МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 2

Вероятностная постановка задачи обучения с учителем и некоторые методы

Некоторые обозначения

d	число входных признаков
N	длина обучающей выборки
$\mathscr{X}$	множество объектов
$\mathscr{Y}$	множество ответов (выходов)
$x^{(1)}, x^{(2)}, \ldots, x^{(N)}$	объекты обучающей выборки, $x^{(i)} \in \mathscr{X}$ $(i = 1, 2, \ldots, N)$
$y^{(1)}, y^{(2)}, \ldots, y^{(N)}$	выходы для объектов из обучающей выборки, $y^{(i)} \in \mathscr{Y}$
K	количество классов (в задачах классификации)
$\mathsf{Pr}\, A$	вероятность события A
$\mathsf{Pr}(A \mid B)$	вероятность события A при условии, что наступило событие B
$P_X(x)$	интегральная функция распределения: $P_X(x) = \mathsf{Pr}\,\{X \leq x\}$
$p_X(x)$	плотность вероятности непрерывной случайной величины X
$P(y \mid x)$	условная интегральная функция распределения
$p(y \mid x)$	условная плотность вероятности
$\mathsf{E}\, X$	математическое ожидание случайной величины X
$\mathsf{D}\, X$ или $\mathsf{Var}\, X$	дисперсия случайной величины X
σX	среднее квадратическое отклонение: $\sigma X = \sqrt{\mathsf{D}\, X}$
$\mathbf{\Sigma}(X)$	матрица ковариаций многомерной случайной величины X

2.1. Вероятностная постановка задачи

$\mathscr{X} = \mathbf{R}^d$ — множество объектов (входов) (точнее: множество их описаний)

$\mathscr{Y} = \mathbf{R}$ — множество ответов (выходов)

Будем рассматривать пары (x, y) как реализации $(d+1)$-мерной случайной величины (X, Y), заданной на вероятностном пространстве

$$(\mathscr{X} \times \mathscr{Y},\ \mathbf{A},\ \mathsf{Pr}), \qquad X \in \mathbf{R}^d,\ Y \in \mathbf{R}.$$

j-й признак — бинарный, номинальный, порядковый или количественный дискретный $\Leftrightarrow$ X_j — дискретная с. в.

j-й признак — количественный непрерывный $\Leftrightarrow$ X_j — непрерывная с. в.

Интегральный закон распределения $P_{X,Y}(x, y)$ не известен, однако известна *обучающая выборка*

$$\left\{(x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \ldots,\ (x^{(N)}, y^{(N)})\right\},$$

где $(x^{(i)}, y^{(i)})$ являются независимыми реализациями случайной величины (X, Y).

Требуется найти функцию $f :\ \mathscr{X} \to \mathscr{Y}$, которая по x предсказывает y, $f \in \mathscr{F}$

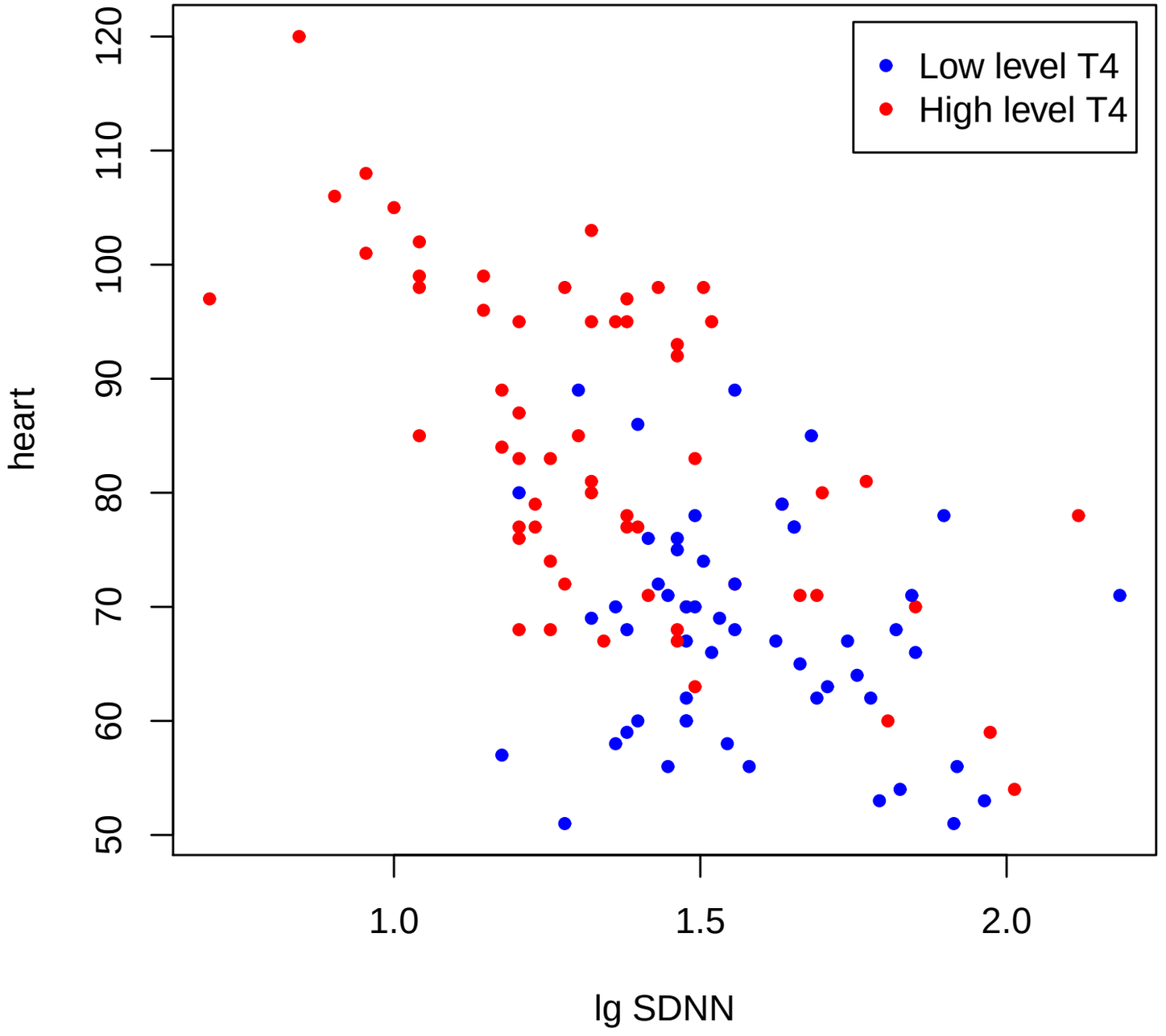
Пример

Имеются данные о 114 лицах с заболеванием щитовидной железы.
У 61 — повышенный уровень свободного гормона T4,
у 53 — уровень гормона в норме.

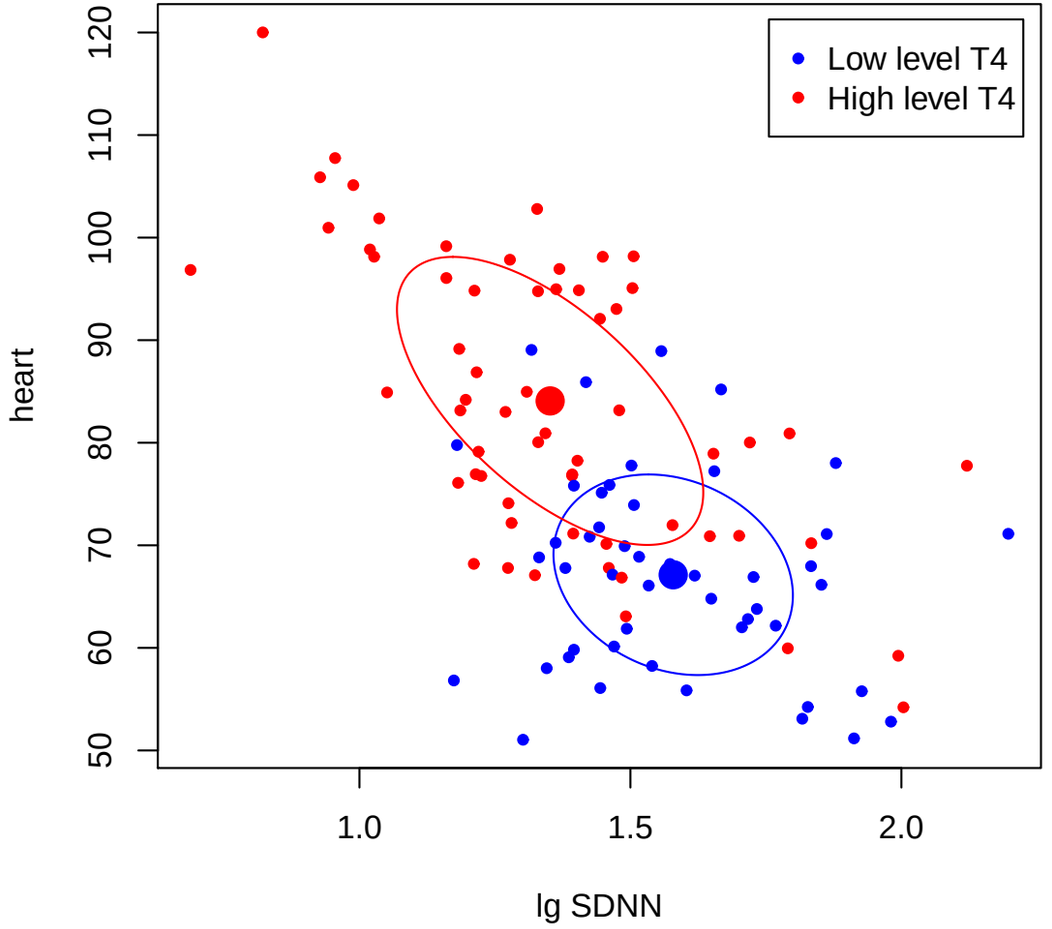
Для каждого пациента известны следующие показатели:

- `heart` — частота сердечных сокращений (пульс),
- `SDNN` — стандартное отклонение длительности интервалов между синусовыми сокращениями RR.

Можно ли научиться предсказывать (допуская небольшие ошибки) уровень свободного T4 по `heart` и `SDNN`?


Low level T4
High level T4
heart
120
110
100
90
80
70
60
50
1.0
1.5
2.0
lg SDNN

Многомерный тест Шапиро-Уилка проверки нормальности распределения
Гипотеза H_0: «X распределено нормально». Пусть $\alpha = 0.05$
Для синих точек (низкий уровень) $W = 0.9809$, p-value $= 0.5527 \Rightarrow$ принимаем
Для красных точек (высокий уровень) $W = 0.9542$, p-value $= 0.02306 \Rightarrow$ отвергаем
Для всей совокупности: $W = 0.9784$, p-value $= 0.06239 \Rightarrow$ принимаем



Пусть дана *функция потерь* (*штраф*) $L(y',\ y) = L(f(x),\ y)$.
x — вход, y — соответствующий выход, $y' = f(x)$ — предсказанное значение
Например, в задачах восстановления регрессии:

- *квадратичная ошибка* (*квадратичная функция потерь*): $L(y',\ y) = (y' - y)^2$;
- *абсолютная ошибка*: $L(y',\ y) = |y' - y|$;
- *относительная ошибка*: $L(y',\ y) = \dfrac{|y' - y|}{|y|}$.

В задачах классификации:

- *симметричная* 0–1 функция штрафа (*индикатор ошибки*):

$$L(y',\ y) = I(y' \neq y) \equiv \begin{cases} 0, & \text{если } y' = y, \\ 1, & \text{если } y' \neq y; \end{cases}$$

- В общем случае функция потерь полностью описывается $K \times K$ матрицей $L = (\ell_{y'y})$, где $\ell_{y'y} = L(y',\ y)$.

Пусть, например, в задаче медицинской диагностики $\mathscr{Y} = \{0, 1\}$, где
$y = 0$ — пациент здоров, $y = 1$ — пациент болен.
$L(1, 1) = L(0, 0) = 0$

- ошибка 1-го рода — ложная тревога — false positive error

  $L(1, 0) = 1$ — болезнь определена у здорового пациента

- ошибка 2-го рода — ложный пропуск — false negative error

  $L(0, 1) = 10$ — болезнь не определена у больного пациента (!)

Аналогично: автоматическое определение почтового спама, техническая диагностика, обнаружение комп. вирусов и т. д.

Мат. ожидание функции потерь

$$R(f) = \mathsf{E}\, L\big(f(X), Y\big) = \int\limits_{\mathscr{X}\times\mathscr{Y}} L\big(f(x), y\big)\, p(x, y)\, dx\, dy$$

— *средний риск*, *средняя ошибка* или *ожидаемая ошибка предсказания* (≈ *метрика качества*).

(Если $L(y', y) = I(y' \neq y)$, то $R(f)$ — вероятность ошибки)

$R(f)$ характеризует *качество*, или *обобщающую способность*, функции f.

Чем меньше $R(f)$, тем качество лучше.

Разумный подход: в качестве f взять функцию из заданного класса $\mathscr{F}$, минимизирующую средний риск $R(f)$ (*принцип минимизации среднего риска*)

НО: Закон $P(x, y)$ не известен и поэтому мы не можем точно вычислить $R(f)$.

- Можем восстановить $P(x, y)$ по выборке («классический» подход)
- Можем аппроксимировать $R(f)$, а затем минимизировать
- ...

(Кроме того, даже если $P(x, y)$ знаем (аппроксимировали), то возникает задача поиска минимума на множестве функций $\mathscr{F}$ — задача вариационного исчисления.)

2.1.1. Восстановление функции распределения вероятности $P(X, Y)$

Будем минимизировать средний риск при $f \in \mathscr{F}$

$$R(f) = \int\limits_{\mathscr{X}\times\mathscr{Y}} L\big(f(x), y\big)\, p(x, y)\, dx\, dy. \qquad (*)$$

1) по имеющейся выборке $\big\{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\big\}$ решается задача восстановления функции распределения $P(x, y)$.

2) восстановленная функция $\widehat{P}(x, y)$ подставляется в (*) вместо $P(x, y)$ и решается задача минимизации (точнее: вариационного исчисления, так как надо минимизировать на множестве функций $\mathscr{F}$).

- В качестве $\widehat{P}(x, y)$ можно взять эмпирическую функцию распределения.

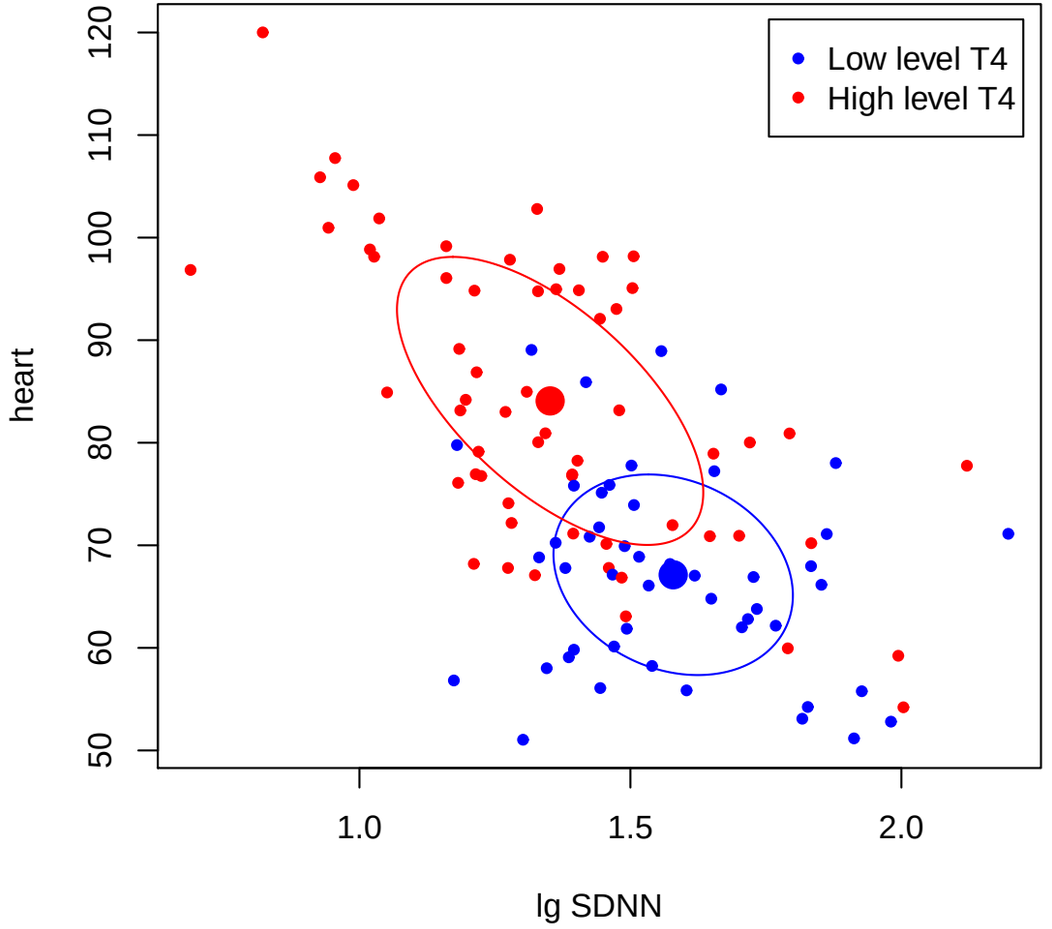
  Согласно теореме Гливенко с ростом N эмпирическая функция распределения равномерно приближается к истинной. Нужна очень большая выборка!

- Параметрические методы (например, дискриминантный анализ).

  Должны много знать о распределении. Выборка должна быть большой.

- Непараметрические методы (например, парзеновские окна).

Многомерный тест Шапиро-Уилка проверки нормальности распределения
Гипотеза H_0: «X распределено нормально». Пусть $\alpha = 0.05$
Для синих точек (низкий уровень) $W = 0.9809$, p-value $= 0.5527 \Rightarrow$ принимаем
Для красных точек (высокий уровень) $W = 0.9542$, p-value $= 0.02306 \Rightarrow$ отвергаем
Для всей совокупности: $W = 0.9784$, p-value $= 0.06239 \Rightarrow$ принимаем



2.1.2. Аппроксимация среднего риска $R(f)$

Элементы обучающей выборки $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ распределены случайно и независимо, каждый согласно закону распределения $P(X, Y)$, поэтому

$$R(f) \approx \widehat{R}(f) = \widehat{R}(f, x^{(1)}, y^{(1)}, \ldots, x^{(N)}, y^{(N)}) = \frac{1}{N} \sum_{i=1}^{N} L\big(f(x^{(i)}), y^{(i)}\big),$$

$\widehat{R}(f)$ — *эмпирический риск* (*эмпирическая ошибка*).

(Если $L(y', y) = I(y' \neq y)$, то $\widehat{R}(f)$ — доля ошибок на обучающей выборке)

Принцип минимизации эмпирического риска: минимизируем $\widehat{R}(f)$ на множестве $\mathscr{F}$.

В курсе будет рассмотрено много практических методов, реализующих этот принцип.

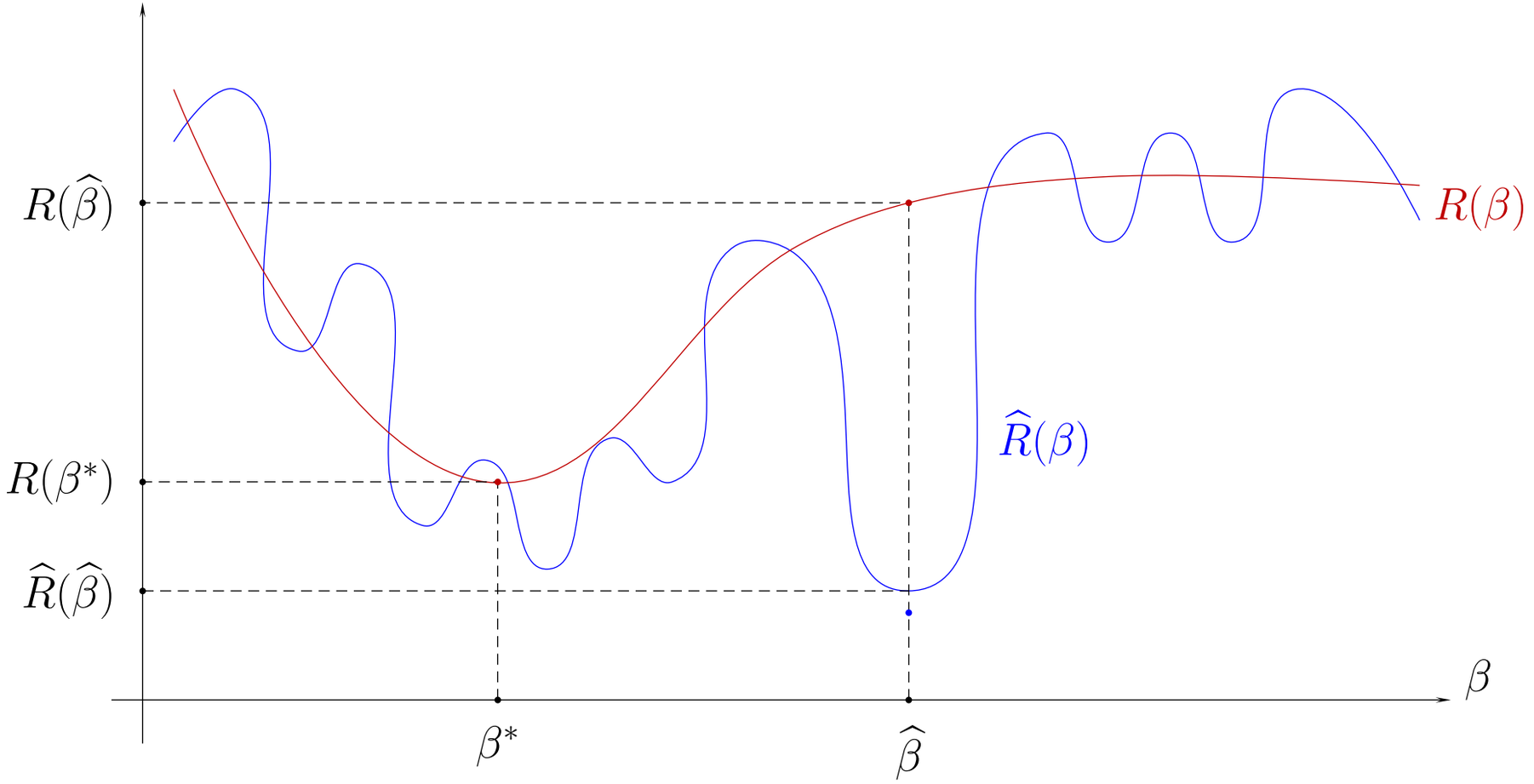
Недостатки: занижение величины риска и возможность переобучения.

Переобучение: ошибка на обучающей выборке много меньше $R(f)$ (и $\widehat{R}_{\text{test}}(f)$)

$$\widehat{R}(f) \equiv \widehat{R}_{\text{train}}(f) \equiv \frac{1}{N} \sum_{i=1}^{N} L\big(f(x^{(i)}), y^{(i)}\big) \ll R(f) \approx \widehat{R}_{\text{test}}(f) \equiv \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} L\big(f(x^{(i)}_{\text{test}}), y^{(i)}_{\text{test}}\big)$$

«Причина» переобучения при реализации метода минимизации эмпирического риска:

$$\mathscr{F} = \{f(\cdot, \beta) : \ \beta \in B\}$$



$$\widehat{R}(\widehat{\beta}) \ll R(\widehat{\beta}) \qquad \text{и} \qquad R(\beta^*) \ll R(\widehat{\beta})$$

2.2. Регрессионная функция

Рассматриваем задачу восстановления регрессии.

$$R(f) = \int\limits_{\mathscr{X}\times\mathscr{Y}} L\big(f(x), y\big)\, p(x, y)\, dx\, dy = \int\limits_{\mathscr{X}} \int\limits_{\mathscr{Y}} L\big(f(x), y\big)\, p(y \,|\, x)\, dy\, p(x)\, dx,$$

т. е.

$$R(f) = \int\limits_{\mathscr{X}} \mathsf{E}\left(L\big(f(x), Y\big) \,|\, x\right) p(x)\, dx$$

Пусть функция потерь — квадратичная:

$$R(f) = \int\limits_{\mathscr{X}} \int\limits_{\mathscr{Y}} \Big(f(x) - y\Big)^2 p(y\,|\,x)\,dy\,p(x)\,dx = \int\limits_{\mathscr{X}} \mathsf{E}\left(\big(f(x) - Y\big)^2\,|\,x\right) p(x)\,dx.$$

Очевидно, минимизировать $R(f)$ можно поточечно:

$$f^*(x) = \operatorname*{argmin}_c \mathsf{E}\left((c - Y)^2 \mid x\right), \tag{1}$$

откуда

$$f^*(x) = \mathsf{E}\,(Y\,|\,x). \tag{2}$$

Это так называемая *регрессионная функция*.

Итак, *в случае квадратичной функции потерь наилучшим предсказанием y в ответ на вход x является условное среднее* (*регрессионная функция*).

$R(f^*)$ назовем *байесовой ошибкой*, или *байесовым риском*, или *неустранимой ошибкой*.

Упражнение 2.1 Доказать, что из (1) следует (2), при этом $R(f^*) = \mathsf{E}_X \mathsf{D}_Y(Y \mid X)$.

Упражнение 2.2 Доказать, что если $L(y',\, y) = |y' - y|$, то минимум среднему риску доставляет условная медиана $f^*(x) = \operatorname{median}(Y\,|\,x)$. Чему равна при этом $R(f^*)$?

2.2.1. Метод ближайшего соседа

Возникает задача аппроксимации условного среднего $\mathsf{E}(Y\,|\,x)$ по имеющейся выборке.

1) Заменим $f^*(x)$ выборочным средним:

$$f(x) = \frac{1}{|I(x)|}\sum_{i\in I(x)} y^{(i)}, \quad \text{где} \quad I(x) = \left\{i:\ x^{(i)} = x\right\},$$

Как правило, такое решение к успеху не приводит, так как обычно x встречается в обучающей выборке не более одного раза.

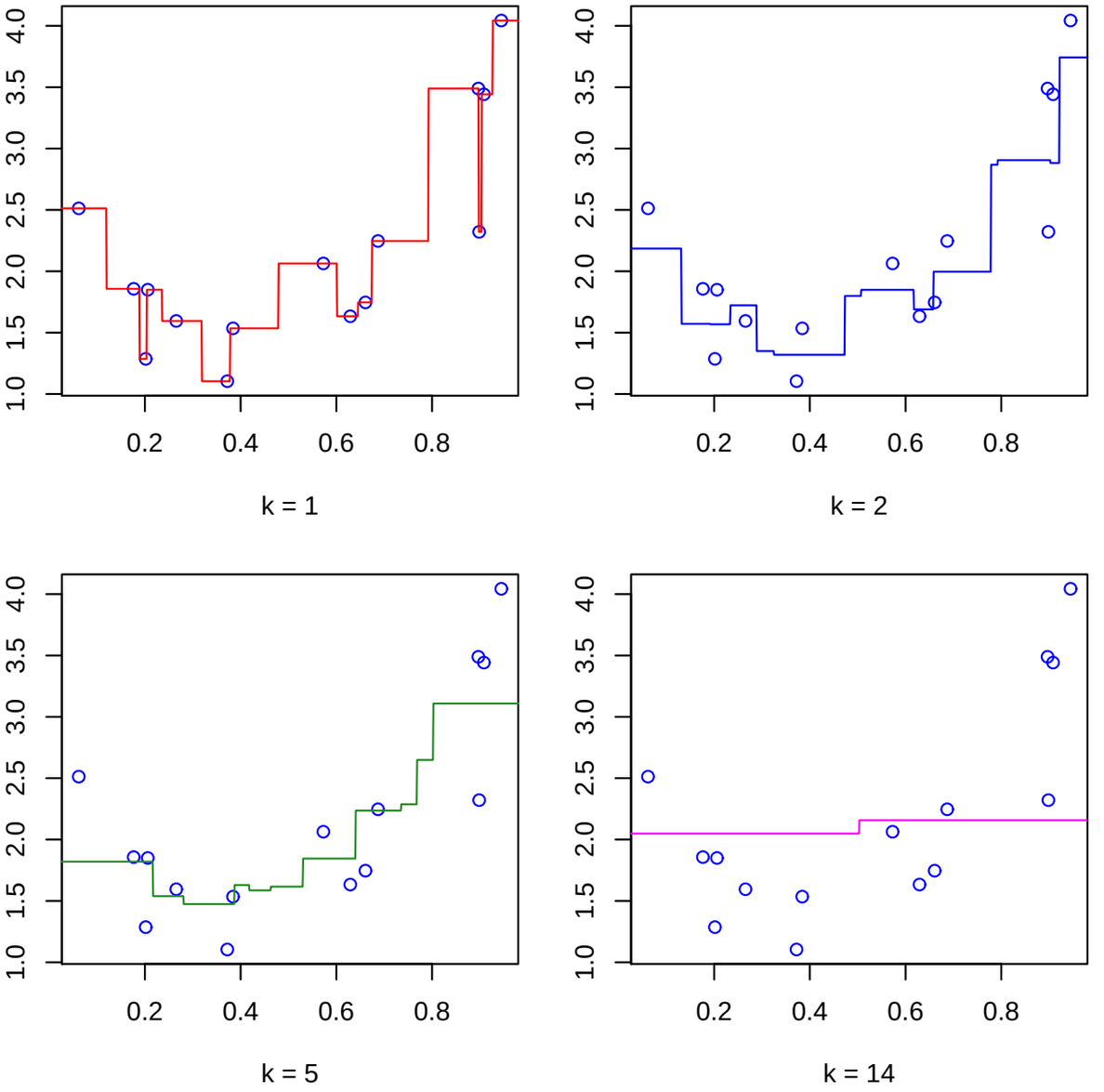
2) В *методе k ближайших соседей* (kNN — k nearest neighbours) вместо выборочного среднего берут

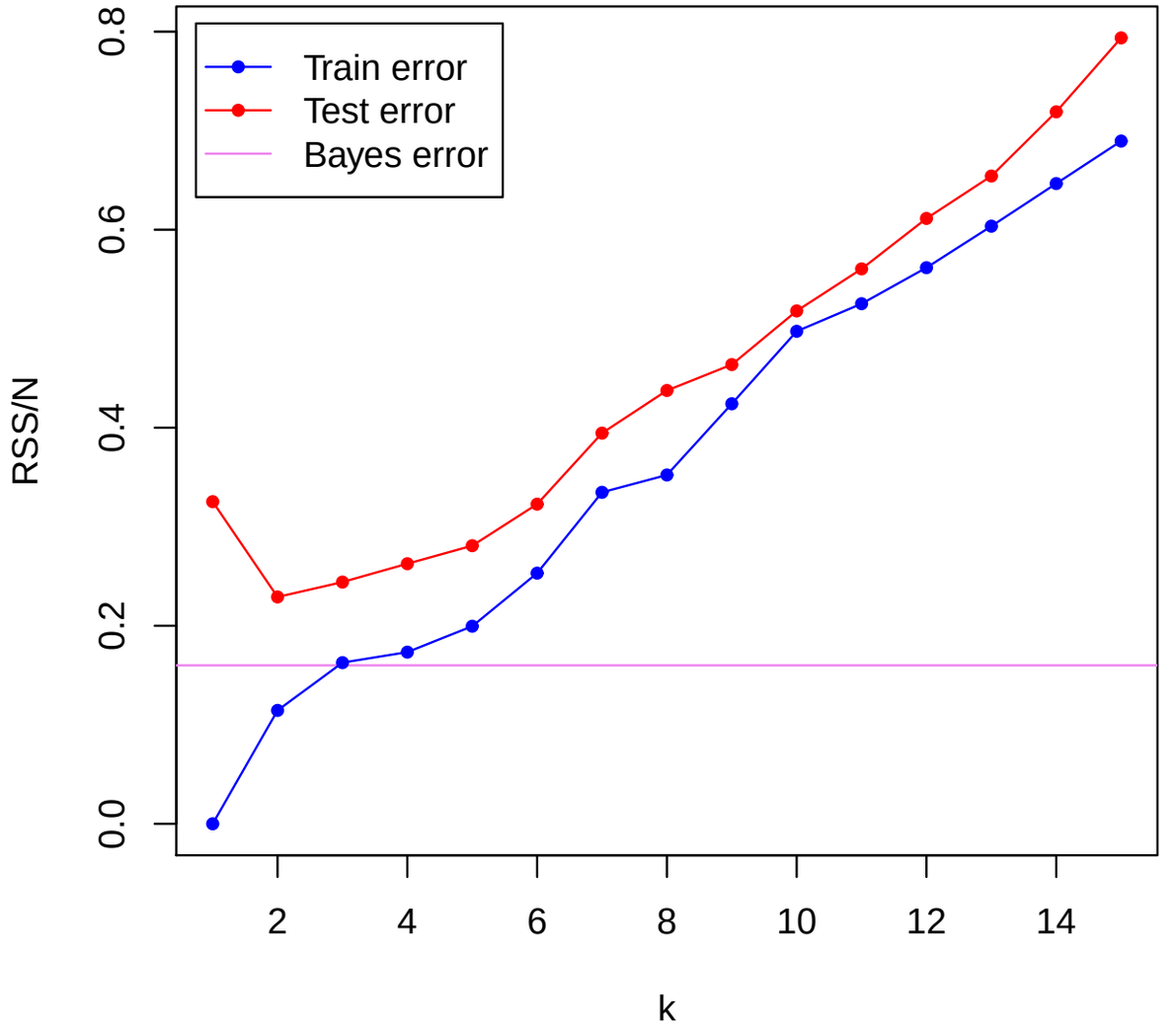
$$f(x) = \frac{1}{k}\sum_{x^{(i)}\in N_k(x)} y^{(i)},$$

где через $N_k(x)$ обозначено множество из k точек обучающей выборки, ближайших (например, по евклидову расстоянию) к x.

Частным случаем является *метод (одного) ближайшего соседа*, в котором $f(x) = y^{(i)}$, где $x^{(i)}$ — ближайшая к x точка из обучающей выборки.

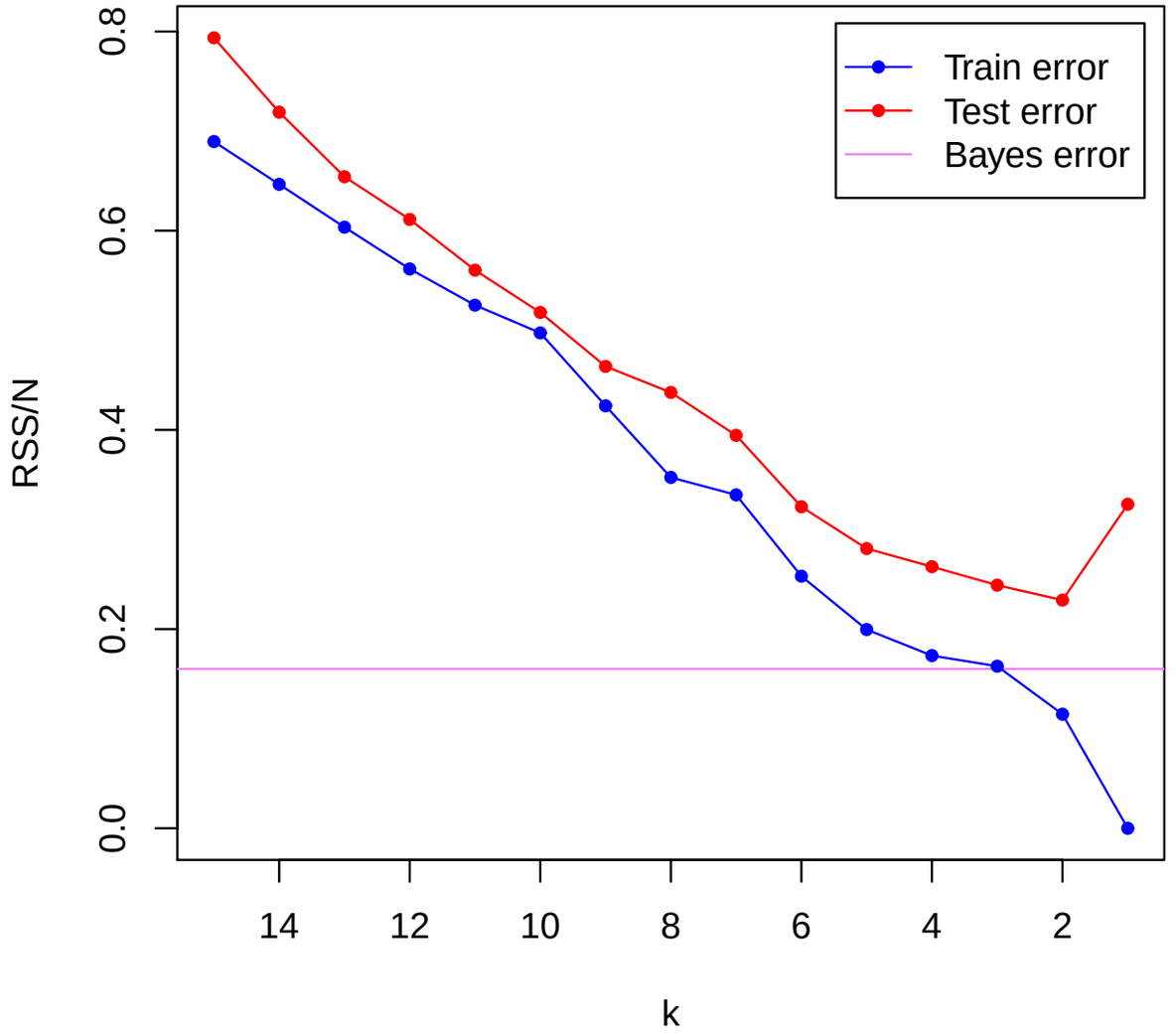
$$y = 8x^2 - 6.4x + 2.5 + N(0, 0.4)$$

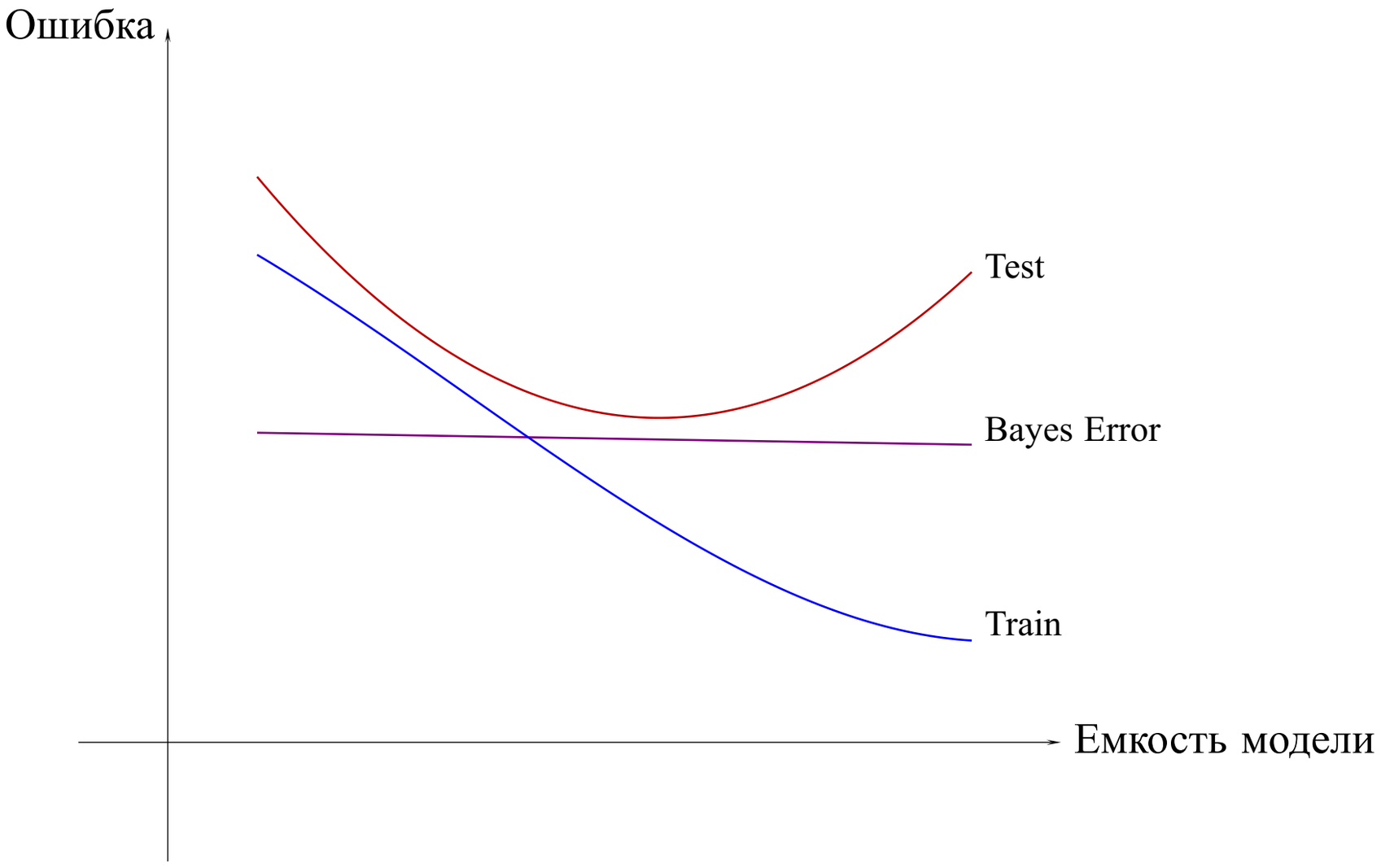




$$R(f^*) = \mathsf{E}_X \mathsf{D}_Y(Y \mid X) = \sigma^2 = 0.16$$

С увеличением k «емкость» («сложность») модели падает,
поэтому развернем горизонтальную ось в обратном направлении
(движение по ней вправо соответствует росту «емкости» модели)




Ошибка
Test
Bayes Error
Train
Емкость модели

Итак, метод ближайших соседей похож на метод восстановления функции распределения вероятности, только теперь аппроксимируется не плотность вероятности, а условное среднее.

2.3. Байесов классификатор

Рассмотрим задачу классификации. $\mathscr{Y} = \{1, 2, \ldots, K\}$.

Минимизируем средний риск

$$R(f) = \int\limits_{\mathscr{X}} \left(\sum_{y=1}^{K} L\left(f(x),\, y\right) \cdot \mathsf{Pr}\left(y \,|\, x\right) \right) p(x)\, dx. \qquad (**)$$

Пусть функция — индикатор ошибки (симметричный 0–1 штраф): $L(y',\, y) = I(y' \neq y)$.

Подынтегральная функция в (**) есть вероятность ошибки (при заданном x),

$$R(f) = \int\limits_{\mathscr{X}} \Big(1 - \mathsf{Pr}\left(Y = f(x) \,|\, x\right)\Big) p(x)\, dx,$$

откуда находим $f^*(x) = \operatorname{argmin} R(f)$:

$$f^*(x) = \operatorname*{argmin}_{y \in \mathscr{Y}} \left(1 - \mathsf{Pr}\left(y \,|\, x\right)\right) = \operatorname*{argmax}_{y \in \mathscr{Y}} \mathsf{Pr}\left(y \,|\, x\right).$$

$$f^*(x) = \underset{y\in\mathscr{Y}}{\operatorname{argmax}} \mathsf{Pr}(y\,|\,x) = \underset{y\in\mathscr{Y}}{\operatorname{argmax}} \frac{p(x\,|\,y)\,\mathsf{Pr}(y)}{p(x)} = \underset{y\in\mathscr{Y}}{\operatorname{argmax}}\, p(x\,|\,y)\,\mathsf{Pr}(y). \qquad (+)$$

Функция $f^*(x)$ называется *байесовым классификатором*.

Средний риск $R(f^*)$ байесова классификатора называется *байесовой ошибкой*, или *байесовым риском*, или *неустранимой ошибкой*.

Байесов классификатор играет в задаче классификации роль, аналогичную той, которую играет регрессионная функция в задаче восстановления регрессии.

$\mathsf{Pr}(y)$ — *априорная вероятность* появления объекта из класса y.

$\mathsf{Pr}(y\,|\,x)$ — *апостериорная вероятность* появления объекта из класса y.

Правило (+) называется *принципом максимума апостериорной вероятности*.

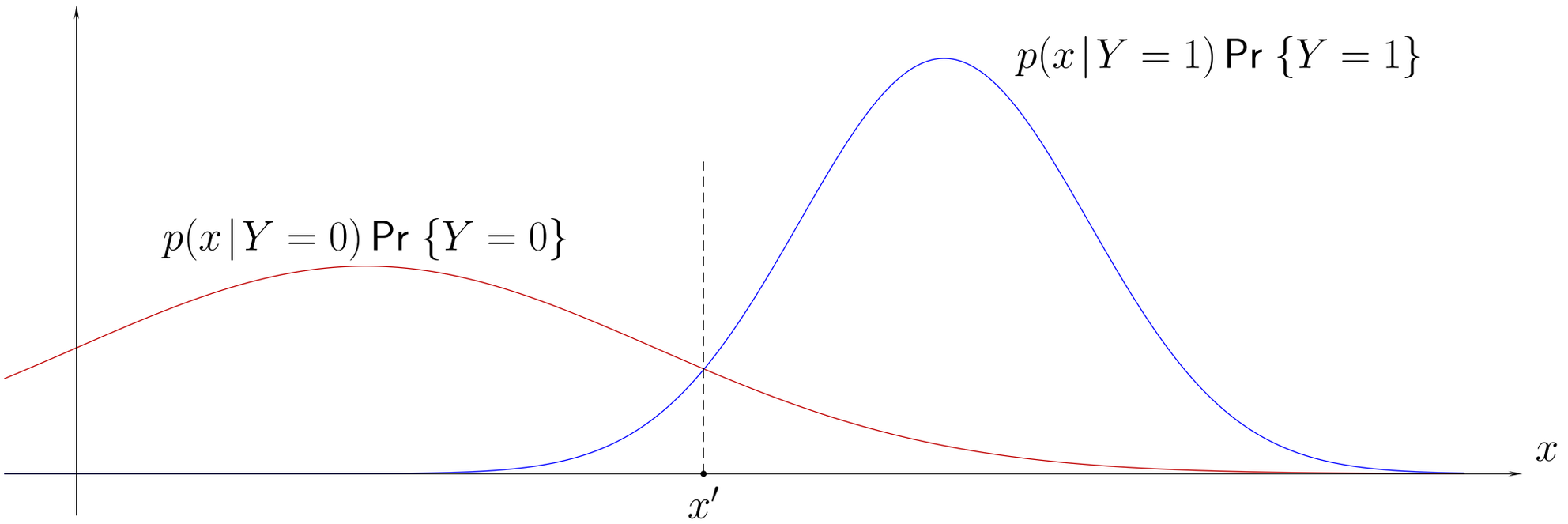
Если классы равновероятны, т. е. $\mathsf{Pr}(y) = 1/K$, то

$$\mathsf{Pr}(y\,|\,x) = \frac{p(x\,|\,y)\,\mathsf{Pr}(y)}{p(x)} = \frac{p(x\,|\,y)}{Kp(x)}$$

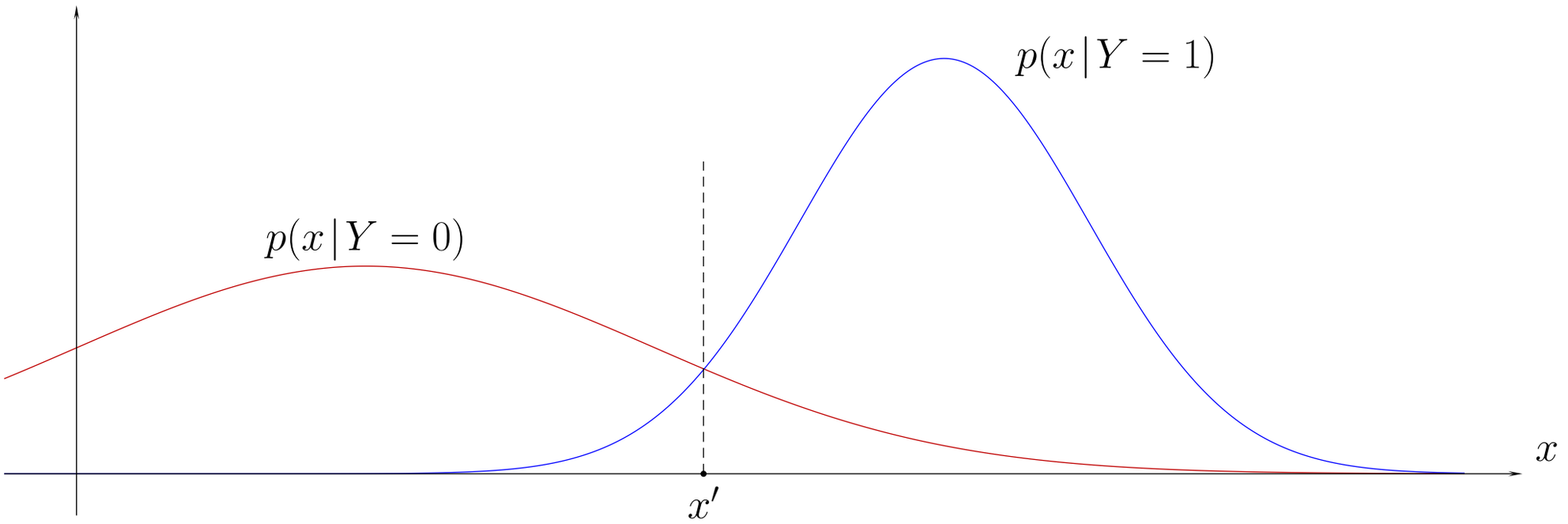
$$f^*(x) = \underset{y}{\operatorname{argmax}}\, p(x\,|\,y). \qquad (++)$$

Плотность $p(x\,|\,y)$ — *правдоподобие* (*likelihood*).
Правило (++) — *метод максимального правдоподобия* (*maximum-likelihood method*).

Принцип максимума апостериорной вероятности. При $x < x'$ полагаем $f(x) = 0$, иначе $f(x) = 1$.



Принцип максимального правдоподобия. При $x < x'$ имеем $f(x) = 0$, иначе $f(x) = 1$.



Чтобы построить байесов классификатор, мы должны знать (или оценить) $\mathsf{Pr}(y \,|\, x)$

Упражнение 2.3 Пусть в задаче классификации с двумя классами $\{0, 1\}$ используется функция потерь $L(y', y)$, такая, что $L(0, 0) = L(1, 1) = 0$, $L(1, 0) = \ell_1$, $L(0, 1) = \ell_0$. Докажите, что в этом случае байесов классификатор $f^*(x)$ удовлетворяет условию

$$f(x) = \operatorname*{argmax}_{y \in \{0,1\}} \ell_y \, \mathsf{Pr}(y \,|\, x).$$

Упражнение 2.4 Выразить байесов классификатор $f^*(x)$ для задачи классификации с K классами, если функция потерь равна $L(y',\, y) = \ell_{y'y}$ $(y', y = 1, 2, \ldots, K)$.

Пример

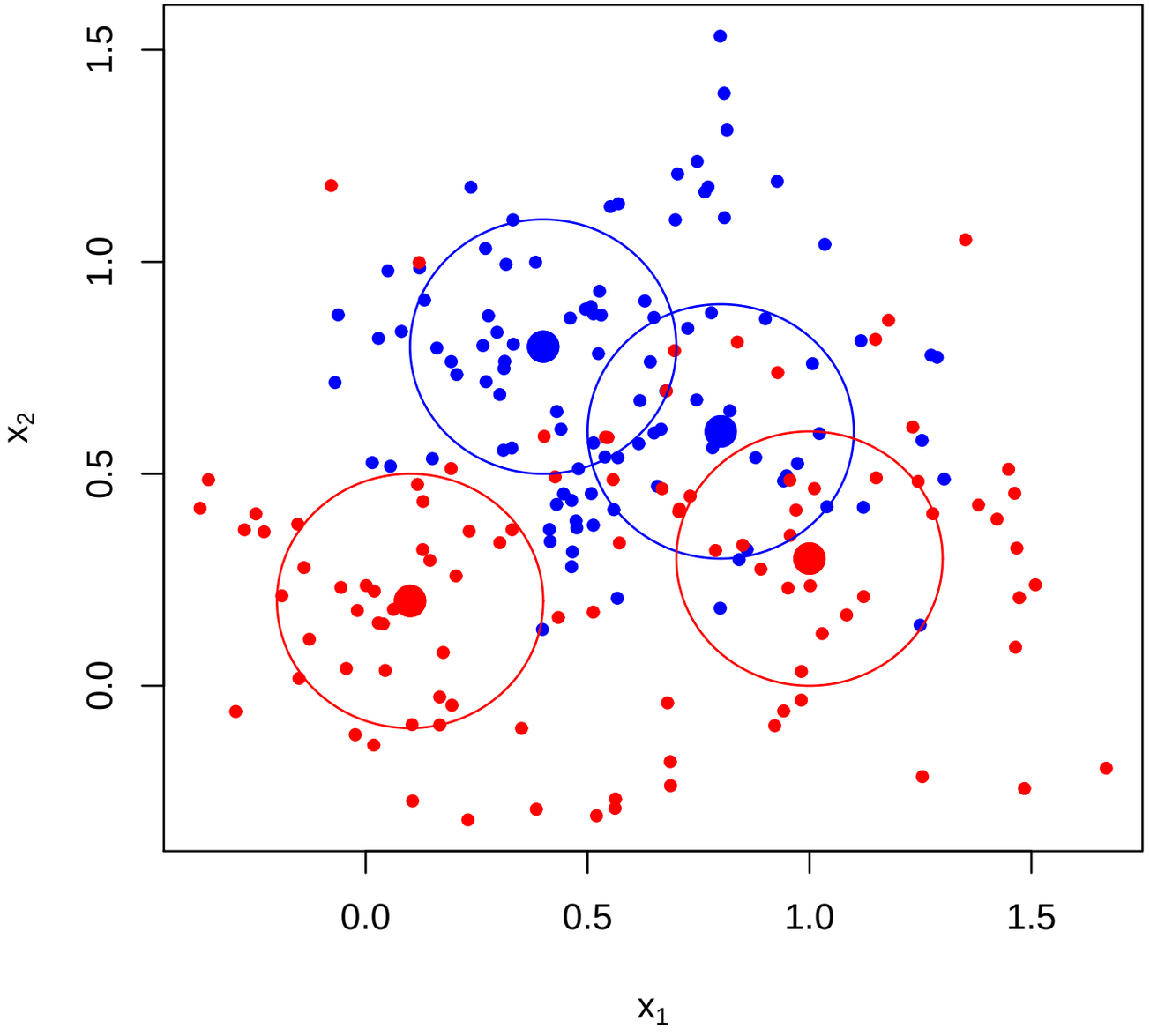
Рассмотрим задачу бинарной классификации.

Каждый класс в обучающей выборке — 100 прецедентов.

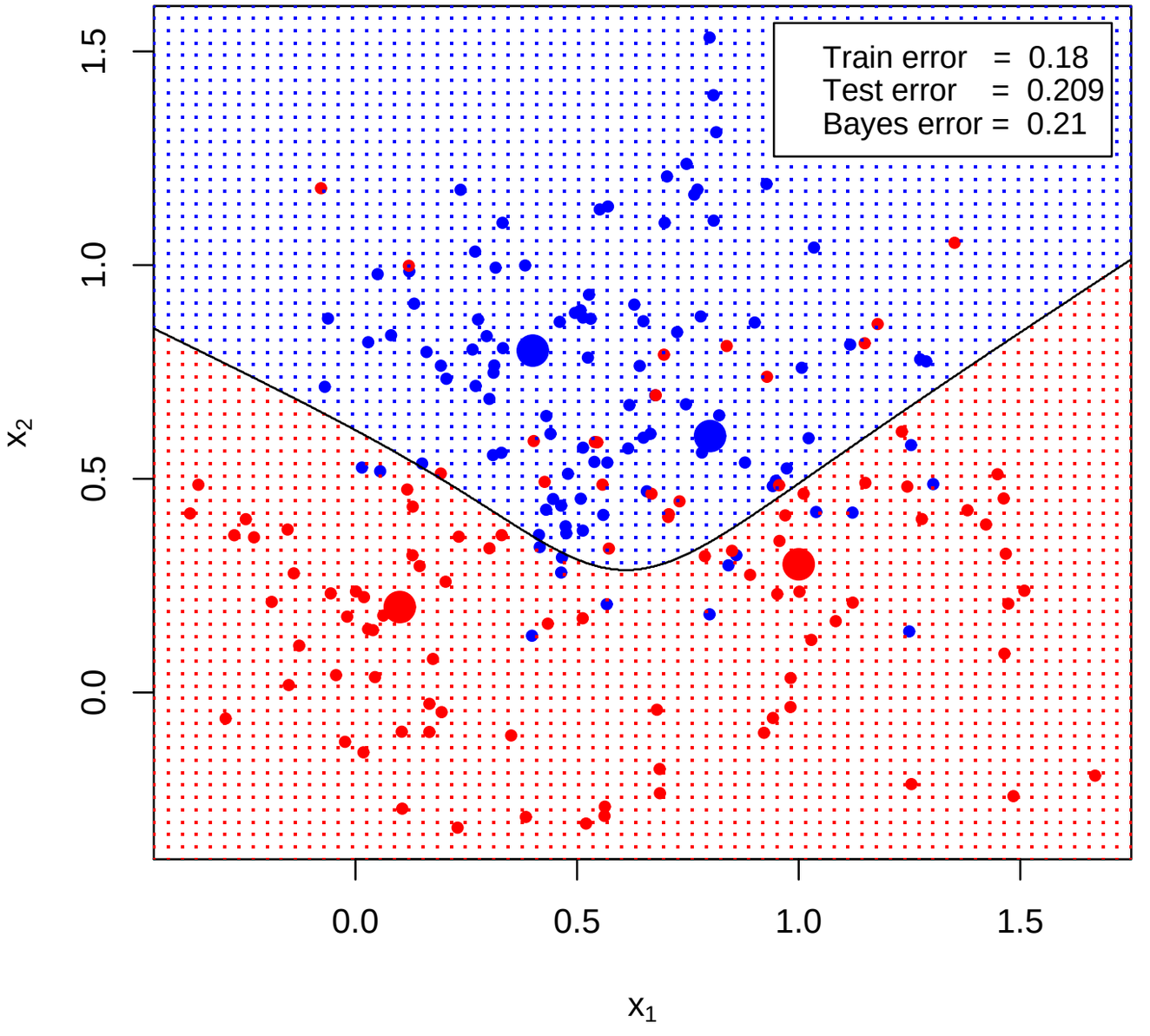
Распределение в каждом классе — смесь (взвешенная сумма) 2-х нормальных распределений (гауссианов).

$\mu_1 = (0.4, 0.8)$, $\mu_2 = (0.8, 0.6)$, $\mu_3 = (1.0, 0.3)$, $\mu_4 = (0.1, 0.2)$.

Матрица ковариации $\sigma^2\mathbf{I}$, где $\sigma = 0.3$.


1.5
1.0
0.5
0.0
x_2
0.0
0.5
1.0
1.5
x_1

Распределение известно, поэтому байесову ошибку можем найти точно.



Таким образом, байесов классификатор — это оптимальный классификатор.

Предполагается, что условные вероятности $\mathsf{Pr}(y|x)$ известны.

Как это можно использовать на практике?

Будем аппроксимировать $\mathsf{Pr}(y|x)$

1) Метод ближайших соседей (для задачи классификации)
2) Восстановление условной плотности вероятности

2.3.1. Метод ближайших соседей для задачи классификации

Будем, как и в задаче восстановления регрессии, для аппроксимации $\mathsf{Pr}(y|x)$ использовать k ближайших (по некоторому, например, евклидову расстоянию) объектов из обучающей выборки. Получаем метод k ближайших соседей для задачи классификации.

Пусть $N_k(x)$ — множество из k ближайших к x (по евклидову расстоянию) точек из обучающей выборки, $I_k(x,y)$ — множество тех точек $x^{(i)}$ из $N_k(x)$, для которых $y^{(i)} = y$.

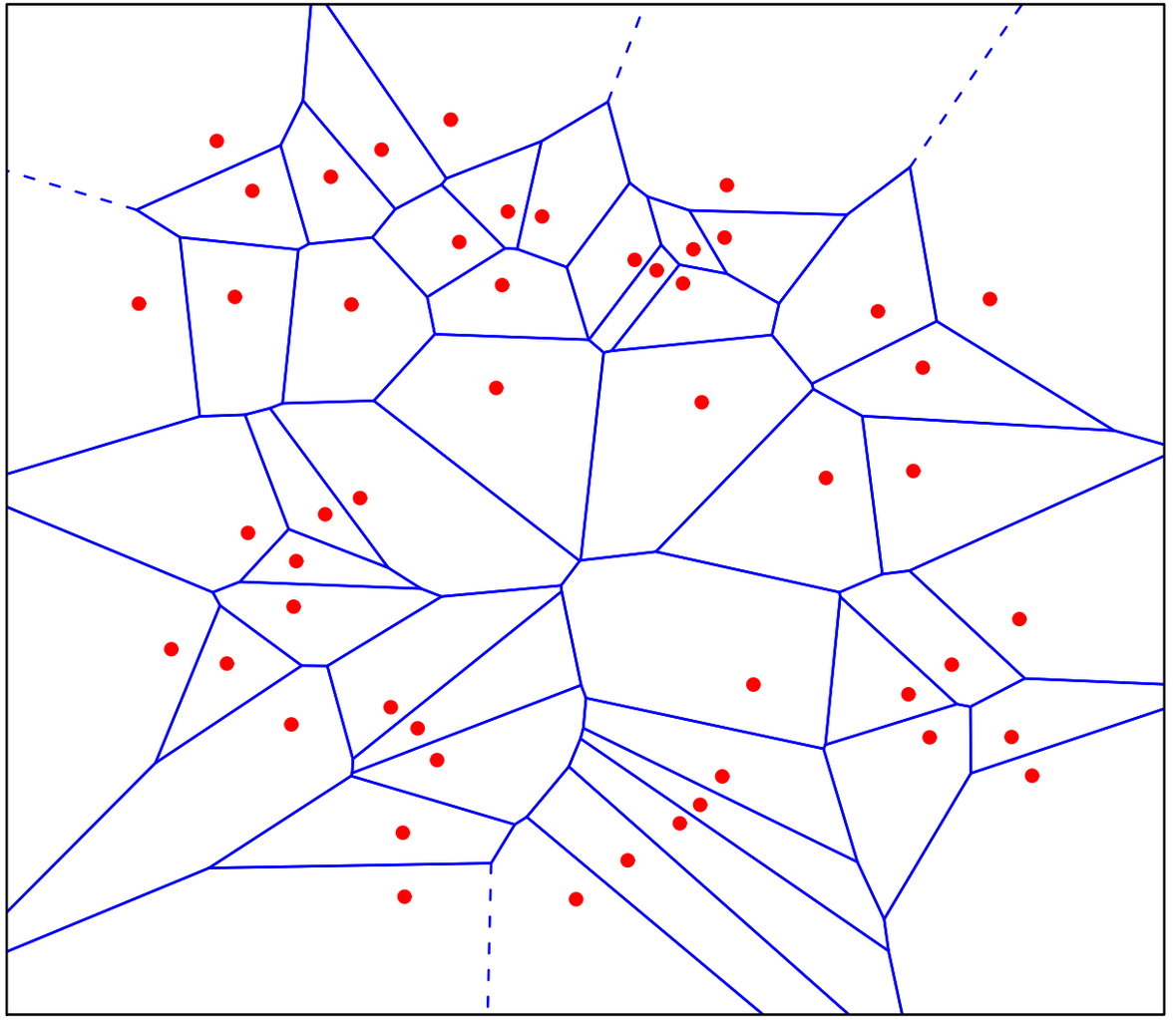
Согласно *методу k ближайших соседей* (kNN — k nearest neighbours) в качестве $f(x)$ берем результат голосования по всем точка из $I_k(x,y)$:

$$f(x) = \operatorname*{argmax}_y |I_k(x,y)|,$$

Частным случаем является *метод* (*одного*) *ближайшего соседа*, в котором $f(x) = y^{(i)}$, где $x^{(i)}$ — ближайший к x объект из обучающей выборки.

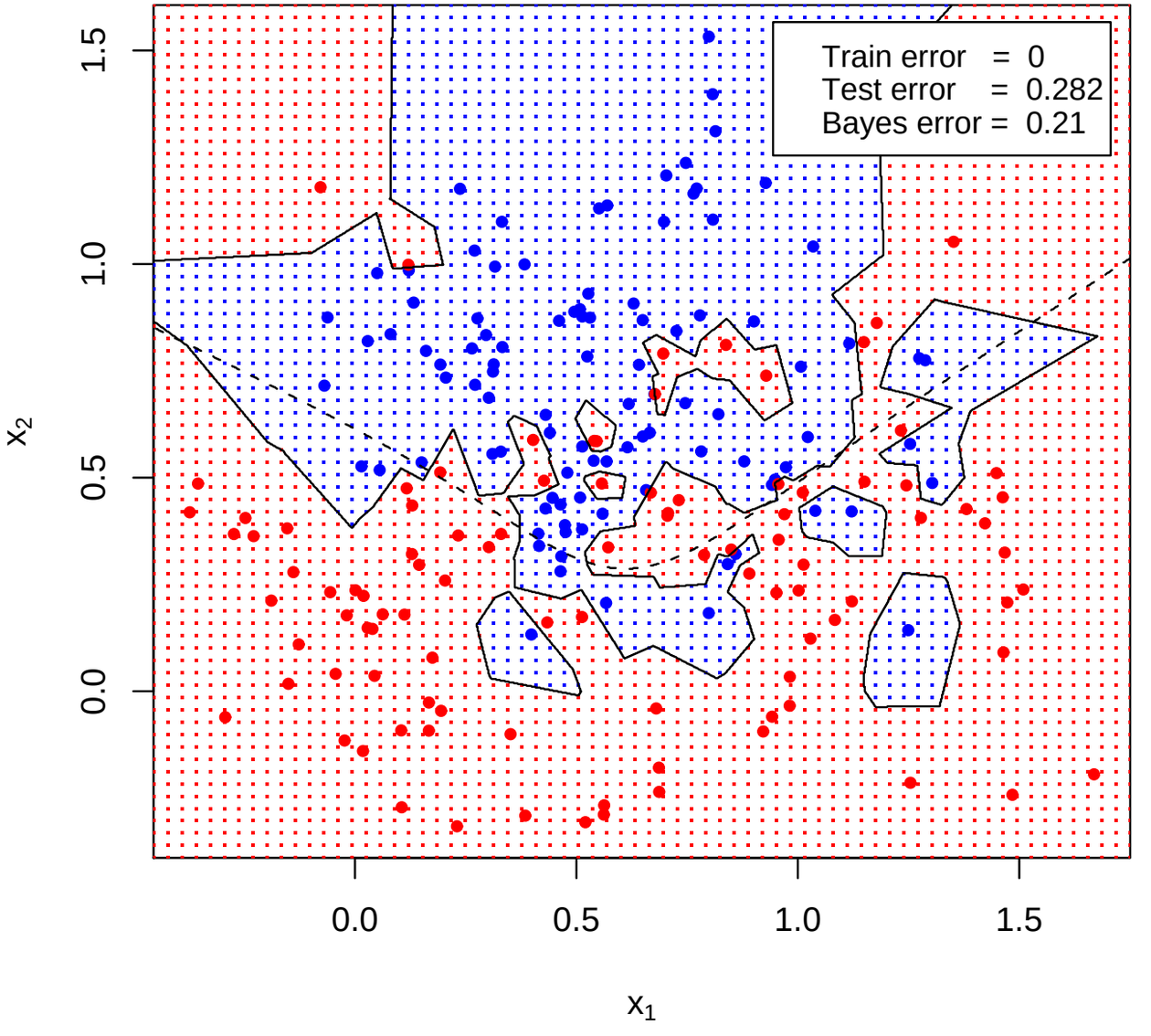
В этом случае D_y представляют собой *области Вороного*

Диаграмма Вороного для набора из 50 точек. Штриховыми линиями отмечены неограниченные участки границы

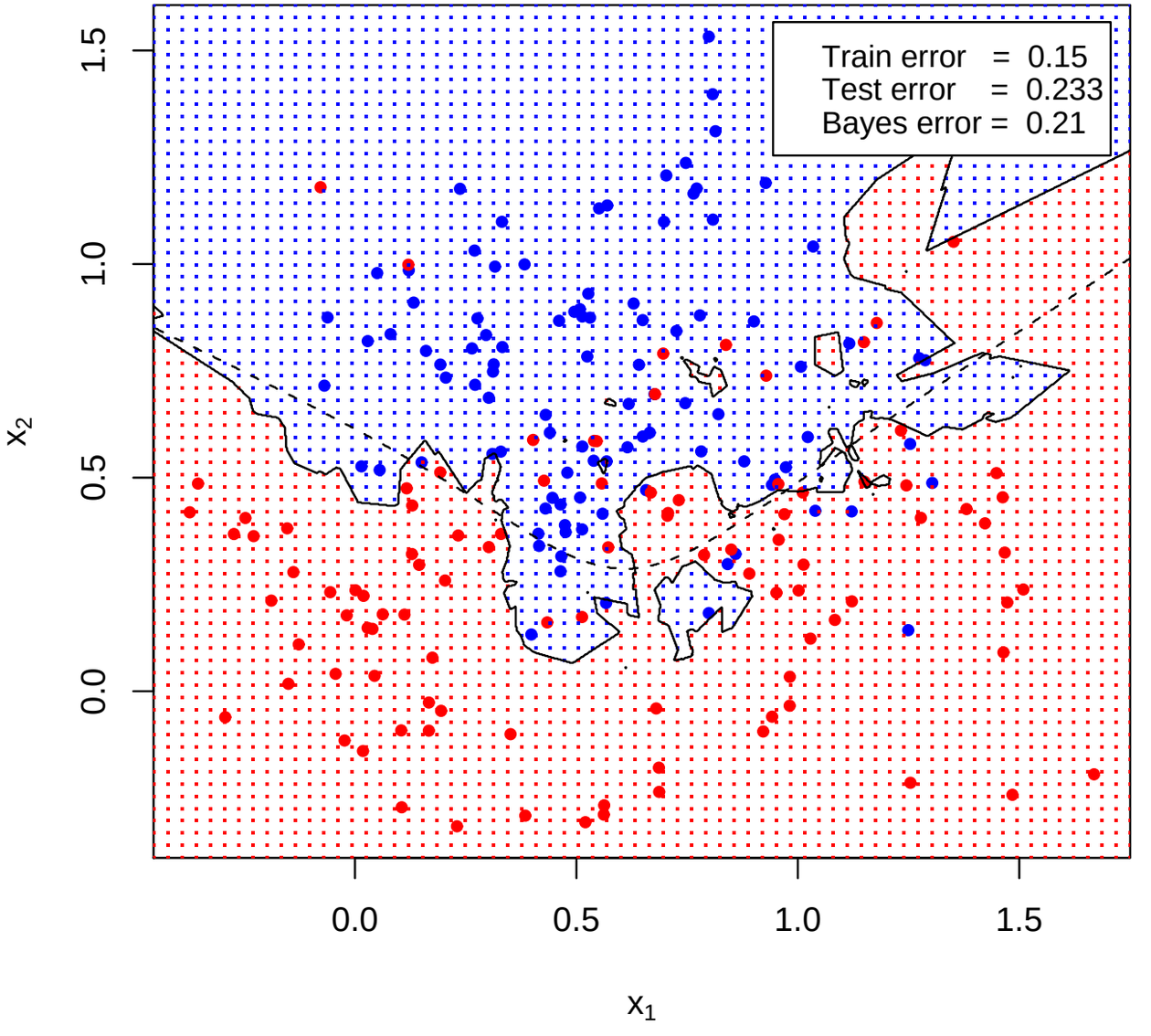


Все сильно зависит от масштаба!

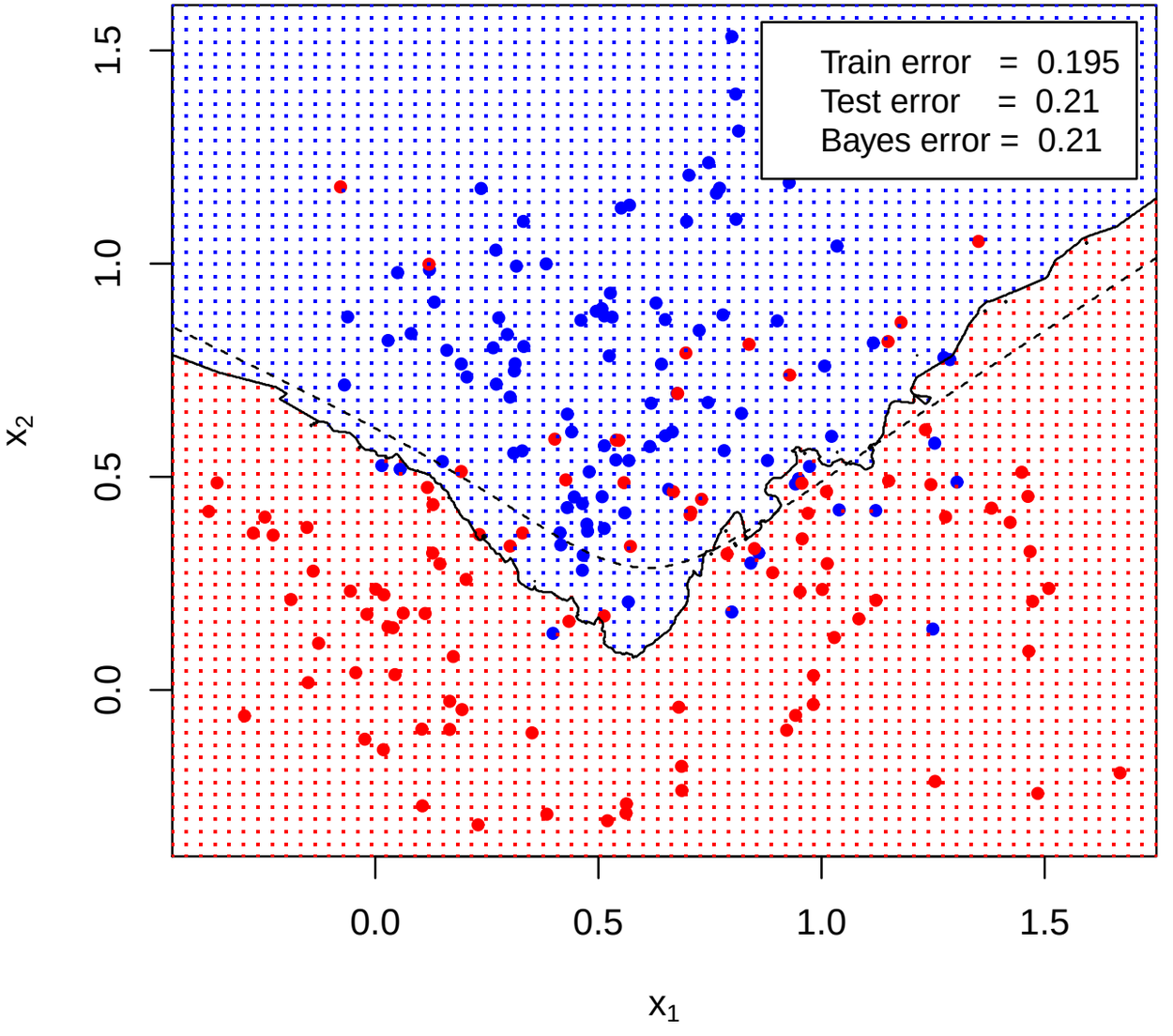
Метод одного ближайшего соседа



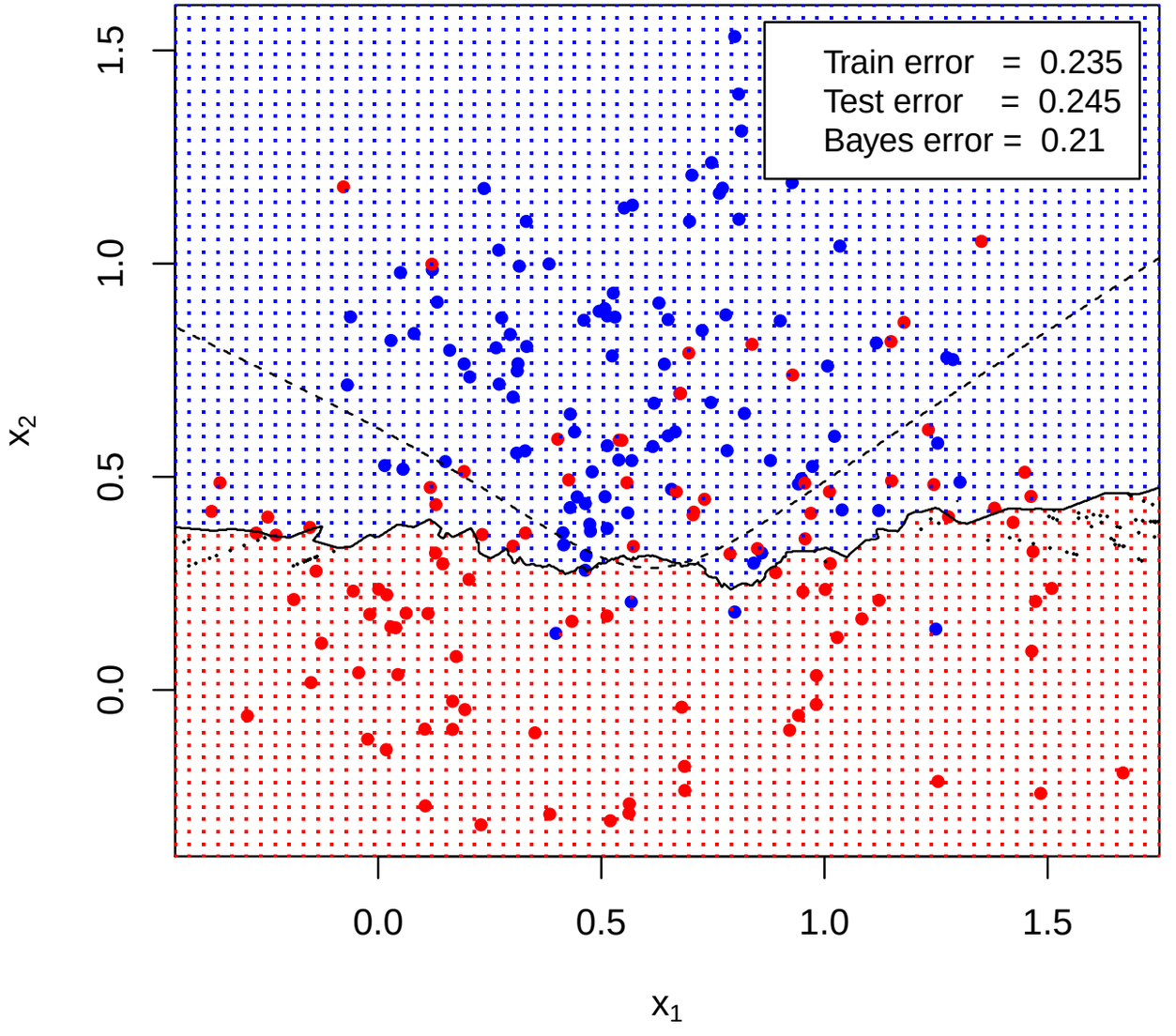
Метод 5 ближайших соседей



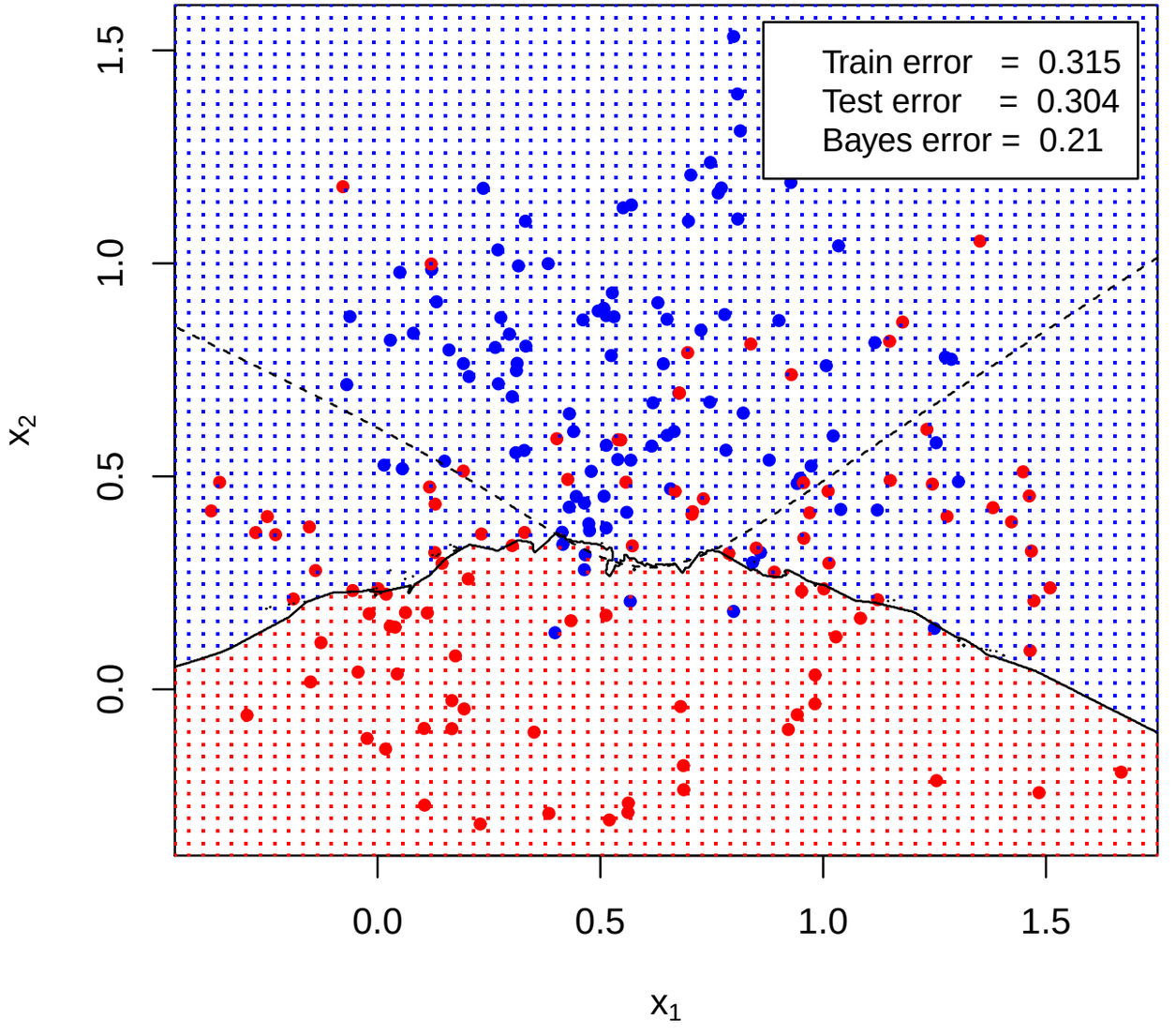
Метод 31 ближайшего соседа

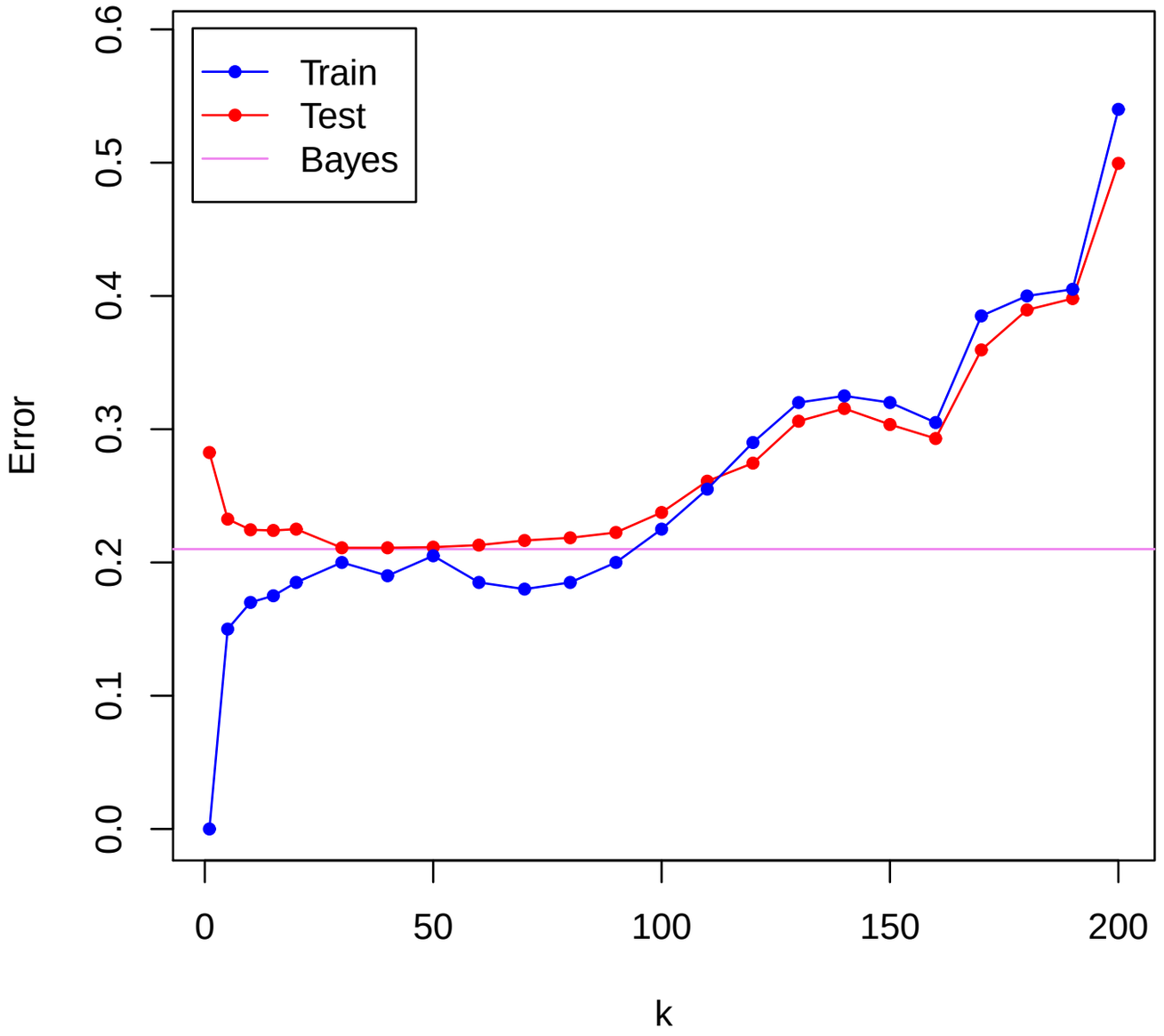


101 ближайший соседа

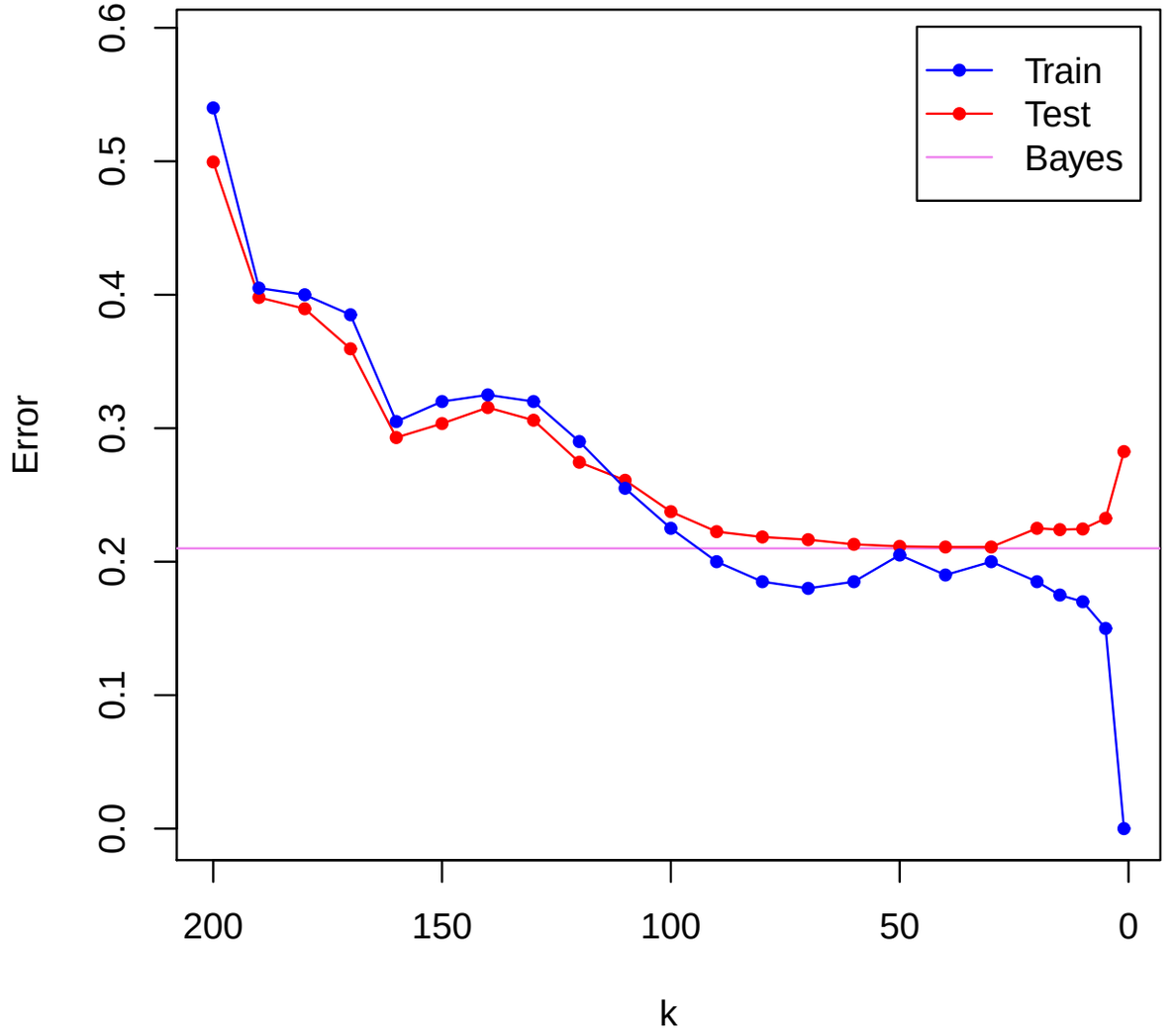


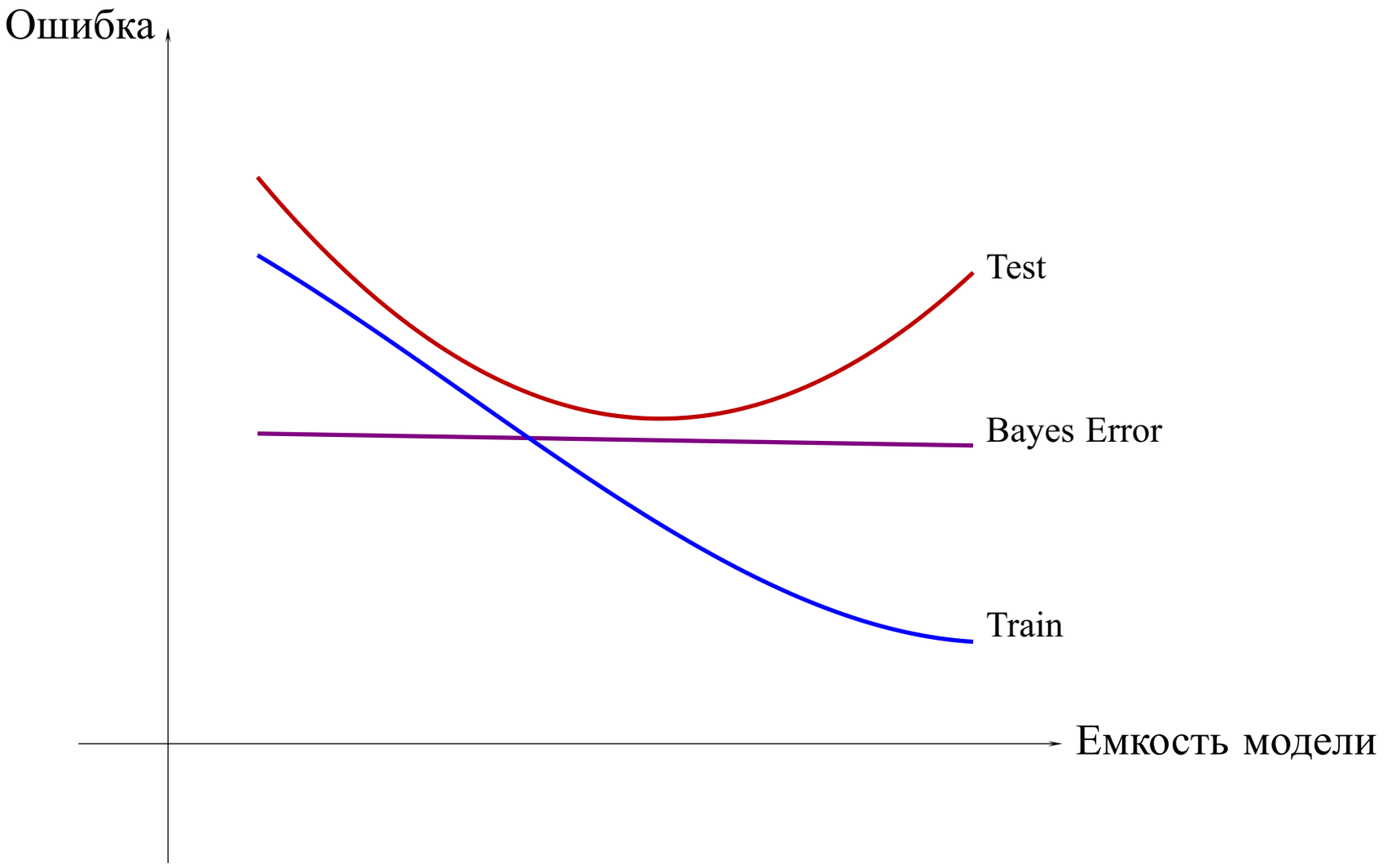
151 ближайший соседа



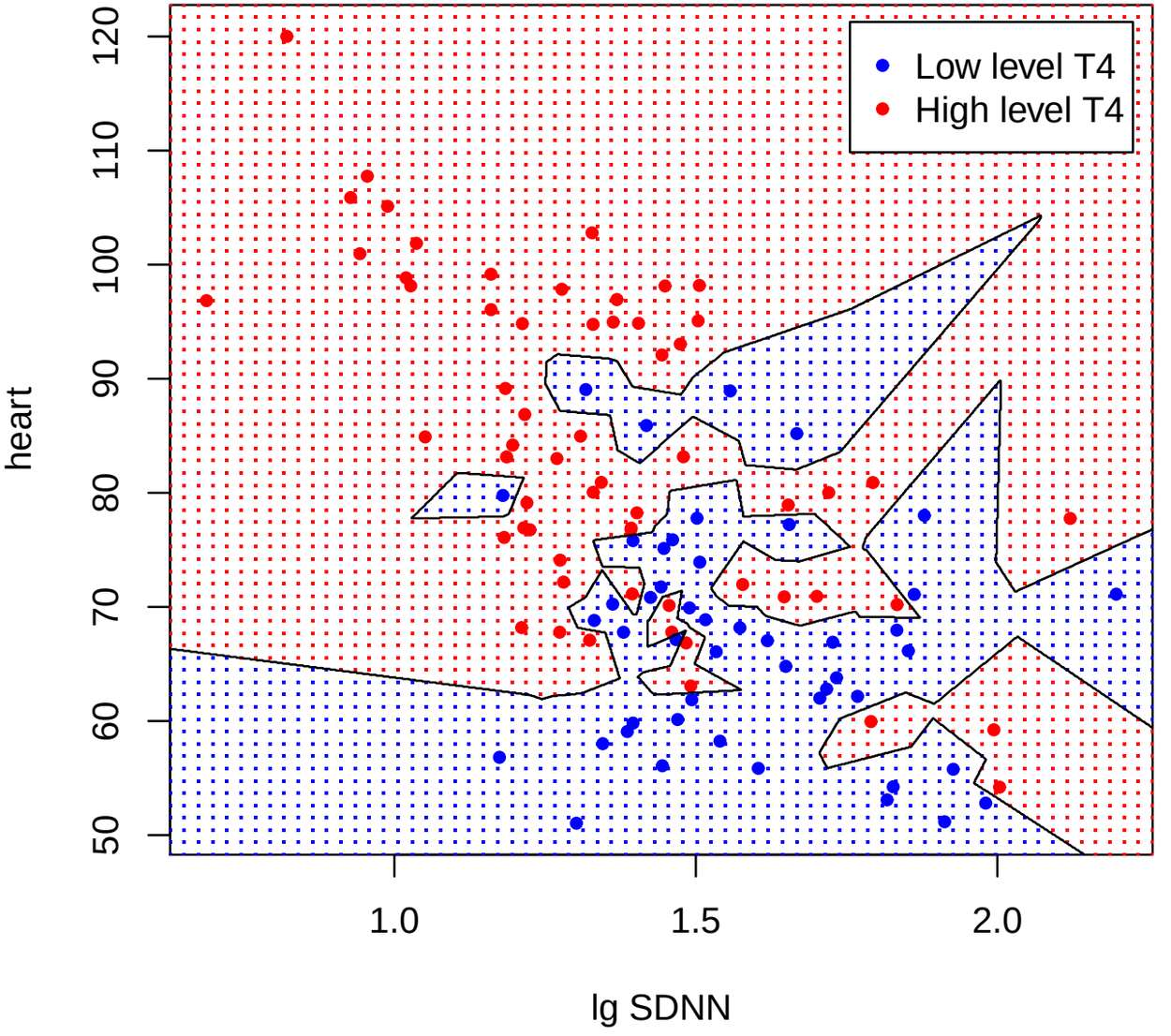

Train
Test
Bayes
Error
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
k

С увеличением k «емкость» («сложность») модели падает,
поэтому развернем горизонтальную ось в обратном направлении
(движение по ней вправо соответствует росту «емкости» модели)

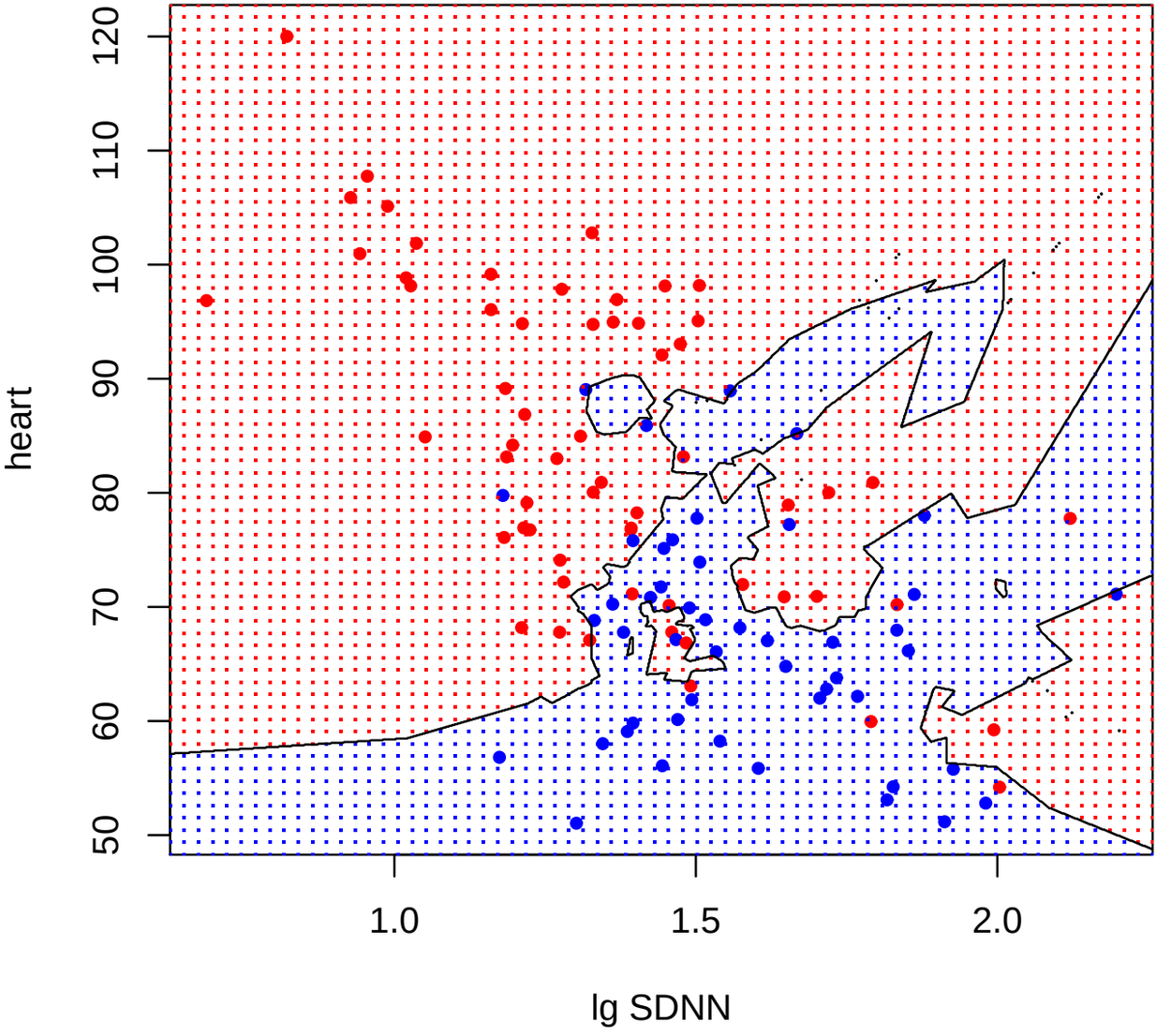



Ошибка
Test
Bayes Error
Train
Емкость модели

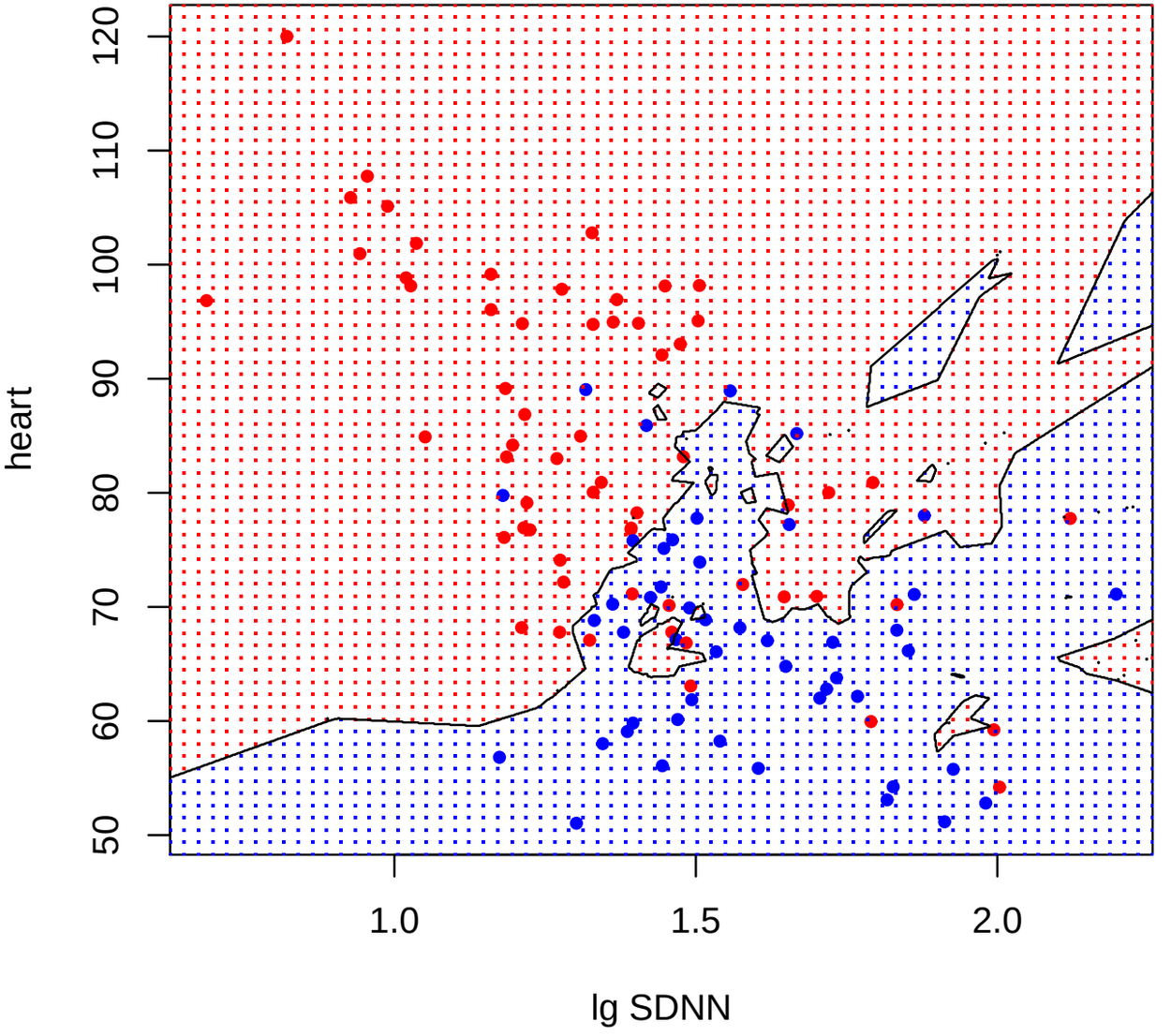
Задача медицинской диагностики. Метод 1 ближайшего соседа. 10-CV ошибка 0.30



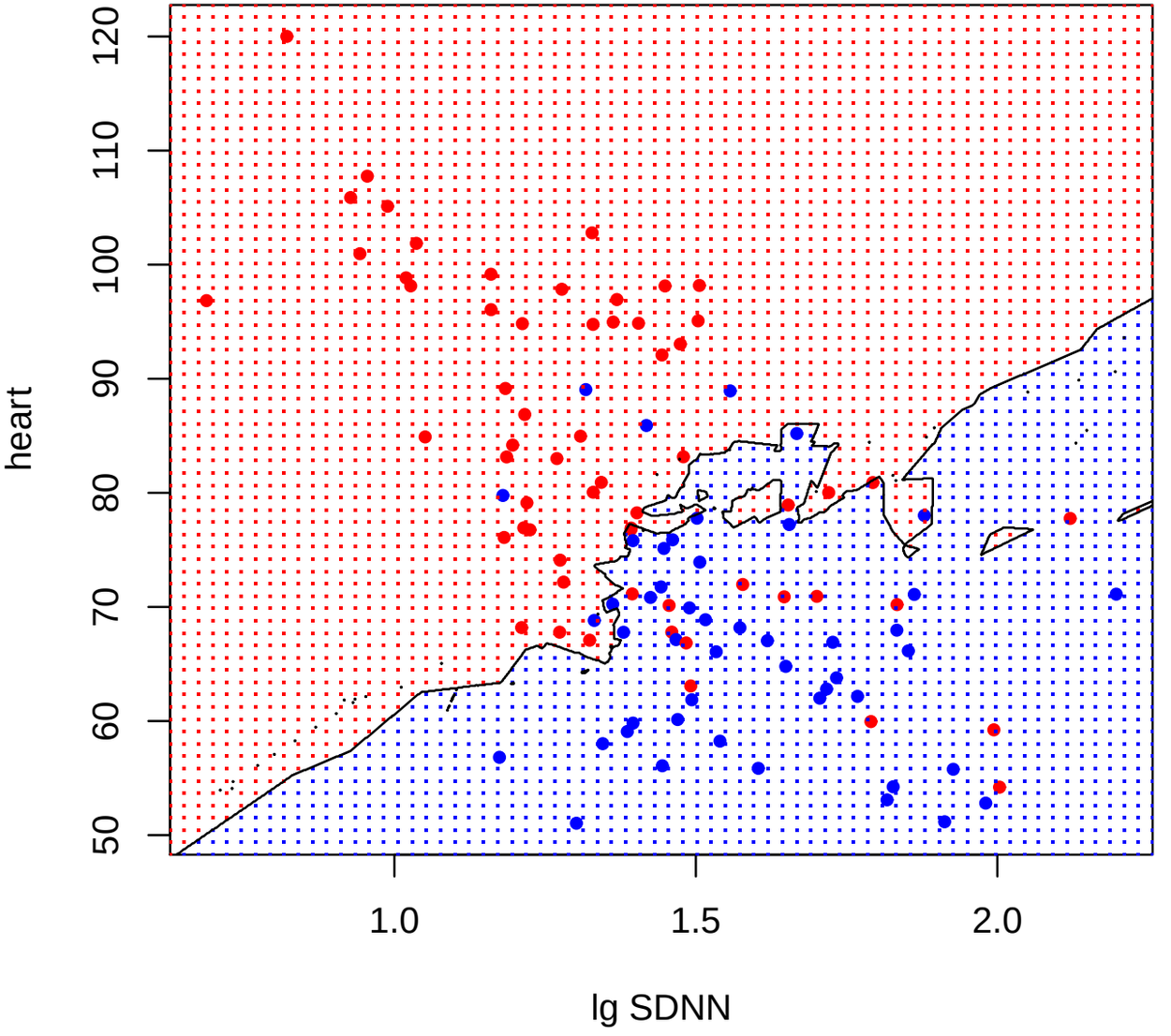
Метод 3 ближайших соседей 10-CV ошибка 0.26



Метод 5 ближайших соседей 10-CV ошибка 0.27



Метод 15 ближайших соседей 10-CV ошибка 0.25



2.3.2. Теорема об оценке риска

Пусть $L(y', y) = I(y' \neq y)$.

При достаточно большом объеме обучающей выборки средний риск классификатора одного ближайшего соседа не более чем в 2 раза превосходит байесов риск, а именно, справделива теорема:

Теорема 2.5 (Cover, Hart, 1967) *Пусть R^* — оптимальное (байесовское) значение среднего риска для некоторой задачи классификации на K классов. Тогда с ростом размера выборки N ожидаемый риск R для метода одного ближайшего соседа сходится к R^o, такому, что*

$$R^* \leq R^o \leq R^* \cdot \left(2 - \frac{K}{K-1} R^*\right) \leq 2R^*.$$

Доказательство. (набросок)

X' — объект из обучающей выборки, ближайший к X, а Y' — соответствующий выход

$k^* = \operatorname{argmax} \Pr(k \mid x)$ — байесово решение (самый популярный класс) в точке x

$r^*(x) = 1 - \Pr(k^* \mid x)$ — предельный условный байесовский риск

$r(x)$ — предельный условный средний риск классификатора ближайшего соседа:

$$r(x) = \lim_{N\to\infty} \Pr\left(Y \neq Y' \mid x\right) = \lim_{N\to\infty} \int_{\mathscr{X}} \Pr\left(Y \neq Y' \mid x, x'\right) p(x' \mid x)\, dx' =$$

$$= \lim_{N\to\infty} \int_{\mathscr{X}} \left(1 - \sum_{k=1}^{K} \Pr(k \mid x) \Pr(k \mid x')\right) p(x' \mid x)\, dx' =$$

$$= \int_{\mathscr{X}} \left(1 - \sum_{k=1}^{K} \Pr(k \mid x) \Pr(k \mid x')\right) \delta(x' - x)\, dx' = 1 - \sum_{k=1}^{K} \Pr^2(k \mid x) =$$

$$= 1 - \Pr^2(k^* \mid x) - \sum_{k\neq k^*} \Pr^2(k \mid x) \leq 1 - \Pr^2(k^* \mid x) - \frac{1}{K-1} \left(\sum_{k\neq k^*} \Pr(k \mid x)\right)^2 =$$

$$= 1 - \Pr^2(k^* \mid x) - \frac{1}{K-1}\Big(1 - \Pr(k^* \mid x)\Big)^2 = 1 - \Big(1 - r^*(x)\Big)^2 - \frac{1}{K-1} r^*(x) = 2r^*(x) - \frac{K}{K-1} r^*(x)^2.$$

Умножая обе части неравенства на $p(x)$ и интегрируя по x, получаем требуемое. ∎

Замечание 2.6 Верхние и нижние оценки в теореме являются точными. В частности, верхняя оценка

$$R^o \leq R^* \cdot \left(2 - \frac{K}{K-1} R^*\right)$$

достигается в случае, когда плотности вероятности $p(x \mid y)$ не зависят от y.

В этом случае $\mathsf{Pr}\{y \mid x\} = \mathsf{Pr}\{y\}$ и $r^*(x)$ не зависят от x.

Замечание 2.7 К сожалению, сходимость R к R^o может быть очень медленной и сложно определить, насколько R близка к R^o.

Замечание 2.8 Результаты аналогичны для метода k ближайших соседей.

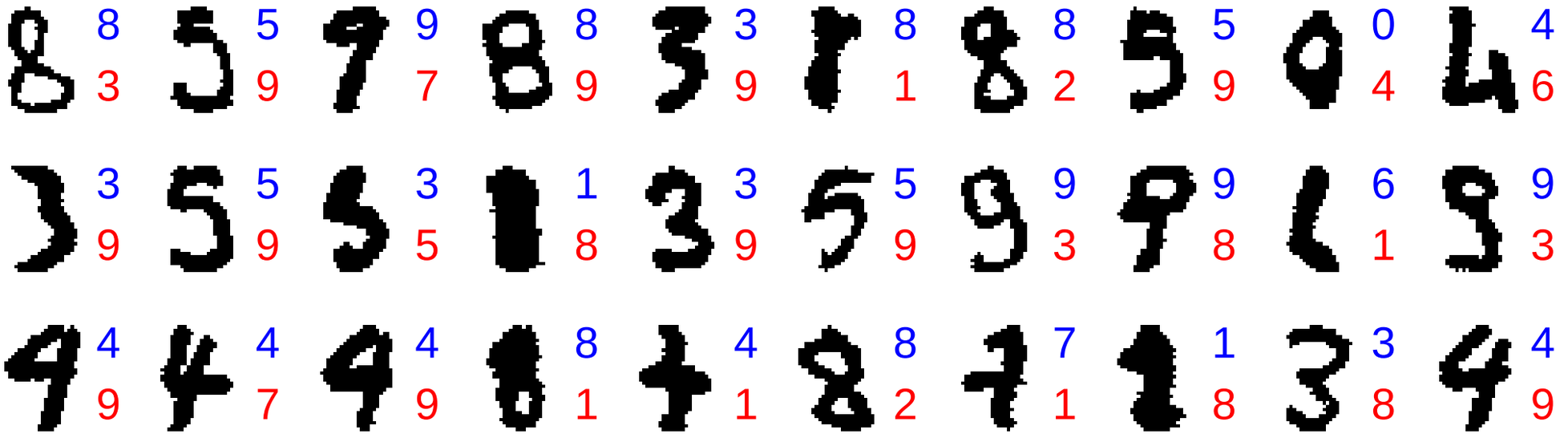
Задача распознавания рукописных цифр

Выборка размера 1934 была случайным образом разбита на две группы: обучающую и тестовую — по 967 объектов в каждой.

k	*Ошибка*	
	на обучающей выборке	*на тестовой выборке*
1	0	0.031
2	0.017	0.046
3	0.014	0.031
5	0.026	0.033
10	0.034	0.038
15	0.042	0.046
20	0.051	0.050

Все случаи неправильной классификации цифр из тестовой выборки в случае $k = 1$.

Красная цифра — ответ классификатора, синяя — верный ответ.



Как правило, метод ближайшего соседа имеет проблемы при большой размерности (если признаки количественные).

2.4. Плюсы и минусы метода kNN

Плюсы

- Простой метод
- Для ряда задач показывает неплохие результаты
- Достаточно устойчив к выбросам (при подходящем выборе k)
- Работает как с числовыми, так и номинальными признаками
- Быстрый (если использовать специальные структуры данных: kd-деревья и т.д.)

Минусы

- Сколько соседей брать?
- Какую метрику использовать?
- Необходимо хранить всю выборку
- Подвержен «проклятию размерности»