МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

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

Лекция 3

Оценка качества и выбор модели

3.1. Экспериментальная оценка качества обучения и выбор модели

3.1.1. Оценка качества обучения

model assessment (model evaluation)

Алгоритм обучения по обучающей выборке $D \in \mathscr{D}$ строит решающую функцию $f \in \mathscr{F}$

Мы можем вычислить эмпирический риск $\widehat{R}(f)$, но нужно уметь оценивать $R(f)$

Как правило, $\widehat{R}(f) < R(f)$

Как оценить $R(f)$?

Случайно разделим все имеющиеся данные

- на обучающую (train) выборку $(x_{\text{train}}^{(1)}, y_{\text{train}}^{(1)}),\ (x_{\text{train}}^{(2)}, y_{\text{train}}^{(2)}),\ \ldots,\ (x_{\text{train}}^{(N_{\text{train}})}, y_{\text{train}}^{(N_{\text{train}})})$,
- на тестовую (test) выборку $(x_{\text{test}}^{(1)}, y_{\text{test}}^{(1)}),\ (x_{\text{test}}^{(2)}, y_{\text{test}}^{(2)}),\ \ldots,\ (x_{\text{test}}^{(N_{\text{train}})}, y_{\text{test}}^{(N_{\text{test}})})$.

Обучающая выборка используется для построения моделей $f \in \mathscr{F}$.

Тестовая — для оценки среднего риска $R(f)$ решающей функции f.

$$\widehat{R}(f) = \widehat{R}_{\text{train}}(f) = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} L\big(f(x_{\text{train}}^{(i)}),\ y_{\text{train}}^{(i)}\big)$$

$$\widehat{R}_{\text{test}}(f) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} L\big(f(x_{\text{test}}^{(i)}),\ y_{\text{test}}^{(i)}\big)$$

$$\widehat{R}_{\text{train}}(f) < R(f) \approx \widehat{R}_{\text{test}}(f)$$

Метод перекрестного контроля. 1-й вариант

Разбиение на обучающую и тестовую выборки можно проделать L раз (каждый раз случайно).

В результате получим L решающих функций: $f_1, f_2, \ldots, f_L$

и L значений эмпирического риска: $\widehat{R}_1^{\text{test}}(f_1),\ \widehat{R}_2^{\text{test}}(f_2),\ \ldots,\ \widehat{R}_L^{\text{test}}(f_L)$

Окончательную модель f построим *на основе всех данных*.

Оцениваем риск:

$$R(f) \approx \overline{R} = \frac{1}{L}\sum_{\ell=1}^{L} \widehat{R}_\ell^{\text{test}}(f_\ell).$$

Очевидно, что оценка несмещенная.

Построение интервальных оценок для оценки риска в этом методе — пока нерешенная задача.

Метод перекрестного контроля по M блокам (M-fold CV)

Случайным образом разобьем исходную выборку на M непересекающихся примерно равных по размеру частей:

Последовательно каждую из этих частей рассмотрим в качестве тестовой выборки, а объединение остальных частей — в качестве обучающей выборки:

Случайным образом разобьем исходную выборку на M непересекающихся примерно равных по размеру частей:

Последовательно каждую из этих частей рассмотрим в качестве тестовой выборки, а объединение остальных частей — в качестве обучающей выборки:

Случайным образом разобьем исходную выборку на M непересекающихся примерно равных по размеру частей:

Последовательно каждую из этих частей рассмотрим в качестве тестовой выборки, а объединение остальных частей — в качестве обучающей выборки:

Случайным образом разобьем исходную выборку на M непересекающихся примерно равных по размеру частей:

Последовательно каждую из этих частей рассмотрим в качестве тестовой выборки, а объединение остальных частей — в качестве обучающей выборки:

Случайным образом разобьем исходную выборку на M непересекающихся примерно равных по размеру частей:

Последовательно каждую из этих частей рассмотрим в качестве тестовой выборки, а объединение остальных частей — в качестве обучающей выборки:

В результате получим M решающих функций: $f_1, f_2, \ldots, f_M$

и M значений эмпирического риска: $\widehat{R}_1^{\text{cv}}(f_1),\ \widehat{R}_2^{\text{cv}}(f_2),\ \ldots,\ \widehat{R}_M^{\text{cv}}(f_M)$

Окончательную решающую функцию f построим на основе *всех* данных.

$$R(f) \approx \widehat{R}^{\text{cv}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{R}_m^{\text{cv}}(f_m)$$

Всю процедуру перекрестного контроля можно повторить L раз, каждый раз разбивая случайно выборку на M частей.

Тогда оценка риска будут вычисляться на основе $M \cdot L$ оценок эмпирического риска

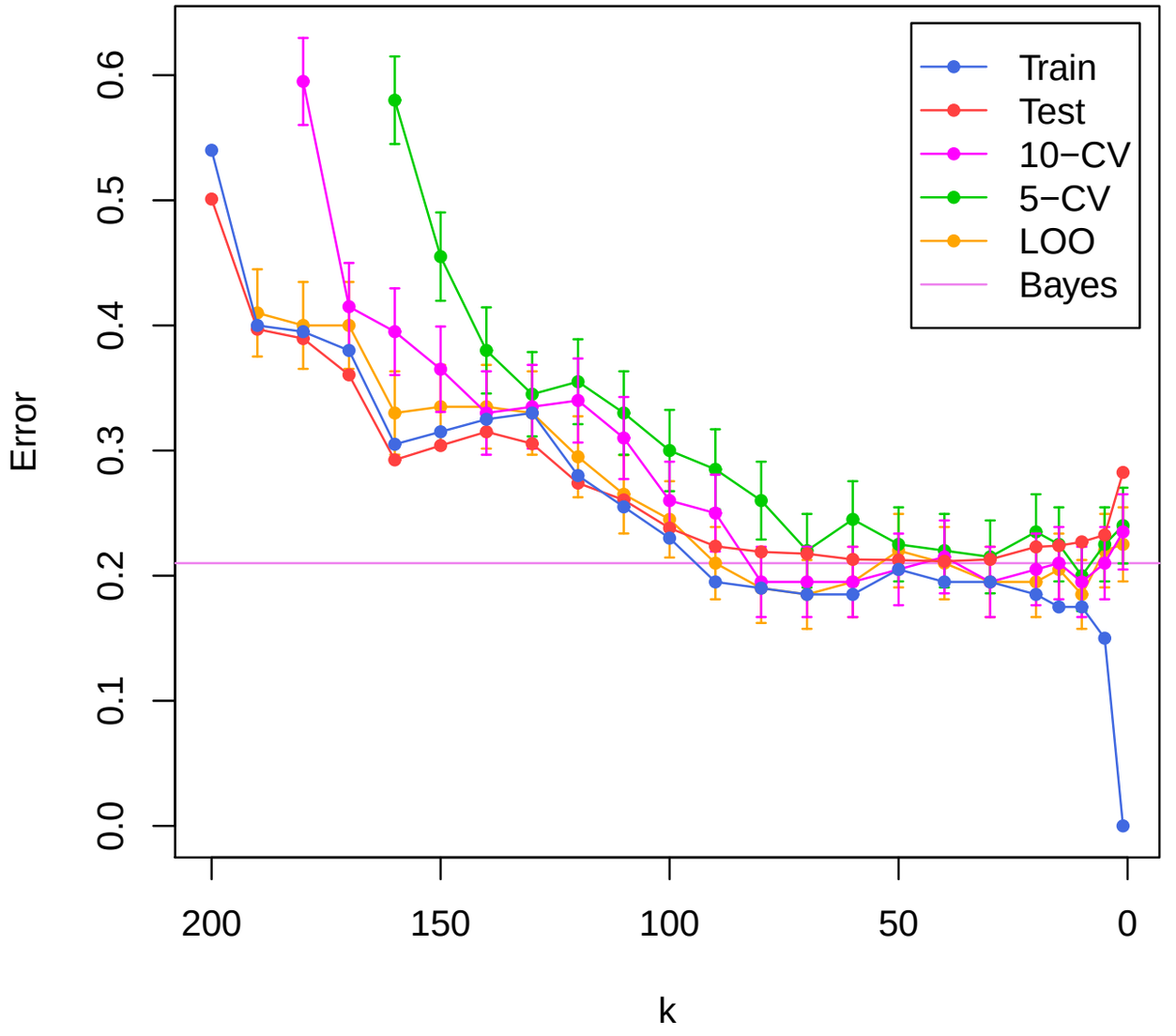
Какое M выбирать?

Часто используемые значения M: $M = 5$ или $M = 10$.

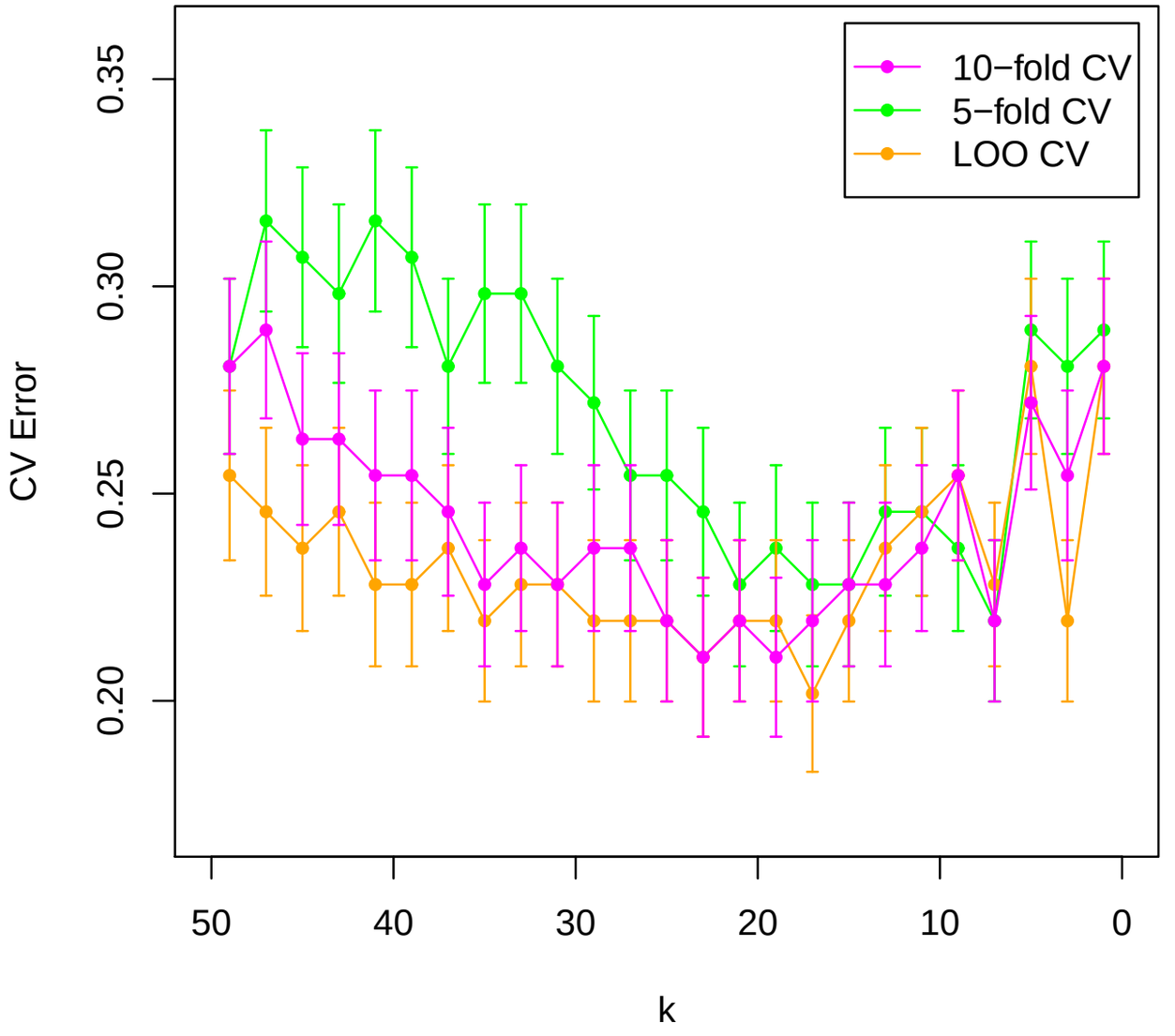
$M = N$ — метод перекрестного контроля *с одним отделяемым элементом* (*leave-one-out cross-validation*, LOO).

LOO — самый точный, но требует много времени.

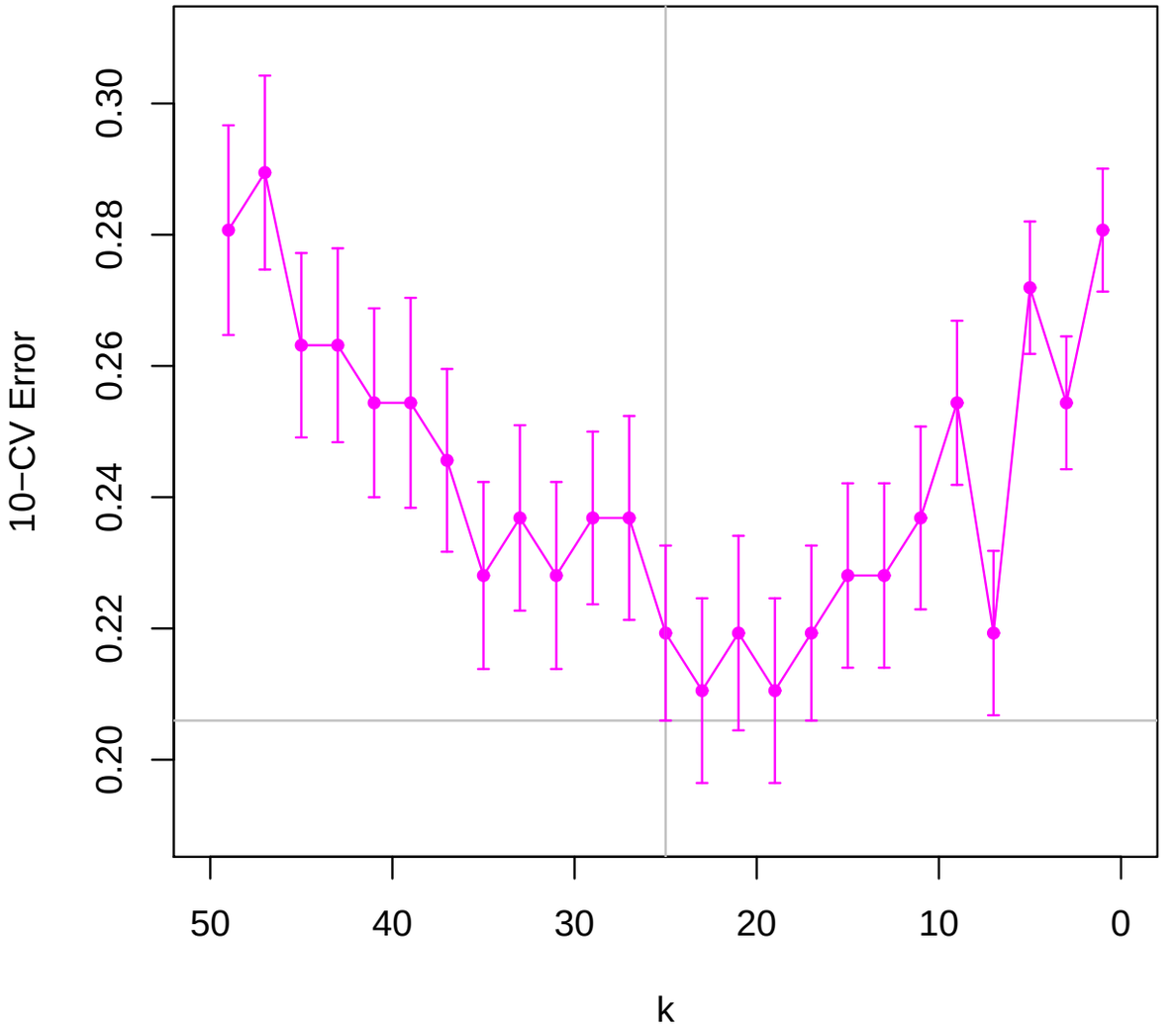
Перекрестный контроль для задачи классификации методом k ближайших соседей.



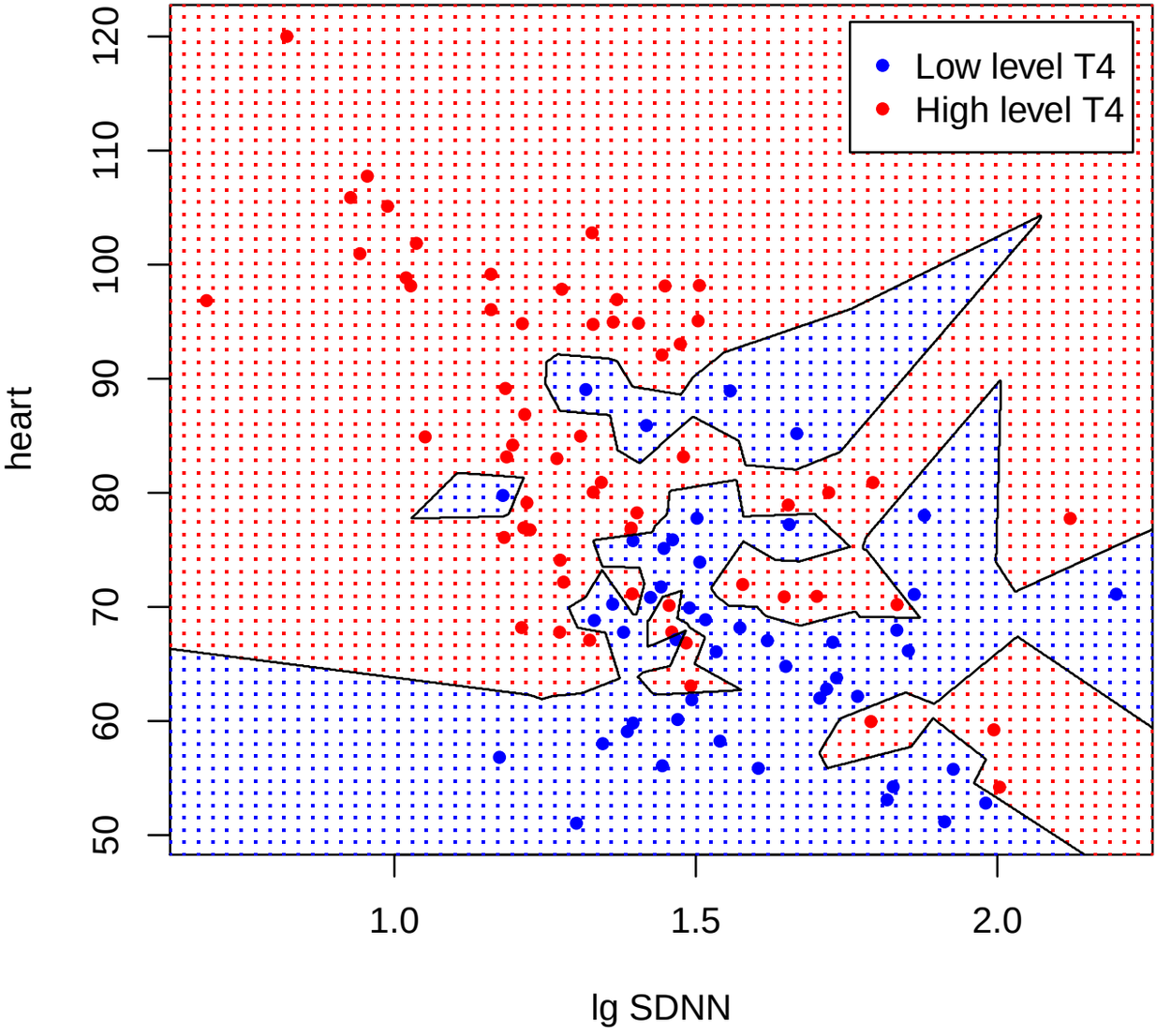
Медицинская диагностика $0.5 \cdot \mathrm{se}$



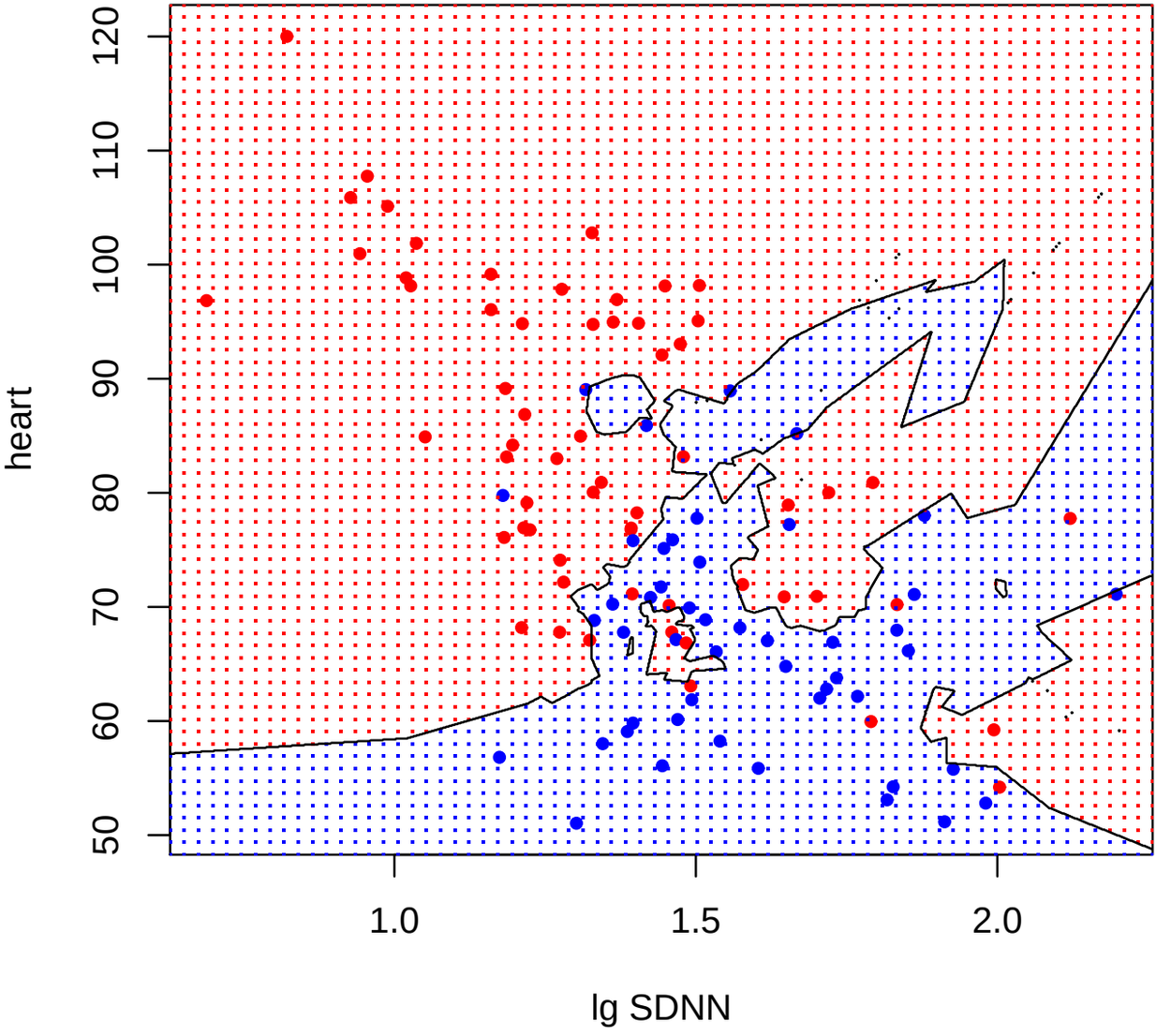
Правило одного se $kk[i.min] = 19$ $err[i.min] = 0.2105263$ $se[i.min] = 0.03835154$ $kk[i.opt] = 37$ $err[i.opt] = 0.245614$ $se[i.opt] = 0.04049339$



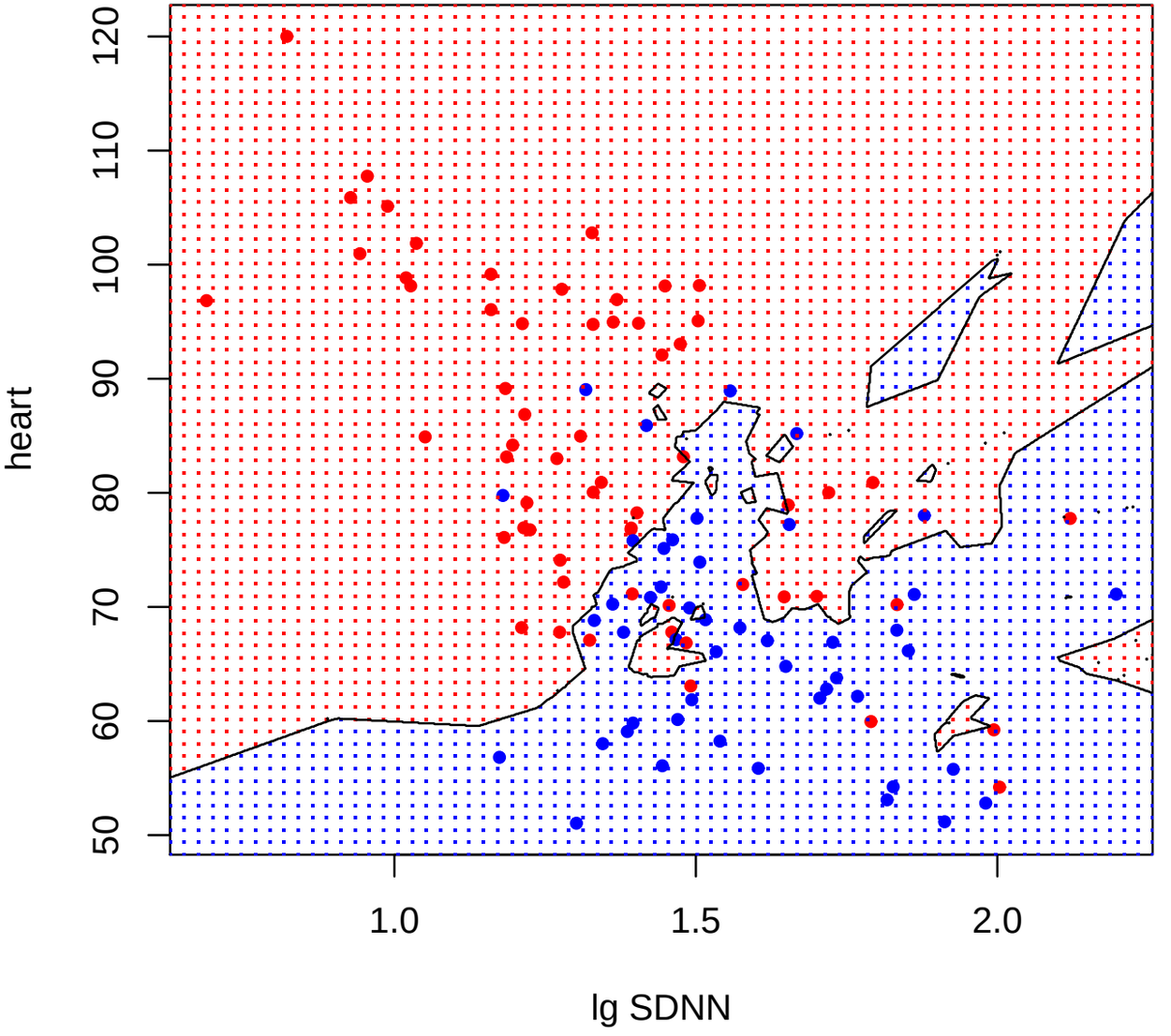
Задача медицинской диагностики. Метод 1 ближайшего соседа. 10-CV ошибка 0.30



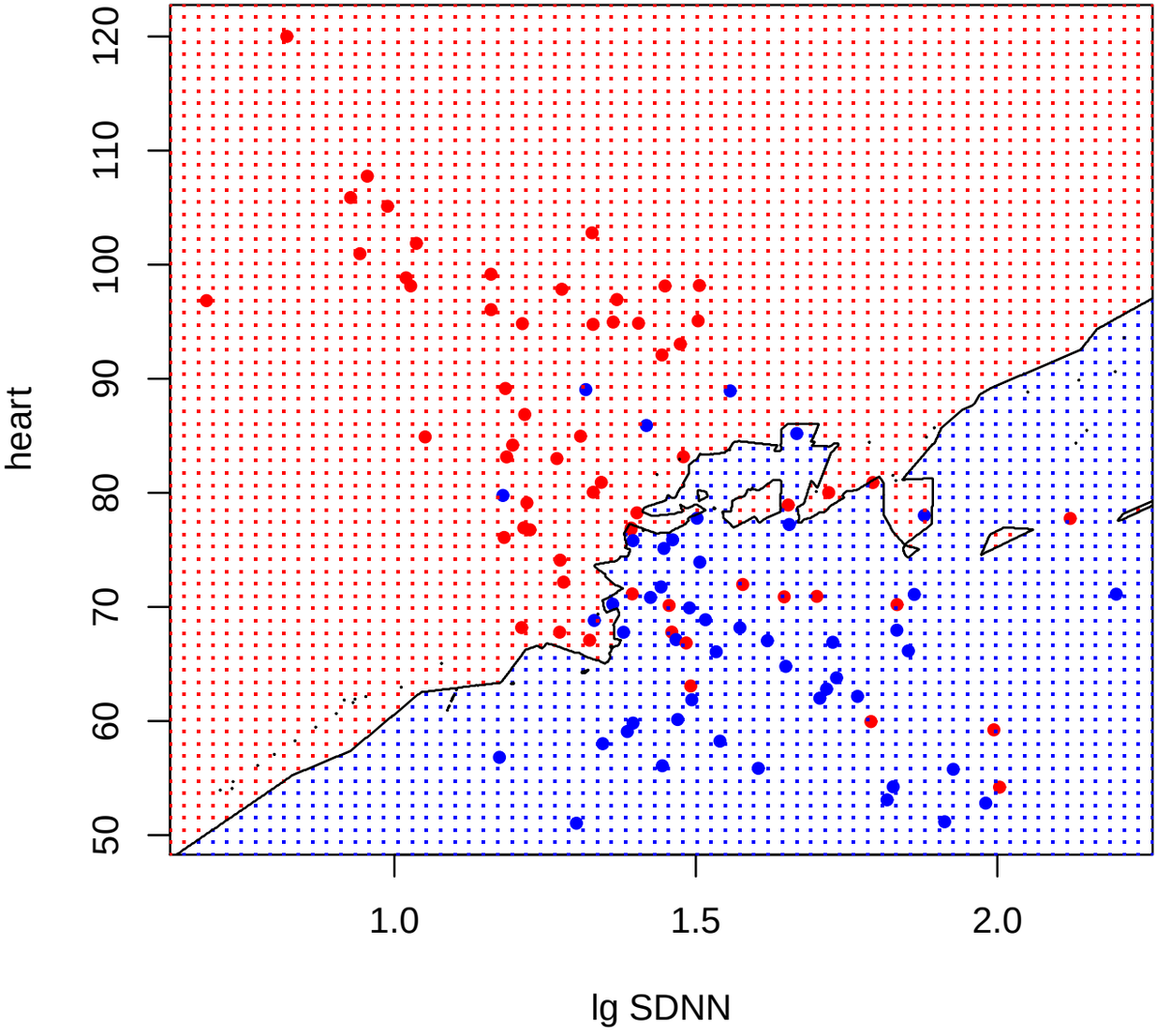
Метод 3 ближайших соседей 10-CV ошибка 0.26



Метод 5 ближайших соседей 10-CV ошибка 0.27



Метод 15 ближайших соседей 10-CV ошибка 0.25



Какое значение M использовать?

Метод скользящего контроля характеризует не конкретную решающую функцию f, а весь метод обучения.

При $M = N$ (LOO) обучающие выборки похожи друг на друга, следовательно, решающие функции f_m, по-видимому, будут похожи на функцию f, построенную по всей выборке, поэтому LOO обычно дает лучшие оценки среднего риска.

Однако LOO требует много времени.

При небольших значениях M ($M = 5, 10$) оценка скользящего контроля $\widehat{R}^{\text{cv}}(f_i)$ может сильнее отличаться от среднего риска $R(f)$ функции, построенной по всей обучающей выборке, так как

1) обучающие выборки сильнее отличаются друг от друга;

2) обучающие выборки могут оказаться слишком маленькими, чтобы построить хорошую решающую функцию.

3.1.2. Выбор модели

(model selection)

Алгоритм обучения по обучающей выборке $D \in \mathscr{D}$ строит решающую функцию $f \in \mathscr{F}$

Пусть α — гиперпараметр (м.б. вектором) алгоритма (метода) обучения.

Для одной и той же обучающей выборки D, но разных α будем получать разные решающие функции $f(x, \alpha)$.

Какую из них выбрать?

Хотелось бы, конечно, найти $f^* = f(\cdot, \alpha^*)$, где

$$\alpha^* = \operatorname*{argmin}_{\alpha} R\big(f(\cdot, \alpha)\big)$$

Но приходится довольствоваться $f_o = f(\cdot, \alpha_o)$, где

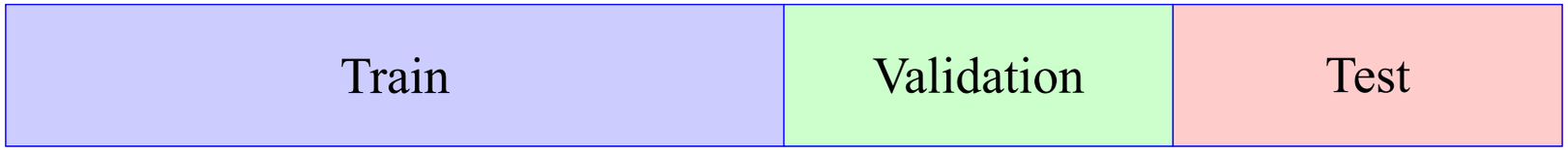
$$\alpha_o = \operatorname*{argmin}_{\alpha} \widehat{R}^{\text{test}}\big(f(\cdot, \alpha)\big) \qquad \text{или} \qquad \alpha_o = \operatorname*{argmin}_{\alpha} \widehat{R}^{\text{cv}}\big(f(\cdot, \alpha)\big).$$

Теперь тестовая выборка стала обучающей! Поэтому, как правило,

$$\widehat{R}^{\text{test}}\big(f(\cdot, \alpha_o)\big) < R(f_o)$$

Если данных достаточно, то их делят

- на обучающую (train) выборку,
- на проверочную (validation) выборку,
- на тестовую (test) выборку.

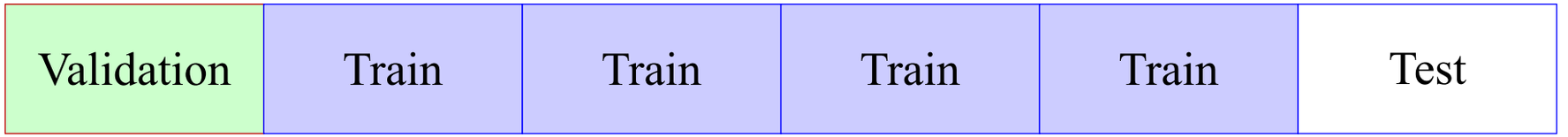


Обучающая выборка используется для построения моделей $f(\cdot, \alpha) \in \mathscr{F}$ для разных α.

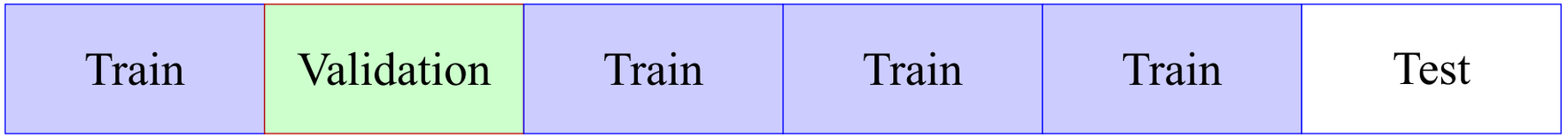
Проверочная — для оценки оценки среднего риска каждой из построенной модели и выбора наилучшей модели в $\mathscr{F}$.

Тестовая — для оценки ошибки предсказания выбранной модели.

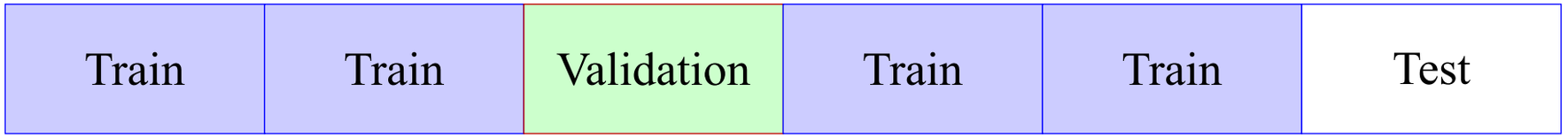
Для выбора наилучшей модели можно использовать перекрестную проверку.



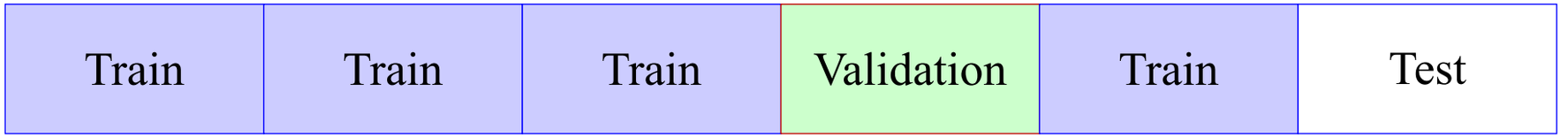
Для выбора наилучшей модели можно использовать перекрестную проверку.



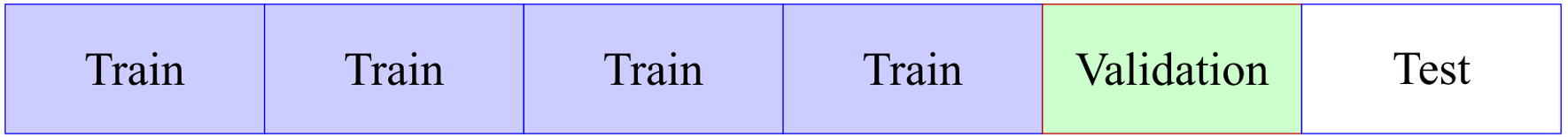
Для выбора наилучшей модели можно использовать перекрестную проверку.



Для выбора наилучшей модели можно использовать перекрестную проверку.



Для выбора наилучшей модели можно использовать перекрестную проверку.



Для выбора наилучшей модели можно использовать перекрестную проверку.

Для каждого α запускаем процедуру обучения — перекрестного контроля.

Выбираем $f_o = f(\cdot,\, \alpha_o)$, где

$$\alpha_o = \operatorname*{argmin}_{\alpha} \widehat{R}^{\text{cv}}\big(f(\cdot,\, \alpha)\big).$$

Оцениваем риск на тестовой выборке.

Для построения окончательной модели используем найденный гиперпараметр α_o и *всю* выборку