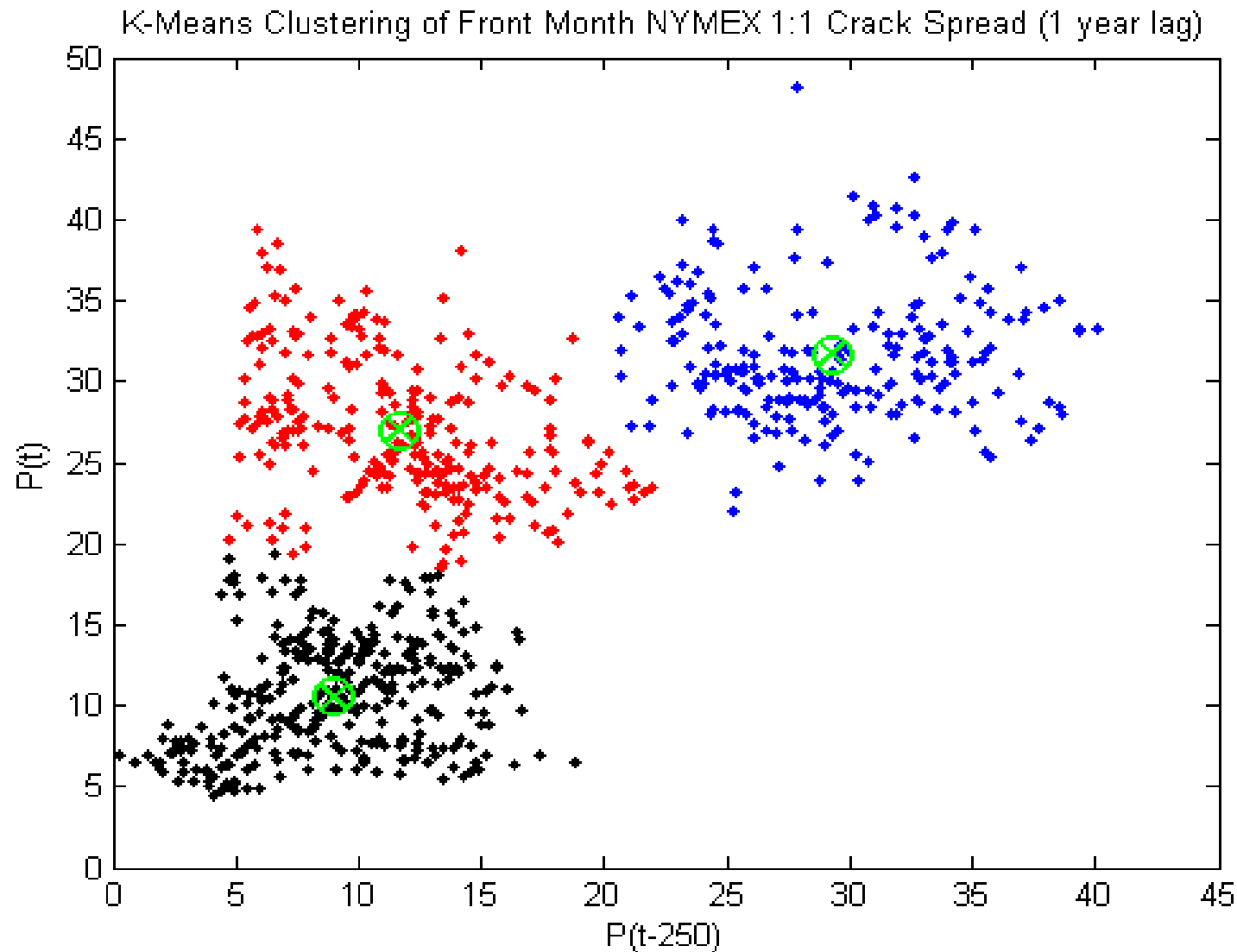
$$\arg \min_{\mathbf{c}} \sum_{i=1}^k \sum_{\mathbf{x} \in c_i} d(\mathbf{x}, \mu_i) = \arg \min_{\mathbf{c}} \sum_{i=1}^k \sum_{\mathbf{x} \in c_i} \|\mathbf{x} - \mu_i\|_2^2$$

$$\mu_i = \text{some value}, i = 1, \dots, k$$

$$c_i = \{j : d(\mathbf{x}_j, \mu_i) \leq d(\mathbf{x}_j, \mu_l), l \neq i, j = 1, \dots, n\}$$

$$\mu_i = \frac{1}{|c_i|} \sum_{j \in c_i} \mathbf{x}_j, \forall i$$

Метод k средних (k-means)



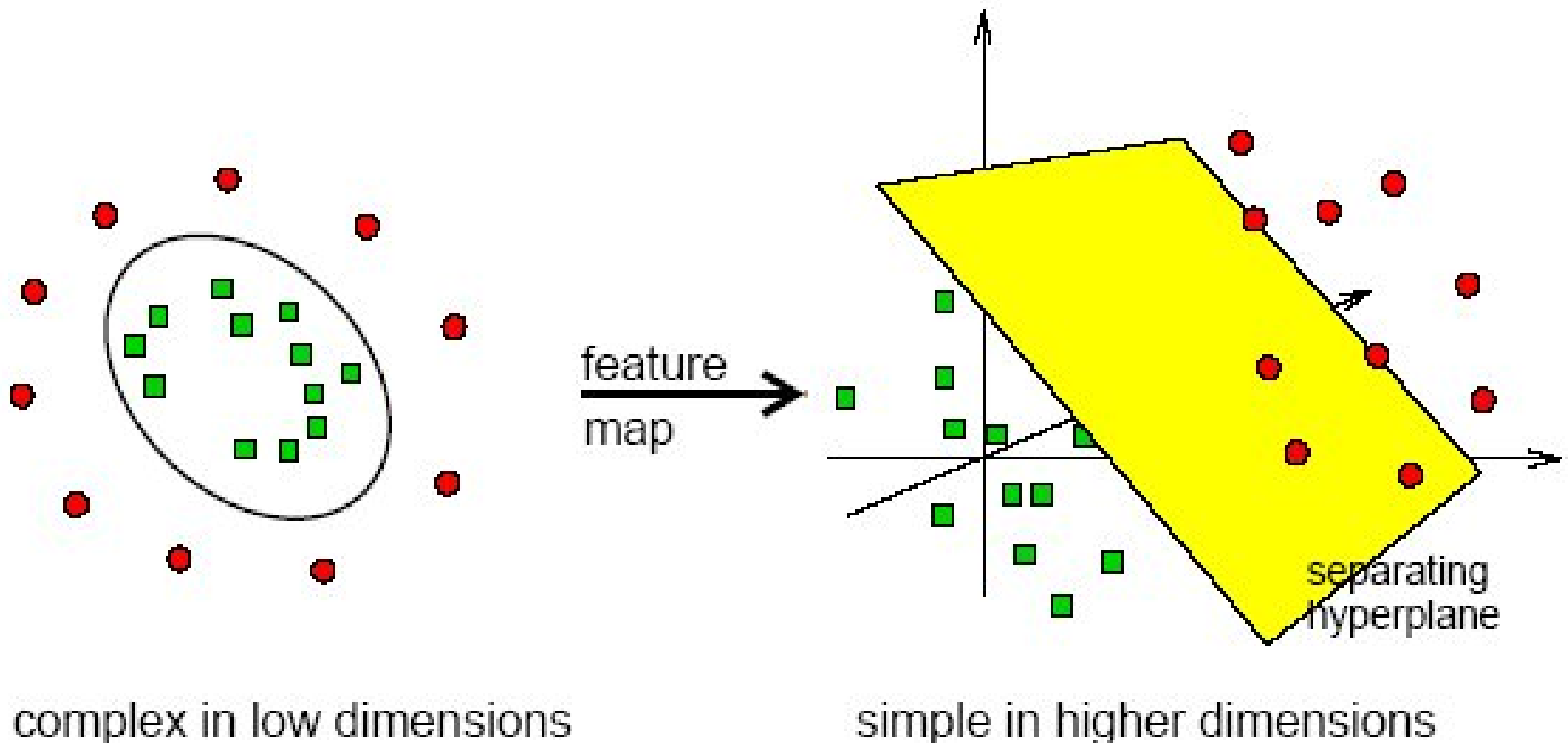
Support Vector Machine

- Ищет максимально чистое разделение в пространстве, заданном преобразованием w

$$\arg \min_{\mathbf{w}, b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i - b) - 1] \right\}$$

Support Vector Machine

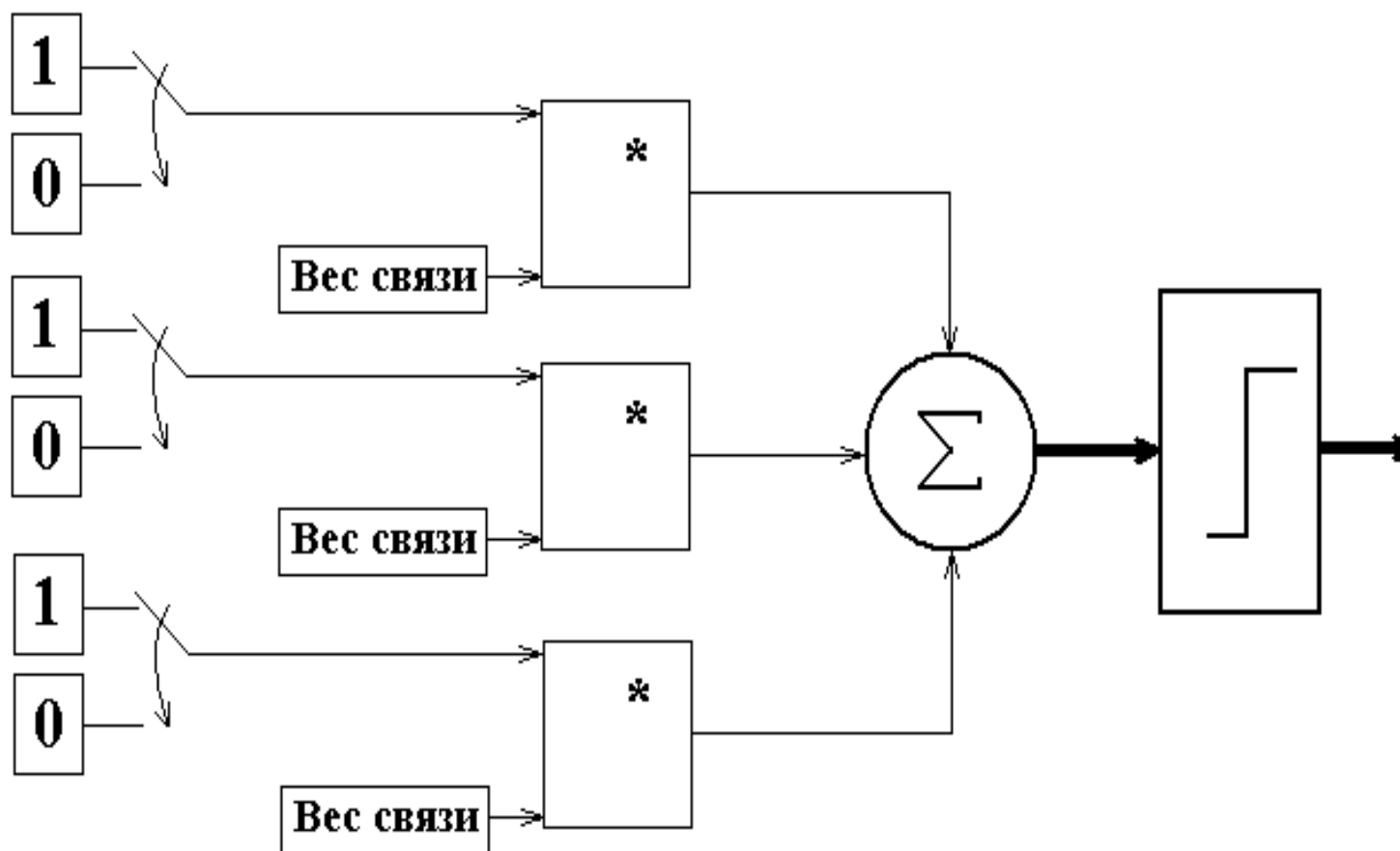
Separation may be easier in higher dimensions



Перцептрон

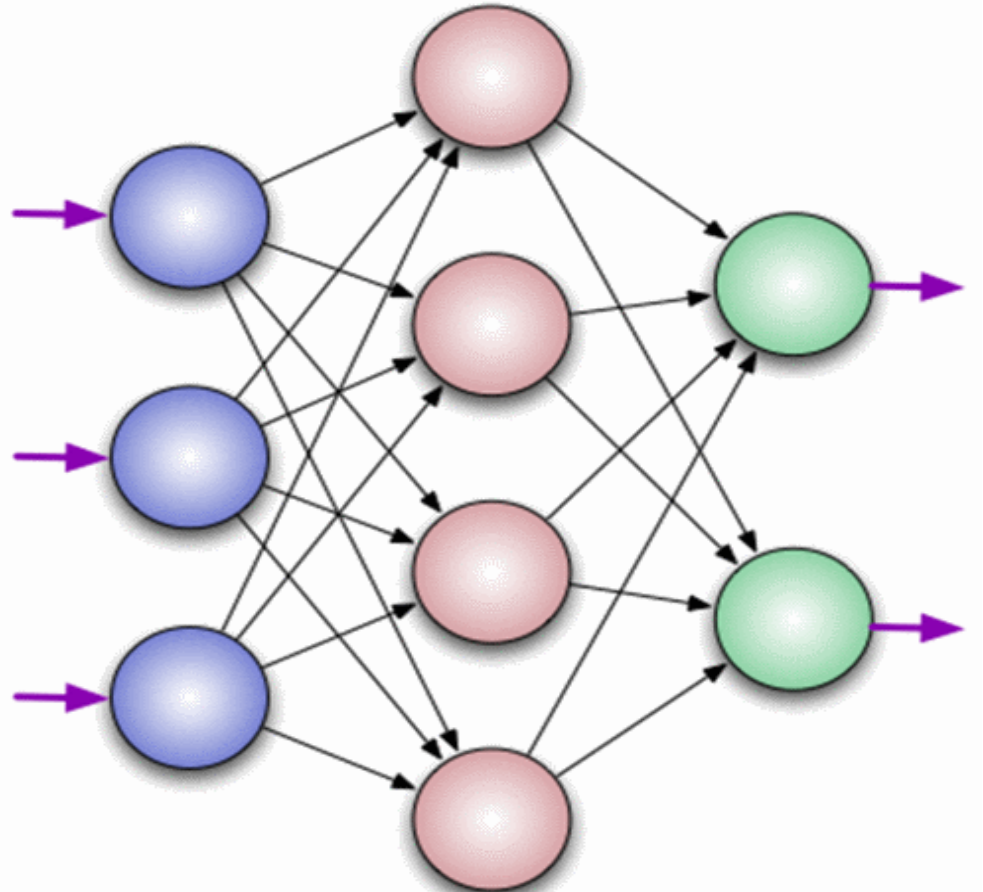
- Рецепторы (S-элементы)
 - Могут принимать два значения -1 и 1
- Ассоциаторы (А-элементы)
 - Возбуждаются по порогу
- Реактор (R-элемент)
 - Подсчитывает линейную форму от входов

Перцептрон



Нейронные сети

- Deep Learning
- Входы
- Выходы
- Скрытый слой



Метод главных компонент (Principal Component Analysis)

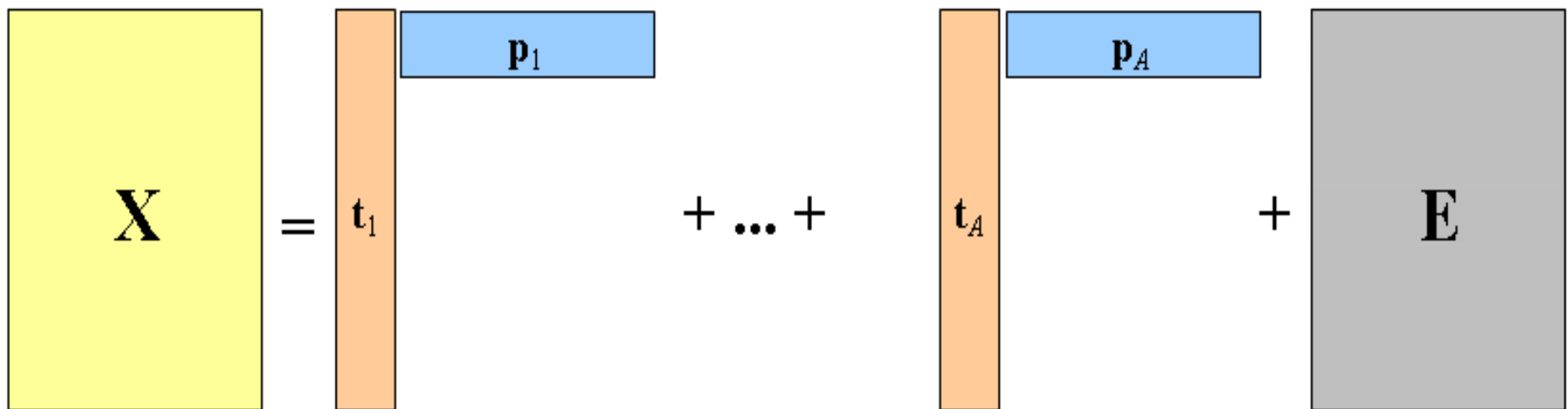
- Формальные переменные t

$$t_a = p_{a1}x_1 + \dots + p_{aJ}x_J$$

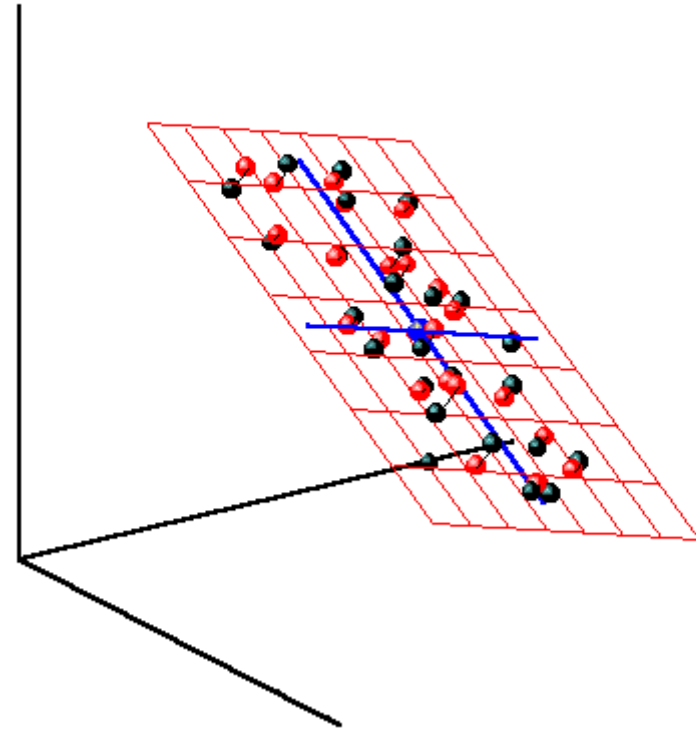
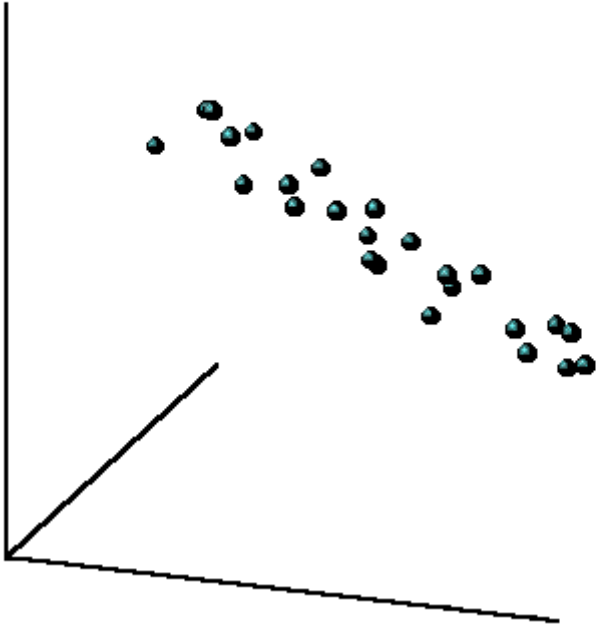
- Разложение изначальной матрицы на две:
 - T (scores – счета)
 - P (loading - нагрузки)

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} = \sum_{a=1}^A \mathbf{t}_a \mathbf{p}_a^t + \mathbf{E}$$

Метод главных компонент (Principal Component Analysis)



Метод главных компонент (Principal Component Analysis)



Решающие деревья

- Лучше воспринимаемы человеком
- Способы работать не на всем признаковом пространстве

Решающие деревья

```
if wage increase first year  $\leq$  2.5 then
  if working hours  $\leq$  36 then class good
  else if working hours  $>$  36 then
    if contribution to health plan is none then class bad
    else if contribution to health plan is half then class good
    else if contribution to health plan is full then class bad
else if wage increase first year  $>$  2.5 then
  if statutory holidays  $>$  10 then class good
  else if statutory holidays  $\leq$  10 then
    if wage increase first year  $\leq$  4 then class bad
    else if wage increase first year  $>$  4 then class good
```


Как работает алгоритм C4.5

- Информация – уменьшение энтропии

$$info(S) = - \sum_{j=1}^k \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right)$$

$$info_X(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i)$$

Как работает алгоритм C4.5

- Приращение информации (information gain)
 - Оно же Kullback-Leibler divergence

$$gain(X) = info(T) - info_X(T)$$

Градиентный бустинг

$$F_M(x) = \sum_{m=1}^M b_m h(x; a_m), b_m \in R, a_m \in A.$$

$$F_m(x) = F_{m-1}(x) + b_m h(x; a_m), b_m \in R, a_m \in A.$$

$$Q = \sum_{i=1}^N L(y_i, F_m(x_i)) \rightarrow \min .$$

Спасибо за внимание!

Вопросы?

ЕМ-алгоритм

- Expectation – по существующему вектору параметров вычисляем вектор скрытых переменных
- Maximization – по существующему вектору скрытых переменных с помощью минимизации эмпирического риска вычисляем вектор параметров