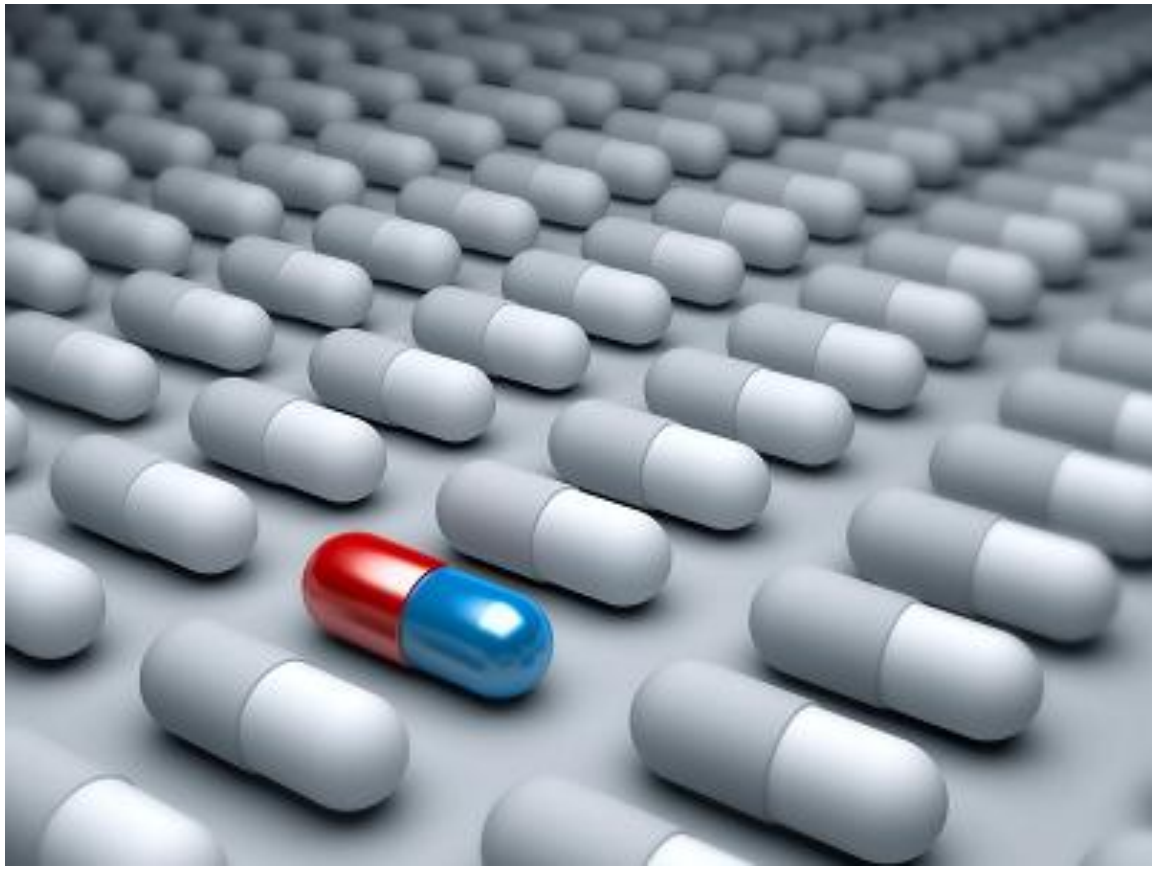
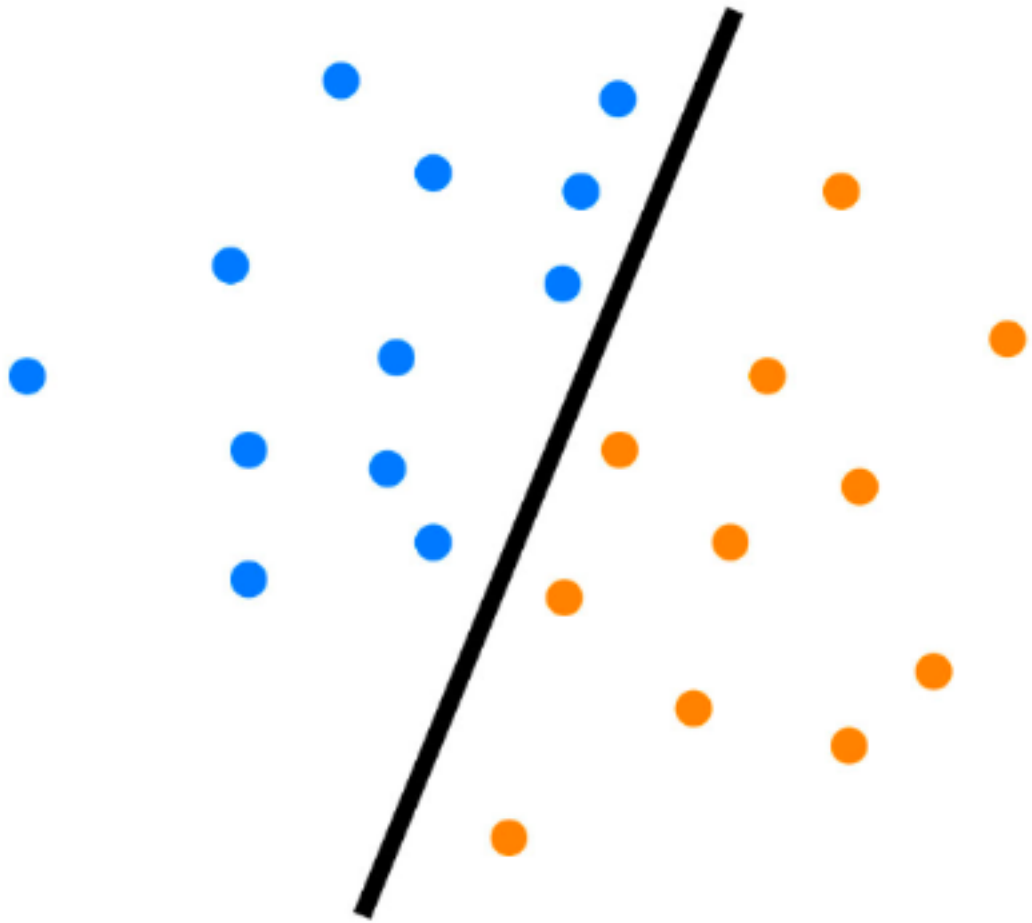
Построение медицинских тест-систем с использованием метода опорных векторов и метода ближайших центроидов





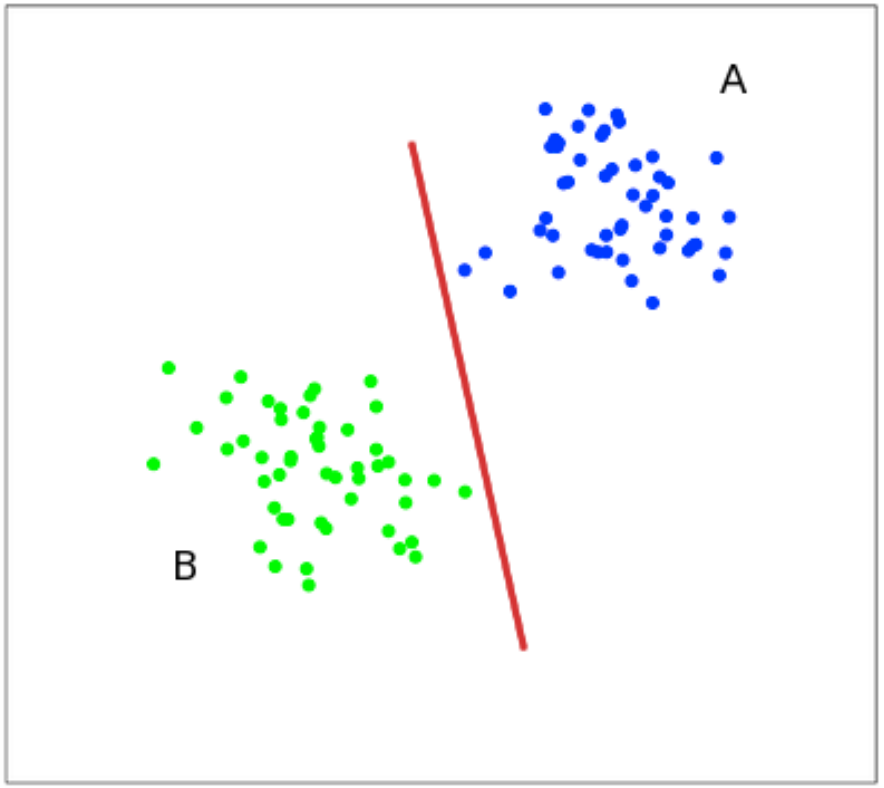
Задача классификации

- $X = \mathbb{R}^n$ - пространство объектов
- $Y = \{-1, +1\}$ - классы
- $(x_1, y_1), \dots, (x_m, y_m)$ - обучающая выборка
- $F: X \to Y$ - классификатор

Support Vector Machine

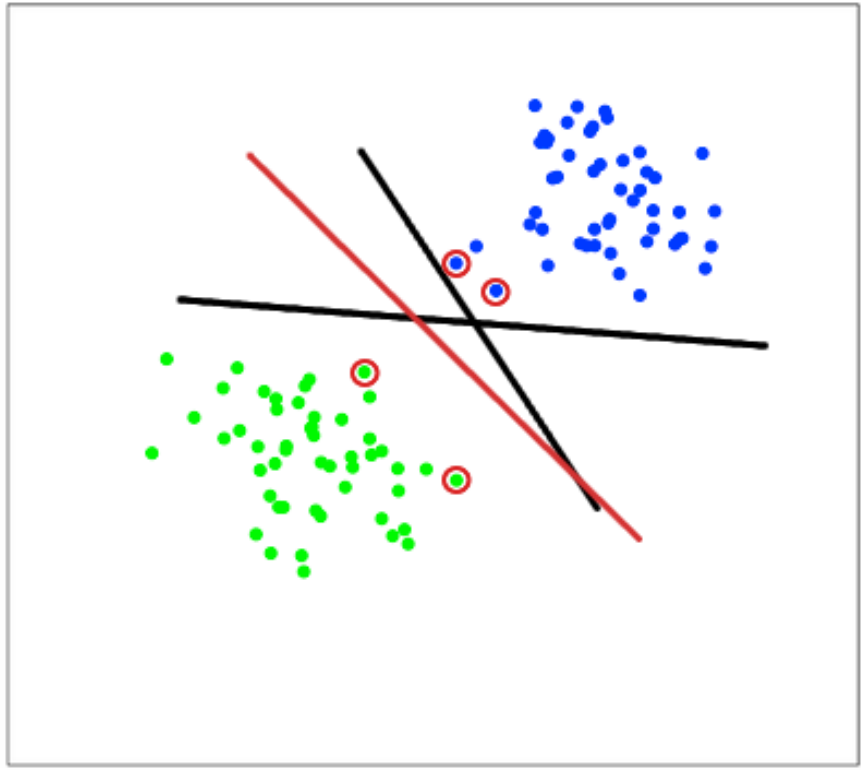
Идея метода (на примере):

- Даны точки на плоскости, разбитые на 2 класса
- Проведем линию, разделяющую эти два класса (разделяющая гиперплоскость)
- Все новые точки (не из обучающей выборки) автоматически классифицируются следующим образом: точки выше прямой – класс A, точки ниже прямой – класс B



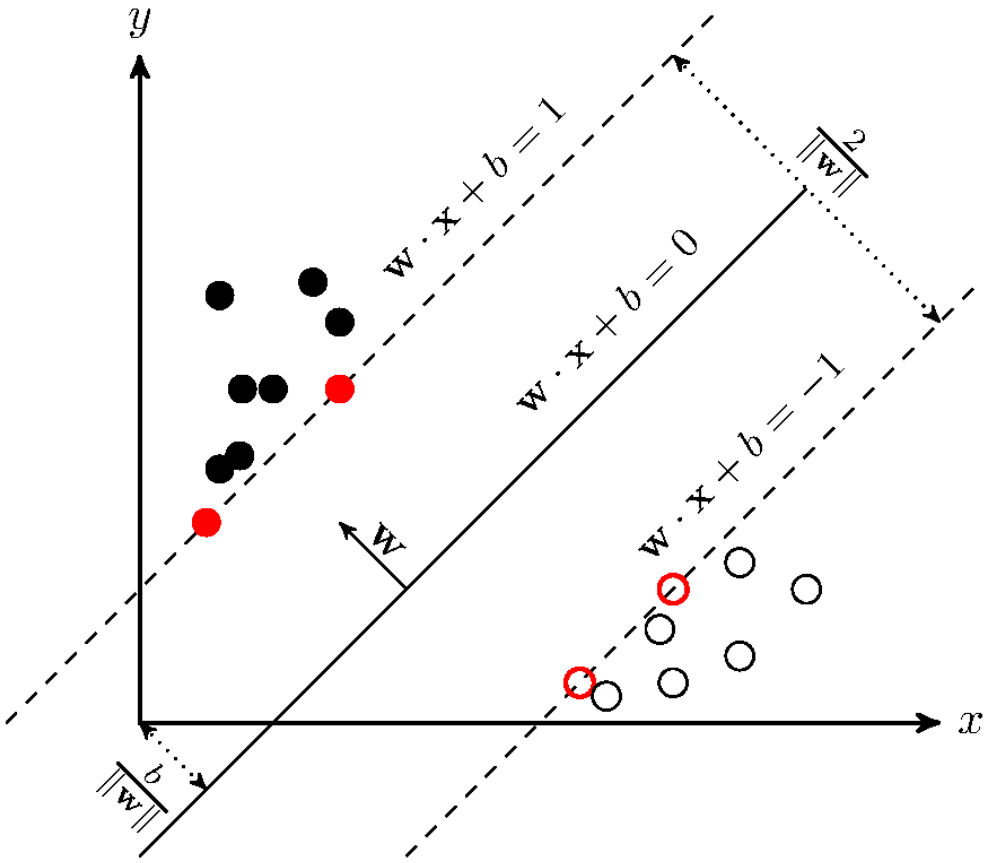
Support Vector Machine

- Красная прямая – оптимальная разделяющая гиперплоскость
- Векторы, помеченные красным – опорные векторы



Support Vector Machine

- $(x_1, y_1), \dots, (x_m, y_m), x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$
- $F(x) = sign(\langle w, x \rangle - w_0)$, w – нормаль к разделяющей гиперплоскости, w0 – вспомогательный параметр
- $$\begin{cases} \arg\min_{w, w_0} \frac{1}{2}\|w\|^2, \\ y_i(\langle w, x_i \rangle - w_0) \geq 1, i = 1, \dots, m \end{cases}$$



Support Vector Machine

Линейно неразделимая выборка:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \to \min\limits_{w,w_0,\xi} \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, m \\ \xi_i \geq 0, i = 1, \dots, m \end{cases}$$

$$\arg\min_{w,w_0} C\sum_{i=1}^{m}(1 - y_i(\langle w, x_i\rangle - w_0) + \frac{1}{2}\|w\|^2$$

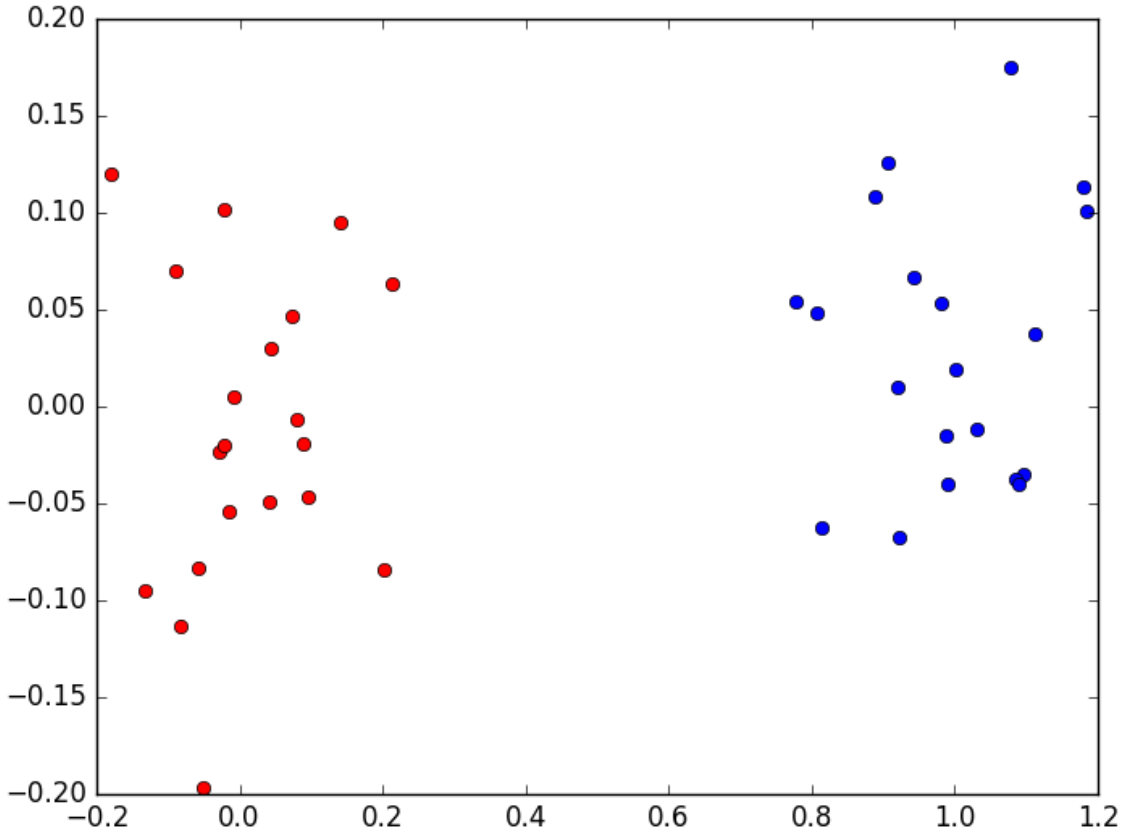
Метод ближайших центроидов

- $(x_1, y_1), \dots, (x_m, y_m), x_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$
- Вычислим центры каждого класса:

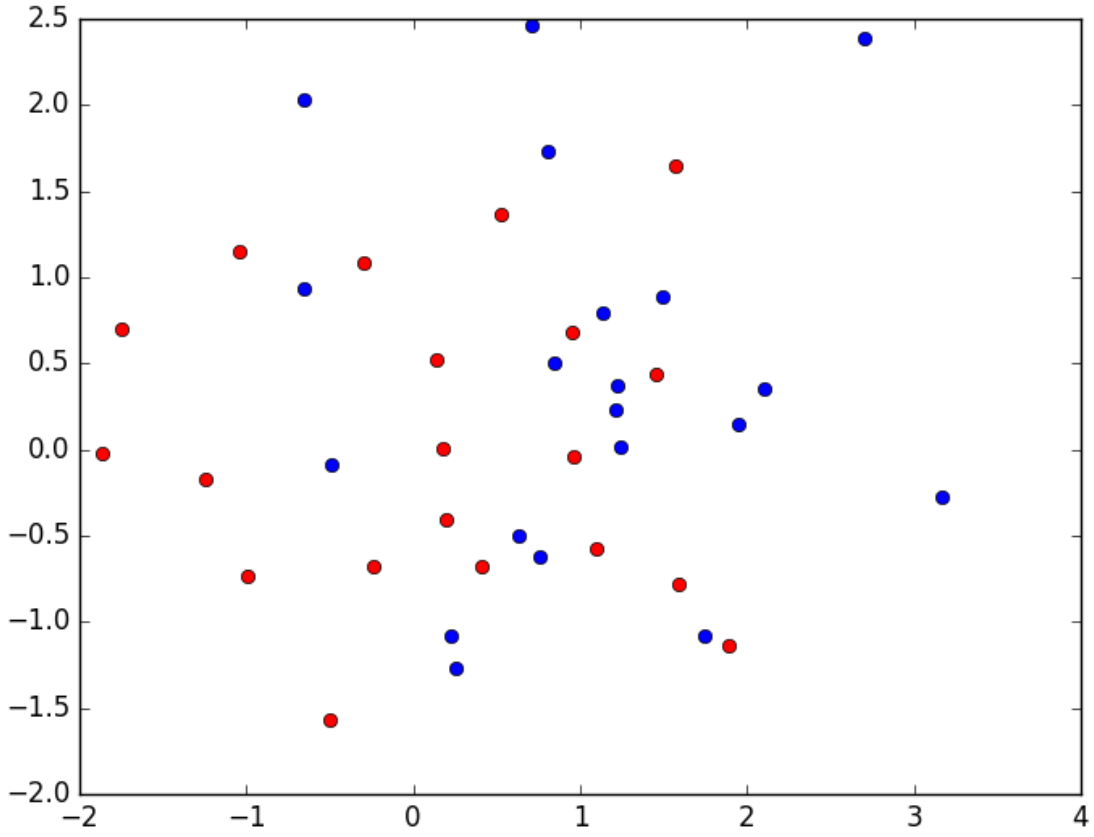
$\mu_l = \frac{1}{|C_l|} \sum_{i \in C_l} x_i,\ C_l$ – множество индексов объектов класса $l \in Y$

- $\hat{y} = \arg\min_{l \in Y} \|\mu_l - x\|$

Эксперименты на модельных данных (1)

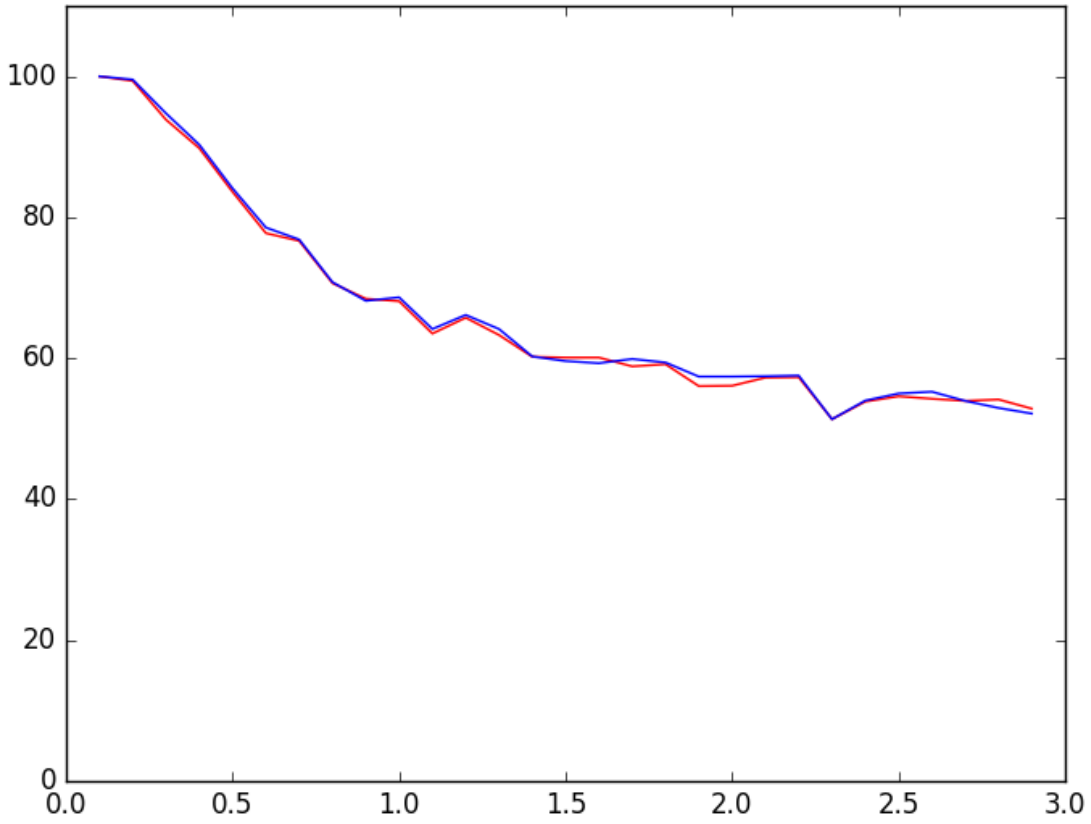


Матожидание 1 класса: (0, 0)
Матожидание 2 класса: (1, 0)
Среднеквадратическое отклонение обоих классов – 0,1



Матожидание 1 класса: (0, 0)
Матожидание 2 класса: (1, 0)
Среднеквадратическое отклонение обоих классов – 1,0

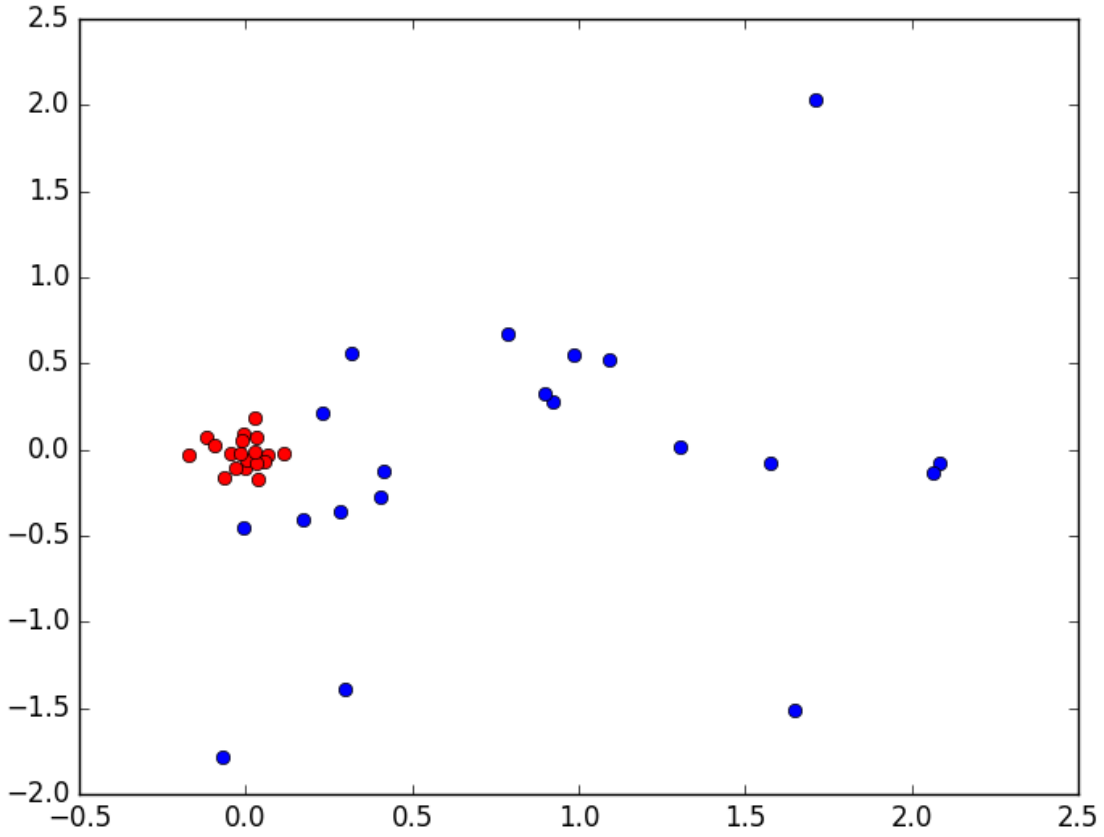
Эксперименты на модельных данных (1)



Качество классификации для двух методов при одинаковом значении среднеквадратического отклонения для обоих классов. Синяя кривая – метод ближайшего центроида, красная – SVM.

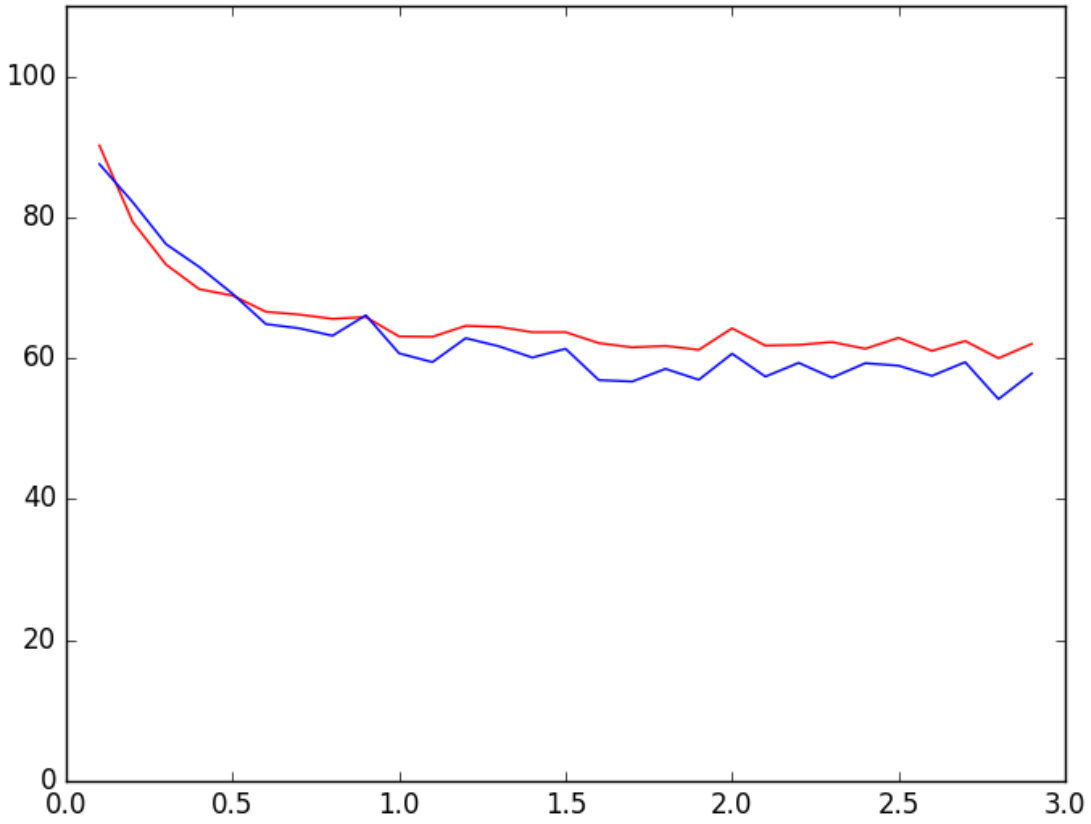
- Оба классификатора справляются со своей задачей одинаково хорошо.

Эксперименты на модельных данных (2)



Матожидание 1 класса: (0, 0)
Матожидание 2 класса: (1, 0)
Среднеквадратическое отклонение 1 класса – 0,1
Среднеквадратическое отклонение 2 класса – 0,7

Эксперименты на модельных данных (2)



Качество классификации для двух методов при разном значении среднеквадратического отклонения для двух классов. Синяя кривая – метод ближайшего центроида, красная – SVM.

- s = 0.1 – первый класс, s = 0.7 – второй класс
- SVM: 87.5%
- Метод ближайшего центроида: 85.5%
- Метод ближайшего центроида: 100% результат для первого класса (s = 0.1); 75% для второго класса (s = 0.7)

Эксперименты на реальных данных

2 класса: pancreas (43 образца) и ovary (44 образца)

Построение классификаторов по случайным парам генов:

- По 10 тыс. случайных пар строим классификаторы
- Оценка достоверности классификации KFold кросс-валидацией.
- SVM: 47%
- Метод ближайшего центроида: 57%.

Эксперименты на реальных данных

Построение классификаторов по «биологически правильным» генам:

$$s_{gene_i} = \frac{|E[pancreas] - E[ovary]|}{\sqrt{\sum_{i=1}^{n} \frac{(expr(i, pancreas) - E[pancreas])^2}{n}} + \sqrt{\sum_{i=1}^{m} \frac{(expr(i, ovary) - E[ovary])^2}{m}}}$$

- Значения s_{gene_i} расположим в порядке убывания
- Возьмем top-30 генов
- По всем парам, составленных из этих генов, построим классификаторы
- SVM: 88%
- Метод ближайшего центроида: 85%.

Ссылки

- Reis-Filho, J. S, Pusztai, L. (2011). «Gene expression profiling in breast cancer: classification, prognostication, and prediction». Lancet 378(9805): 1812-23
- Galatenko, V. V., Shkurnikov, M. Yu., Samatov, T. R., Galatenko, A. V., Mityakina, I. A., Kaprin, A. D., Schumacher, U. & Tonevitsky, A. G. (2015). «Highly informative marker sets consisting of genes with low individual degree of differential expression». Scientific Reports 5, Article number: 14967
- Manning, Christopher; Raghavan, Prabhakar; Schütze, Hinrich (2008). «Vector space classification». Introduction to Information Retrieval. Cambridge University Press.
- Dabney, A. R., (2005). «Classification of microarrays to nearest centroids». Bioinformatics 21(22): 4148-54
- Barretina, J., Caponigro, G., Stransky, N., et al. (2012). «The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity». Nature 483, 603-607