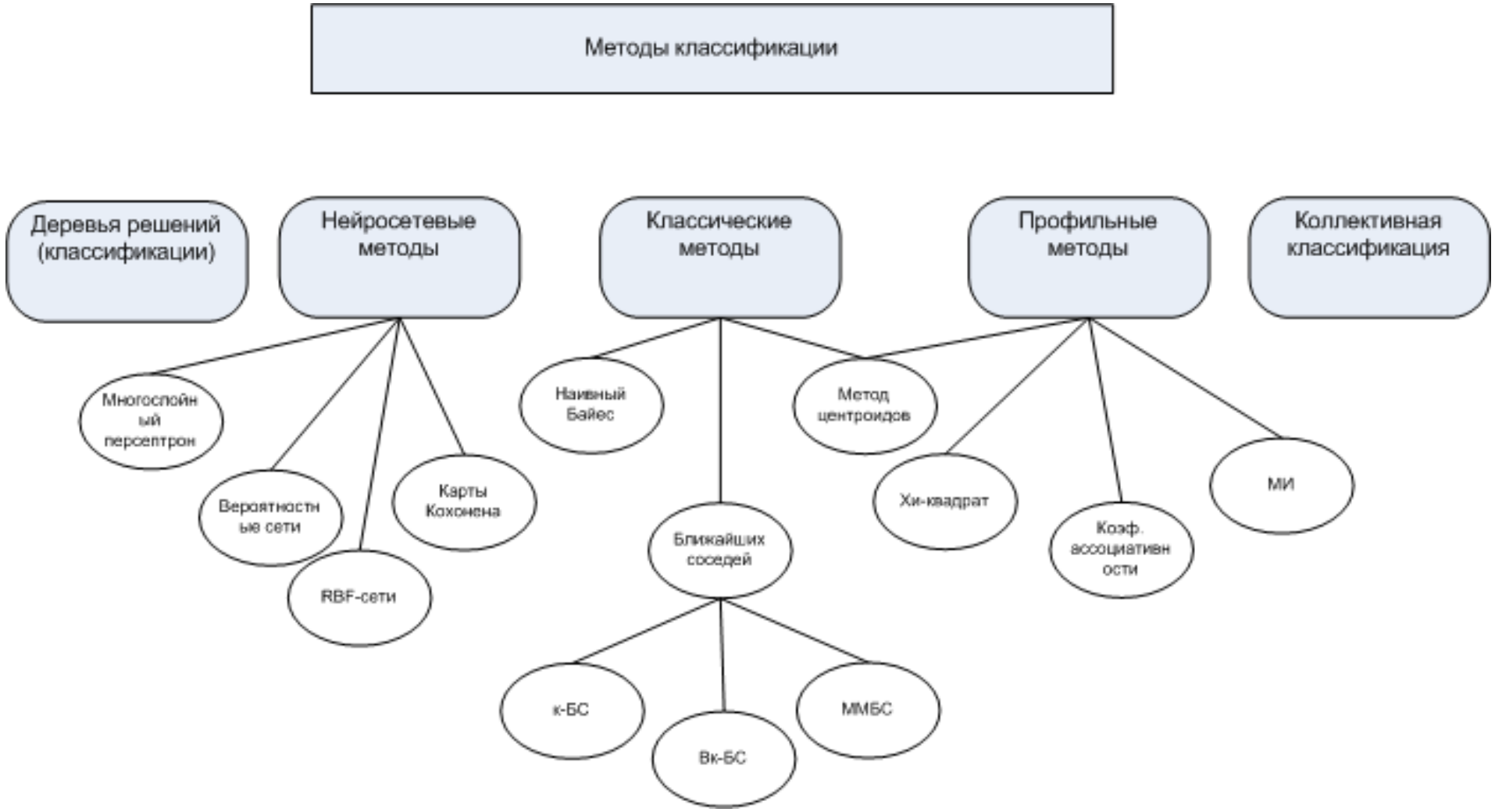
# *Обзор методов классификации*

Курс «Интеллектуальные информационные системы»

Кафедра управления и информатики НИУ «МЭИ»

Осень 2022 г.

# Систематизация методов классификации

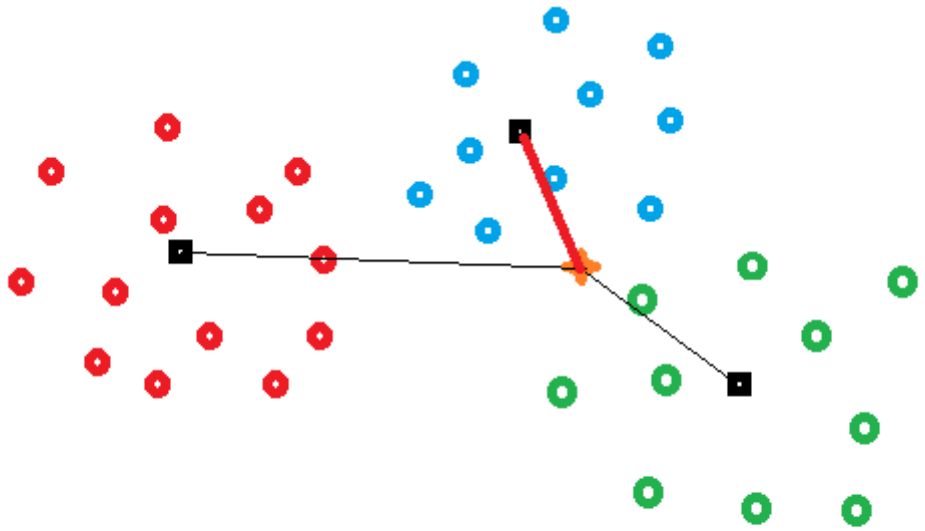


# Центроидный метод

Центроид – вектор со средними значениями весов признаков объектов данного класса. «Центр тяжести».
Классифицируемый объект относится к классу с наиболее близким центроидом.

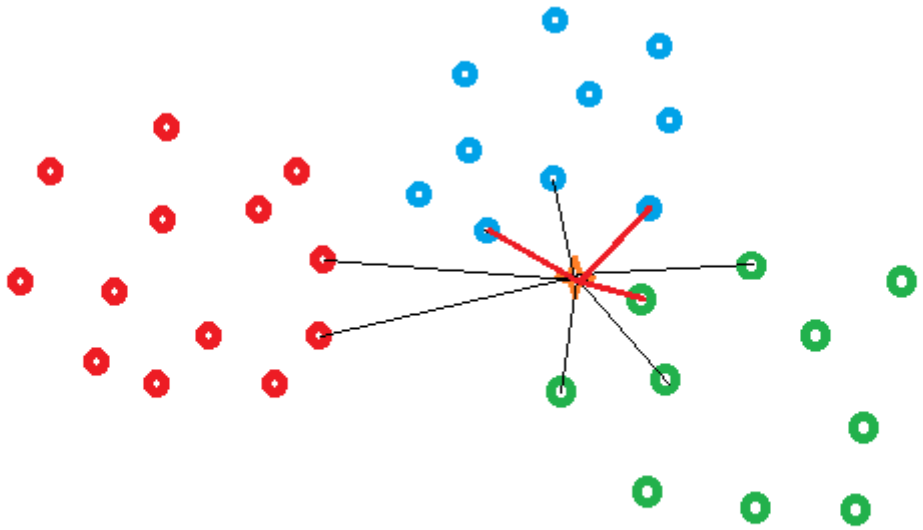
$$\vec{\mathbf{C}}_k = \frac{1}{N_k}\sum_{j=1}^{N_k}\vec{X}_j$$

Роккио: $$\vec{C}_k = \alpha\frac{1}{N_k}\sum_{j=1}^{N_k}\vec{X}_j - \beta\frac{1}{N-N_k}\sum_{l=1}^{N-N_k}\vec{X}_l$$



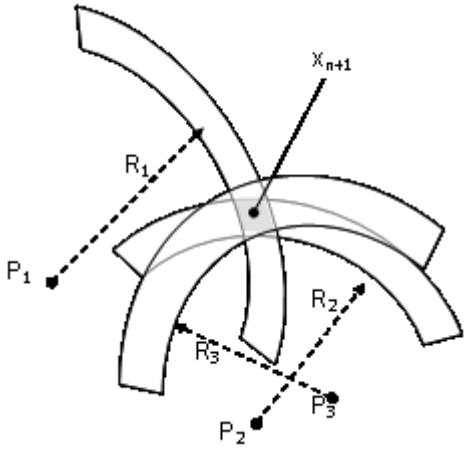
# *Правило ближайшего соседа (БС)*

Классифицируемый объект относится к тому классу, к которому относится ближайший к нему сосед.



# *Семейство методов БС*

- кБС – Решение принимается на основании анализа к ближайших соседей. Обычно к - нечетное число [5;25]
- Взвешенный кБС – наиболее близкие соседи имеют больший вес при голосовании.
- Модифицированный МБС – поиск соседей только определенной области признакового пространства, с целью сокращения вычислительных  операций.



# *Наивный байесовский метод (НБ)*

теорема Байеса: $$P(Q_k \mid \vec{X}) = \frac{P(\vec{X} \mid Q_k)P(Q_k)}{P(\vec{X})}$$

позволяет определить вероятность какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие.

P (“болен”| болен) = 0.9
P (“болен”| здоров) = 0.01
P (болен) = 0.001; P (здоров) = 1 - 0.001 = 0.999
P(«болен») = P (“болен”| болен) * P (болен) + P (“болен”| здоров) * P (здоров) = 0.0109

P (здоров | «болен») $= \frac{P(\text{"болен"}|\text{здоров}) * P(\text{здоров})}{P(\text{"болен"})} = \frac{0.01 * 0.999}{0.0109} \approx 0.917$

# *Наивный байесовский метод (НБ)*

теорема Байеса: $$P(Q_k \mid \vec{X}) = \frac{P(\vec{X} \mid Q_k)P(Q_k)}{P(\vec{X})}$$

- $P(\vec{X})$ - одинакова для различных классов и может быть исключена из дальнейшего рассмотрения
- Допущение: признаки (свойства, термины,…), которыми описывается объект, независимы между собой.
- Данное допущение значительно упрощает задачу, но крайне редко выполняется на практике

$$P(\vec{X} \mid Q_k) = \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

$$P(Q_k \mid \vec{X}) = P(Q_k) \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

# Наивный байесовский метод (2)

$$P(Q_k \mid \vec{X}) = P(Q_k)\prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

- $\hat{P}(Q_k) = \frac{N_k}{N}$ - оценка для P(Qk) – вероятность встретить объект класса Qk в выборке

- $\hat{P}(x^{(i)} \mid Q_k) = \frac{N_{ik}}{N_k}$ - вероятность встретить признак x(i) в классе Qk

- Часто (особенно в задачах Text Mining) используется мультиномиальная реализация:

$$\hat{P}\left(x^{(i)} \middle| Q_k\right) = \frac{\alpha + N_{ik}}{\alpha M + N_k}$$

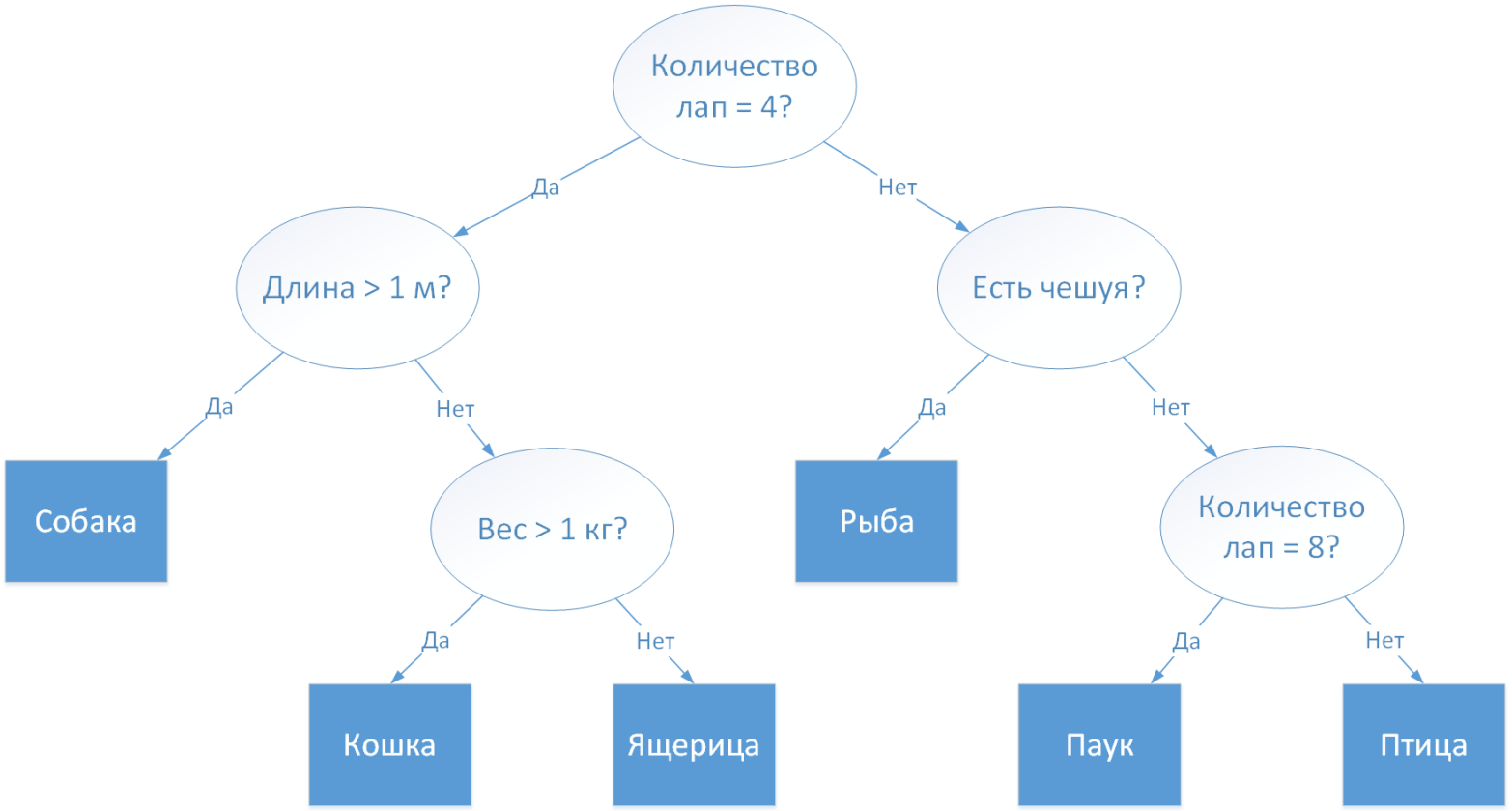
где M – общее количество признаков (терминов) во всех объектах (документах) выборки

$$P\left(Q_k \middle| \vec{X}\right) = argmax\ \frac{N_k}{N}\prod_{i=1}^{M}\frac{\alpha + N_{ik}}{\alpha M + N_k}$$

# *Метод деревьев решений*

Средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей.

В методе деревьев решений проводится последовательное разделение множества объектов на основе значений выбранного признака, в результате чего строится дерево, содержащее нетерминальные узлы (узлы проверок), в которых происходит разбиение по выбранному атрибуту, и терминальные узлы (узлы ответа), в которых должны находиться элементы одного класса.



# *Метод деревьев решений. Критерий прироста информации*

Для выбора наиболее информативного признака, по которому проводится разбиение, в методе деревьев решений чаще всего используется *теоретико-информационный (энтропийный) подход*.

Хотим найти такой признак $x^{(S)}$, при разбиении по которому один из классов имел наибольшую вероятность появления. Это возможно, если величина прироста информации *Gain* будет достигать своего максимума.

$$Gain(x^{(s)},T) = I(T) - I(x^{(s)},T)$$

$$I(T) = \sum_{k=1}^{K} P_k \log_2 \frac{1}{P_k} = -\sum_{k=1}^{K} \frac{N_k}{N} \log_2 \frac{N_k}{N}$$

- среднее количество информации (энтропия), необходимое для определения класса примера из обучающей выборки *T*

$$I(x^{(s)},T) = \sum_{i=1}^{S} \frac{N_s}{N} I(T_s) = \sum_{s=1}^{S} \frac{N_s}{N} \left( -\sum_{k=1}^{K} \frac{N_{ks}}{N_s} \log_2 \frac{N_{ks}}{N_s} \right)$$

- среднее количество информации, необходимое для идентификации класса примера в каждом подмножестве после разбиения по признаку $x^{(S)}$

# Метод деревьев решений. Меры неоднородности

Еще один подход к выявлению признака, по которому стоит проводить разбиение – использовать меры неоднородности ф. Здесь вектор **p** состоит из **m** вероятностей меток встречающихся в некотором подмножестве обучающего множества

$$\phi(\vec{p}) = \sum_{i=1}^{m} p_i(1 - p_i)$$ Индекс Джини (Gini impurity)

$$\phi(\vec{p}) = 1 - \max(\vec{p})$$ Наиболее часто встречаемый класс

На каждой итерации для входного подмножества обучающего множества строится такое разбиение пространства гиперплоскостью (ортогональной одной их осей координат), которое минимизировало бы среднюю меру неоднородности двух полученных подмножеств. Данная процедура выполняется рекурсивно для каждого полученного подмножества до тех пор, пока не будут достигнуты критерии остановки.

# Метод деревьев решений. Пример разбиения