ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
УПРАВЛЕНИЯ

Основан в 1919 году

Методы и технологии машинного обучения

Лекция 2: Параметрические классификаторы для бинарной Y

Светлана Андреевна Суязова (Аксюк)
sa_aksyuk@guu.ru

осенний семестр 2021 / 2022 учебного года

План лекции

- Логистическая регрессия

- Линейный дискриминантный анализ
- Порог отсечения классов и ROC-кривая
- Квадратичный дискриминантный анализ
- Методы создания повторных выборок

ru-xkcd.livejournal.com/75022.html

Почему не обычная регрессия?

Y – категориальная переменная, например, вид ириса:

$$Y = \begin{cases} 1, & setosa \\ 2, & versicolor \\ 3, & virginica \end{cases}$$

Вопросы:

1. Расстояние между категориями?
2. Порядок категорий?

Y – категория; 1, 2, 3 – *метки*

Почему не обычная регрессия?

Y – категория; 1, 2, ... – *метки*

Две категории – **логистическая регрессия**

- Y – бинарный (0 – отсутствие признака; 1 – наличие признака). Y интерпретируется как принадлежность одному из классов; нужно задать порог отсечения.
- Y – оценка вероятности (частость), $Y \in [0, 1]$. Y интерпретируется как вероятность принадлежности одному из классов.

Более двух категорий – нерегрессионные модели.

Логистическая регрессия

Сгенерированные данные по 10000 держателям кредитных карт `Default`

default	student	balance	income
No	No	729.5265	44361.63
No	Yes	817.1804	12106.13
No	No	1073.5492	31767.14
No	No	529.2506	35704.49
No	No	785.6559	38463.50

Задача: предсказать `default` по `balance`.

Логистическая регрессия

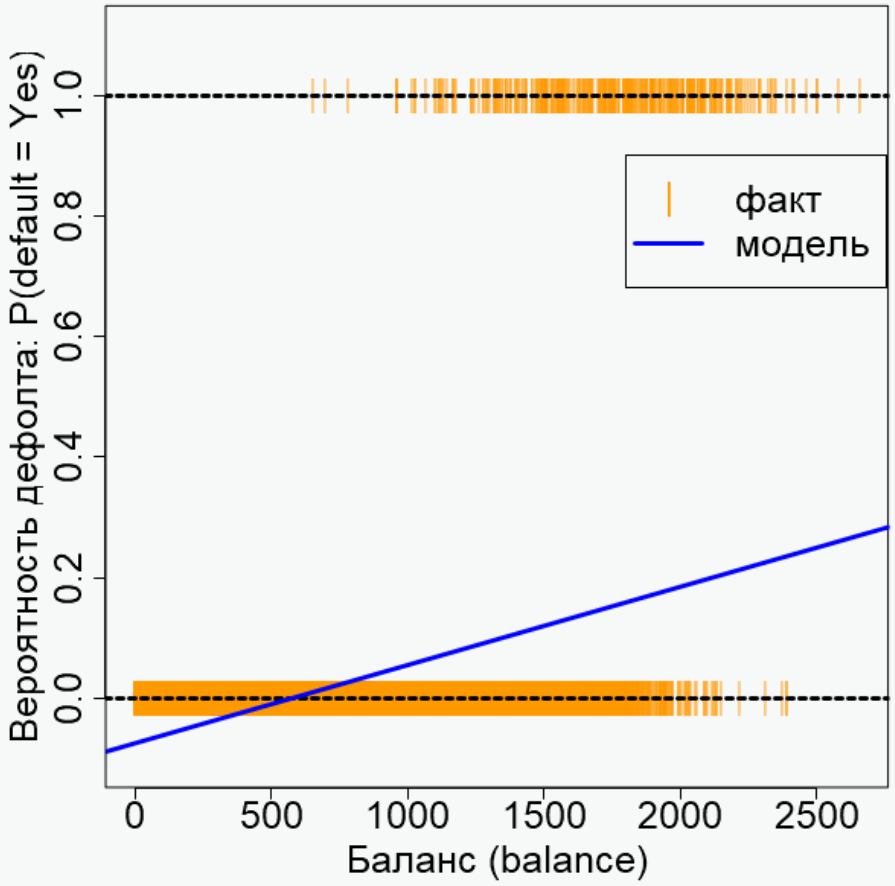
Зависимая переменная модели – условная вероятность:

$$Y = P(\mathrm{default} = \mathrm{Yes} | \mathrm{balance})$$

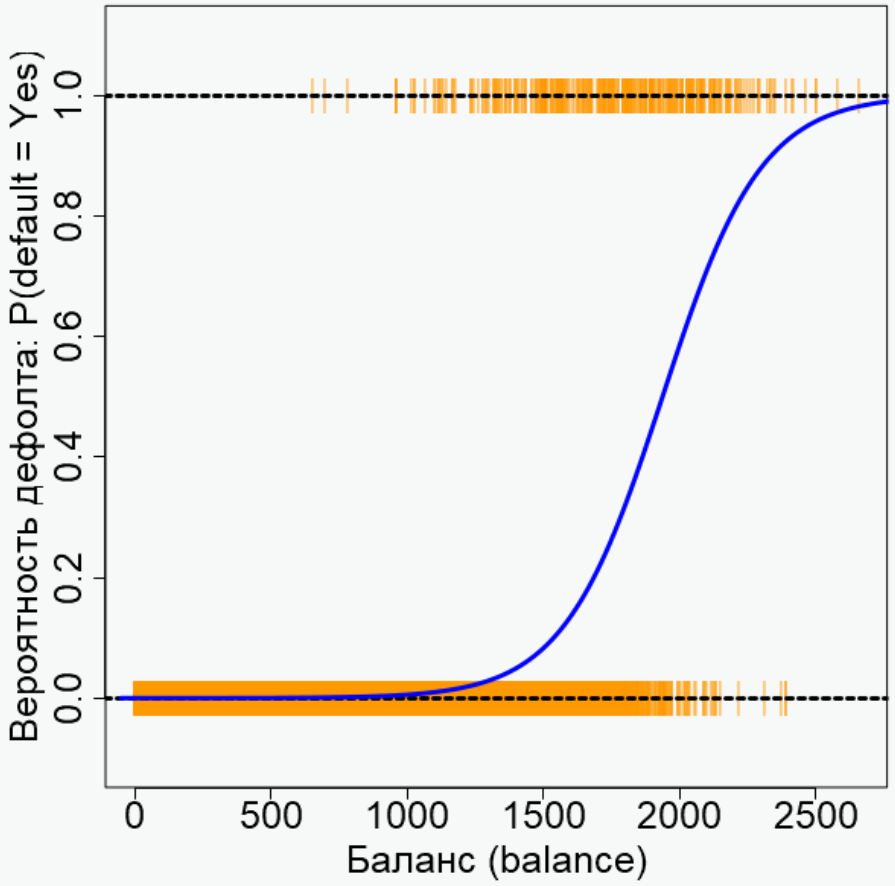
$$P(X) = f(\hat{\beta}_0 + \hat{\beta}_1 \cdot X)$$

Логистическая функция возвращает для любого X значение из интервала от 0 до 1:

$$P(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot X}}$$



Обычная линейная регрессия



Логистическая регрессия

Логистическая регрессия

$$P(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot X}} \Leftrightarrow \frac{P(X)}{1 - P(X)} = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot X}$$

$\frac{P(X)}{1-P(X)} \in (0, \infty)$ – **риск события**:

Например: если $P(X) = 0.8$, это означает, что 8 из 10 человек станут неплательщиками с риском:

$$\frac{P(X)}{1 - P(X)} = \frac{0.8}{1 - 0.8} = 4$$

Логистическая регрессия

$$\ln\left(\frac{P(X)}{1-P(X)}\right) = \ln\left(e^{\hat{\beta}_0+\hat{\beta}_1\cdot X}\right) \Leftrightarrow \ln\left(\frac{P(X)}{1-P(X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

$\ln\left(\frac{P(X)}{1-P(X)}\right)$ – логарифм риска, или *логит*.

- коэффициент $\hat{\beta}_1$ *не отражает изменение* $P(X)$*, вызванное увеличением* X *на 1*;
- скорость изменения $P(X)$ с изменением X на 1 зависит от текущего значения X;
- направление связи интерпретируется как в линейной регрессии: если $\hat{\beta}_1 > 0$, $X \uparrow\uparrow P(X)$; если $\hat{\beta}_1 < 0$, $X \uparrow\downarrow P(X)$.

Логистическая регрессия: оценка параметров

Принцип метода максимального правдоподобия: подобрать $\hat{\beta}_0$, $\hat{\beta}_1$ так, чтобы вероятность дефолта $\hat{p}(x_i)$ для каждого держателя карты максимально близко соответствовала его наблюдаемому статусу.

Функция правдоподобия:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \to \max$$

МНК в линейной регрессии – частный случай ММП.

На примере данных `Default`

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.6513	0.36	-29.49	0.0000
balance	0.0055	0.00	24.95	0.0000

Информационный коэффициент Акаике: $AIC = 1600.45$

$$\ln\left(\frac{P(X)}{1-P(X)}\right) = -10.6513 + 0.0055 \cdot \text{balance}$$

Прогнозы

Вероятность дефолта для $\text{balance} = 1000$: $\hat{p}(1000) = \frac{e^{-10.6513+0.0055\cdot 1000}}{1+e^{-10.6513+0.0055\cdot 1000}} \approx 0.006$

Вероятность дефолта для $\text{balance} = 2000$: $\hat{p}(2000) = \frac{e^{-10.6513+0.0055\cdot 2000}}{1+e^{-10.6513+0.0055\cdot 2000}} \approx 0.586$

Положительный баланс на кредитной карте – долг

Качественная объясняющая переменная: `studentYes`

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.5041	0.07	-49.55	0.0000
studentYes	0.4049	0.12	3.52	0.0004

Информационный коэффициент Акаике: $AIC = 2912.68$

$$\ln\left(\frac{P(X)}{1-P(X)}\right) = -3.5041 + 0.4049 \cdot \text{studentYes}$$

$$\hat{P}(\text{default} = \text{Yes}|\text{student} = \text{No}) = \frac{e^{-3.5041+0.4049\cdot 0}}{1+e^{-3.5041+0.4049\cdot 0}} \approx 0.029$$

$$\hat{P}(\text{default} = \text{Yes}|\text{student} = \text{Yes}) = \frac{e^{-3.5041+0.4049\cdot 1}}{1+e^{-3.5041+0.4049\cdot 1}} \approx 0.043$$

Множественная логистическая регрессия

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.8690	0.49	-22.08	0.0000
studentYes	-0.6468	0.24	-2.74	0.0062
balance	0.0057	0.00	24.74	0.0000
income	0.0000	0.00	0.37	0.7115

Информационный коэффициент Акаике: $AIC = 1579.54$

$$\ln\left(\frac{P(X)}{1-P(X)}\right) = -10.869 - 0.6468 \cdot \text{studentYes} + 0.0057 \cdot \text{balance} + 0 \cdot \text{income}$$

Взаимодействие объясняющих переменных приводит к изменению знака коэффициента при факторе `student`.

Без незначимой переменной `income`

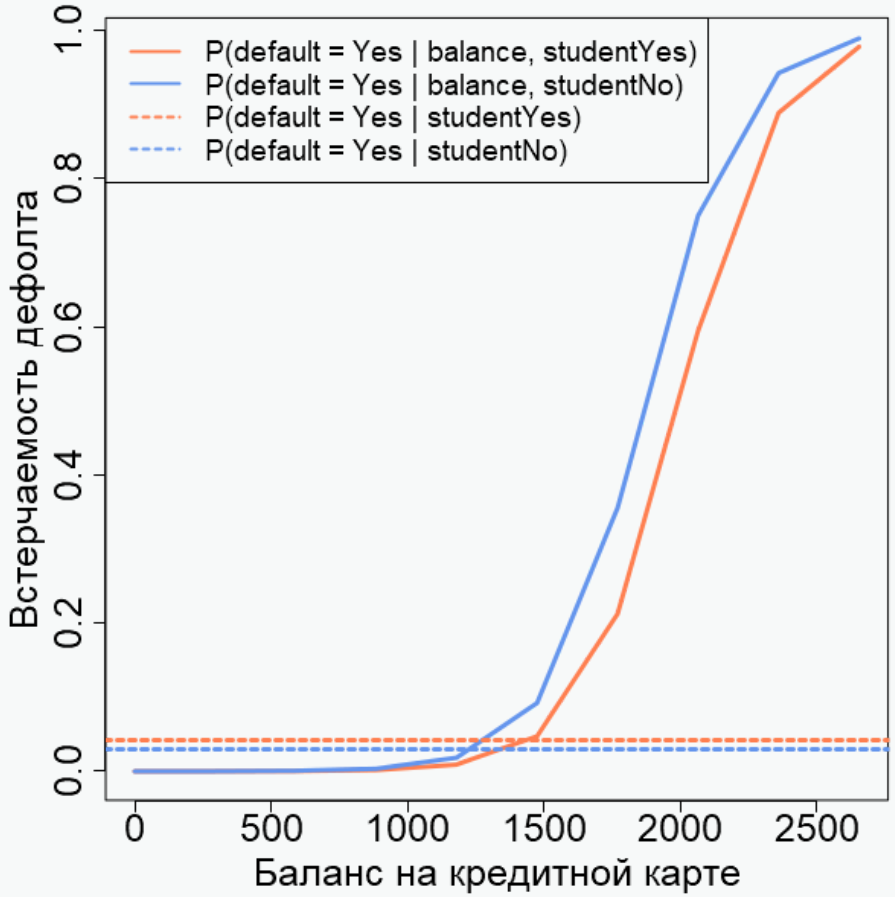
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.7495	0.37	-29.12	0.0000
balance	0.0057	0.00	24.75	0.0000
studentYes	-0.7149	0.15	-4.85	0.0000

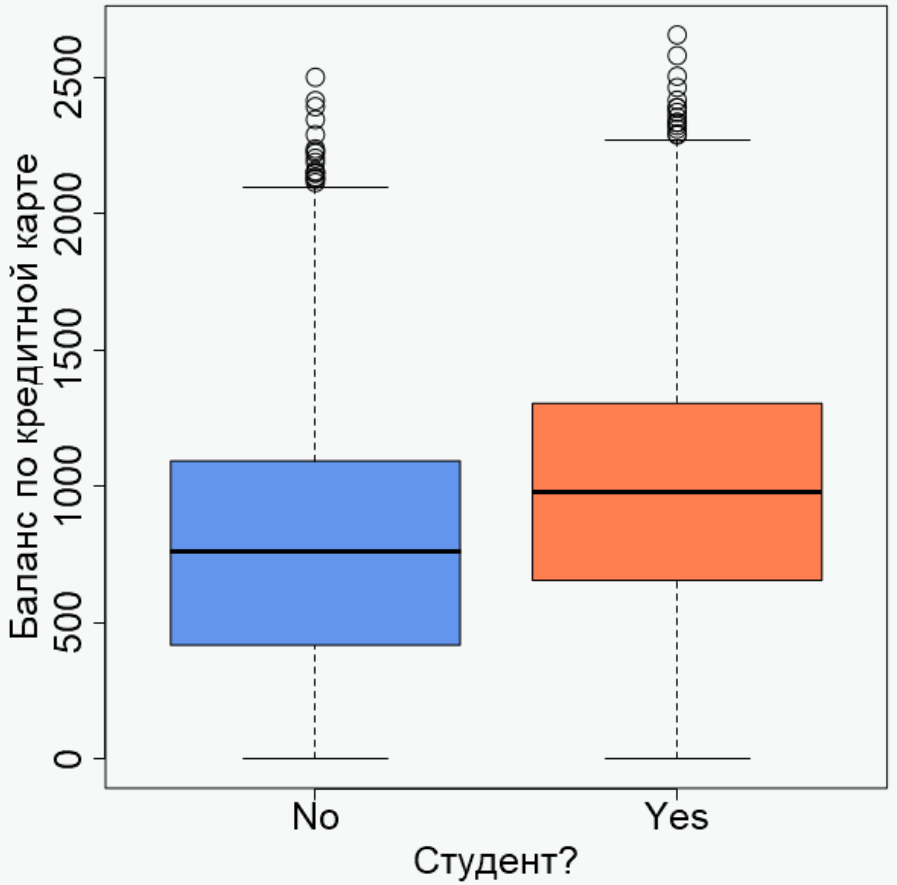
Информационный коэффициент Акаике: $AIC = 1577.68$

$$\hat{p}(1500, 1) = \frac{e^{-10.7495 + 0.0057 \cdot 1500 - 0.7149 \cdot 1}}{1 + e^{-10.7495 + 0.0057 \cdot 1500 - 0.7149 \cdot 1}} \approx 0.054$$

$$\hat{p}(1500, 0) = \frac{e^{-10.7495 + 0.0057 \cdot 1500 - 0.7149 \cdot 0}}{1 + e^{-10.7495 + 0.0057 \cdot 1500 - 0.7149 \cdot 0}} \approx 0.105$$

Смешивание эффектов предикторов





План лекции

- Логистическая регрессия
- **Линейный дискриминантный анализ**
- Порог отсечения классов и ROC-кривая
- Квадратичный дискриминантный анализ
- Методы создания повторных выборок

Дискриминантный анализ

В *логистической регрессии* моделируем условное распределение отклика Y с учётом предикторов X.

В *дискриминантном анализе* – распределение предикторов X отдельно для каждого класса, затем применяем теорему Байеса для получения условной вероятности $P(Y = k|X = x)$.

Почему дискриминантный анализ, а не логит?

1. Когда классы хорошо разделены, оценки логистической регрессии нестабильны.
2. При малом n и нормально распределённых X дискриминантный анализ более устойчив.
3. Дискриминантный анализ применим в задачах, где у Y больше двух классов.

Теорема Байеса для классификации

Отнести наблюдение x к одному из K классов.

π_k – *априорная вероятность* принадлежности наблюдения к классу k

$f(X) = P(X = x|Y = k)$ – функция плотности вероятности X для наблюдения из класса k

Найти *апостериорную вероятность* принадлежности наблюдения x к классу k (*исходя из значения предиктора*):

$$P(Y = k|X = x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l \cdot f_l(x)} \quad (1)$$

Дискриминантная функция по одному признаку

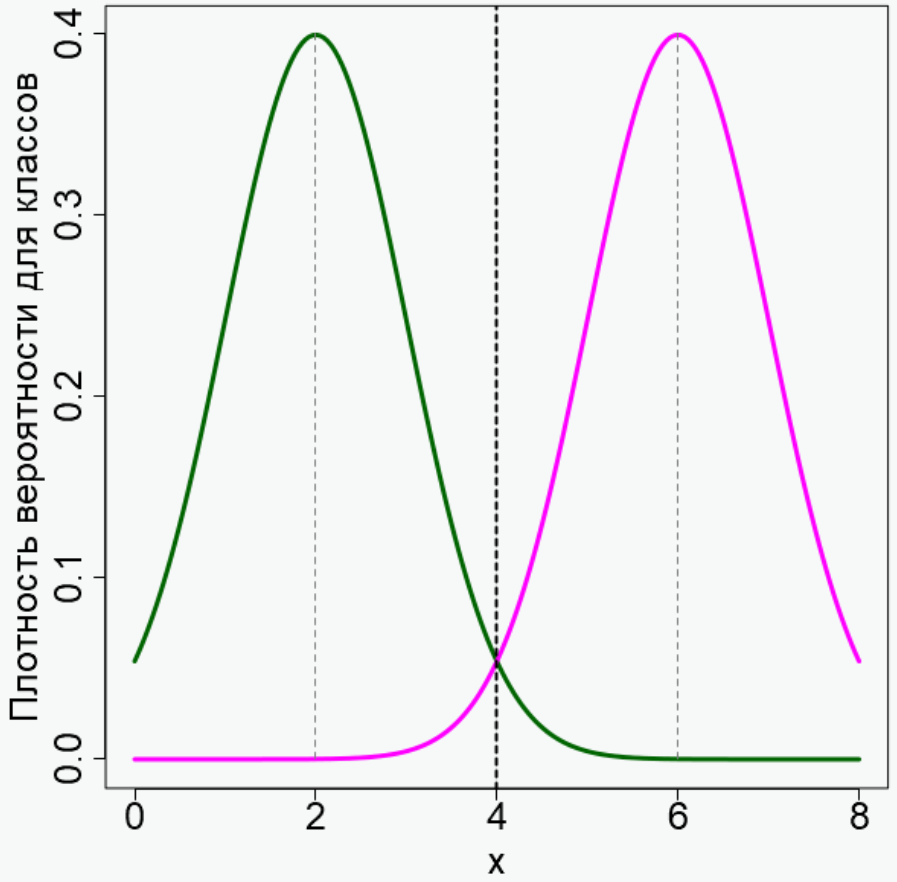
LDA: $f_k(x)$ подчиняются нормальным законам с одинаковой дисперсией σ^2:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2(x-\mu_k)^2}\right), \quad (2)$$

где μ_k – это математическое ожидание класса k. Из подстановки (2) в (1) выводится правило для отнесения наблюдения x к классу k:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \to \max_k$$

Дискриминантная функция по одному признаку



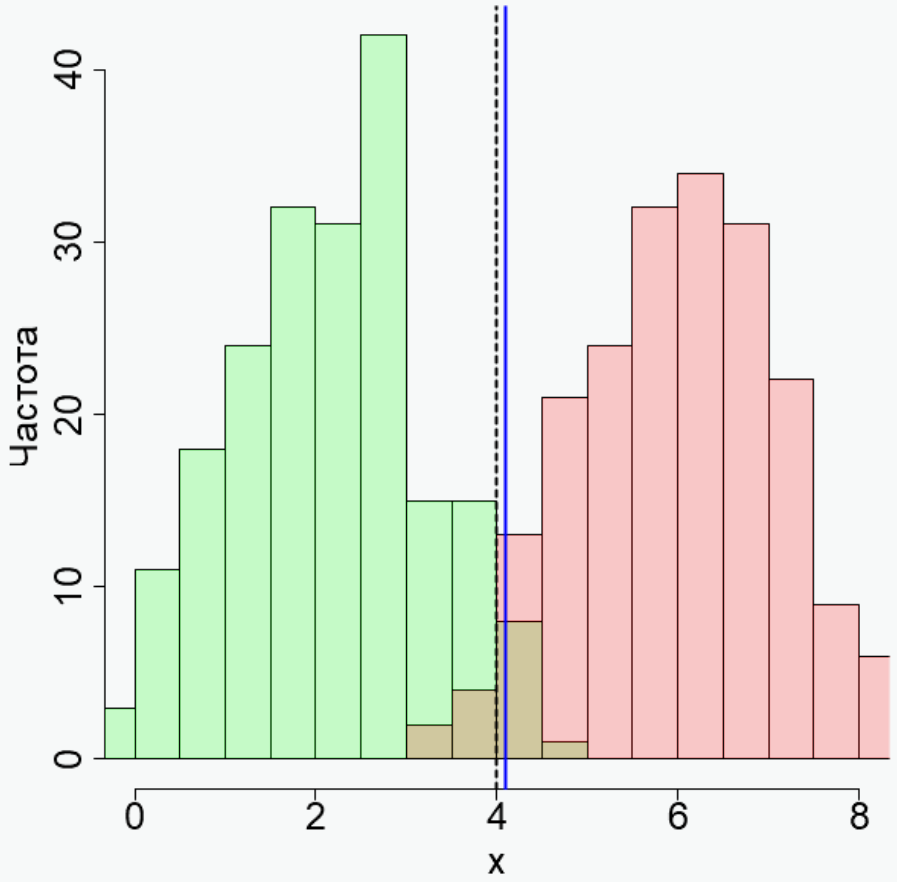
Если $k = 2$ и $\pi_1 = \pi_2$,
решающая граница:

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Пусть $\mu_1 = 2$, $\mu_2 = 6$:

$$x = \frac{2 + 6}{2} = 4$$

Дискриминантная функция по одному признаку



Оценки по выборке:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\mu}_1 = 2.15, \hat{\mu}_2 = 6.03$$

$$x = \frac{2.15 + 6.03}{2} = 4.09$$

Другие оценки по выборке

Априорные вероятности классов: $\hat{\pi_k} = \frac{n_k}{n}$

Дисперсия: $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$

Дискриминантная функция линейна по x:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \to \max_k$$

Дискриминантная функция по двум и более признакам

Функция плотности вероятности для классов – многомерный нормальный закон с ковариационной матрицей объясняющих переменных Σ, *общей для всех классов*.

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \to \max_k$$

План лекции

- Логистическая регрессия
- Линейный дискриминантный анализ
- Порог отсечения классов и ROC-кривая
- Квадратичный дискриминантный анализ
- Методы создания повторных выборок

Данные Default: default = Yes против default = No

LDA, default зависит от balance и student

```
## Prior probabilities of groups:

##    No   Yes
## 0.968 0.032

## Group means:

##      balance   income
## No   806.240 33633.89
## Yes 1738.497 32919.88

## Coefficients of linear discriminants:

##              LD1
## balance 0.00223
## income  0.00001
```

Данные `Default`: `default = Yes` против `default = No`

	No	Yes
No	8208	17
Yes	216	59

Наличие признака: **неплательщик** (`default = Yes`).

Чувствительность: $TPR = \frac{59}{216+59} = 0.215$ – доля истинных неплательщиков, обнаруженных классификатором.

Специфичность: $SPC = \frac{8208}{8208+17} = 0.998$ – доля правильно идентифицированных клиентов, уплативших долг.

Ошибка: $1 - Acc = 1 - \frac{8208+59)}{(8208+17+216+59)} = 0.027$

Ошибка нулевого классификатора (`default = Yes`) *равна 0.033 и немногим больше ошибки LDA*

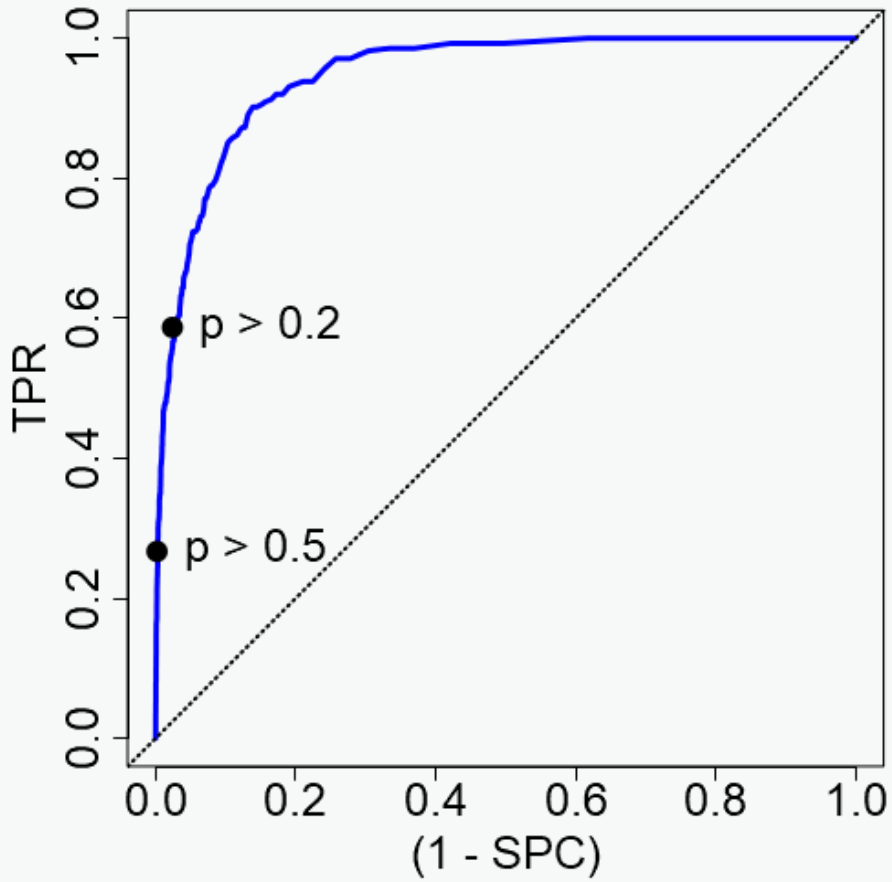
Пример на данных `Default` В текущей модели неплательщик – клиент, для которого вероятность дефолта превышает 0.5.

Чтобы повысить чувствительность модели, снизим порог отсечения, например: $P(\text{default} = \text{Yes}|X = x) > 0.2$. Тогда:

	No	Yes
No	8034	191
Yes	122	153

Чувствительность: $TPR = \frac{153}{122+153} = 0.556$.
Специфичность: $SPC = \frac{8034}{8034+191} = 0.977$.

ROC-кривая



ROC – "радиочастотная характеристика приёмника" (*receiver operating characteristic*) – исторически сложившийся термин.

Кривая отражает зависимость чувствительности (TPR) от специфичности ($1 - SPC$) в зависимости от порога отсечения вероятности (p).

AUC (*area under the curve*) – площадь под кривой, или общее качество классификатора.

Если $AUC = 0.5$ для бинарного классификатора, то это простое случайное угадывание, ROC-кривая – биссектриса первой четверти.

План лекции

- Логистическая регрессия
- Линейный дискриминантный анализ
- Порог отсечения классов и ROC-кривая
- Квадратичный дискриминантный анализ
- Методы создания повторных выборок

Квадратичный дискриминантный анализ (QDA)

Допущения:

- наблюдения в каждом классе распределены по нормальному закону: $X \sim N(\mu_k, \Sigma_k)$;
- ковариационные матрицы для классов отличаются: $\Sigma_i \neq \Sigma_j \ \forall i \neq j, \ i, j = 1, \ldots, k$.

Дискриминирующая функция:

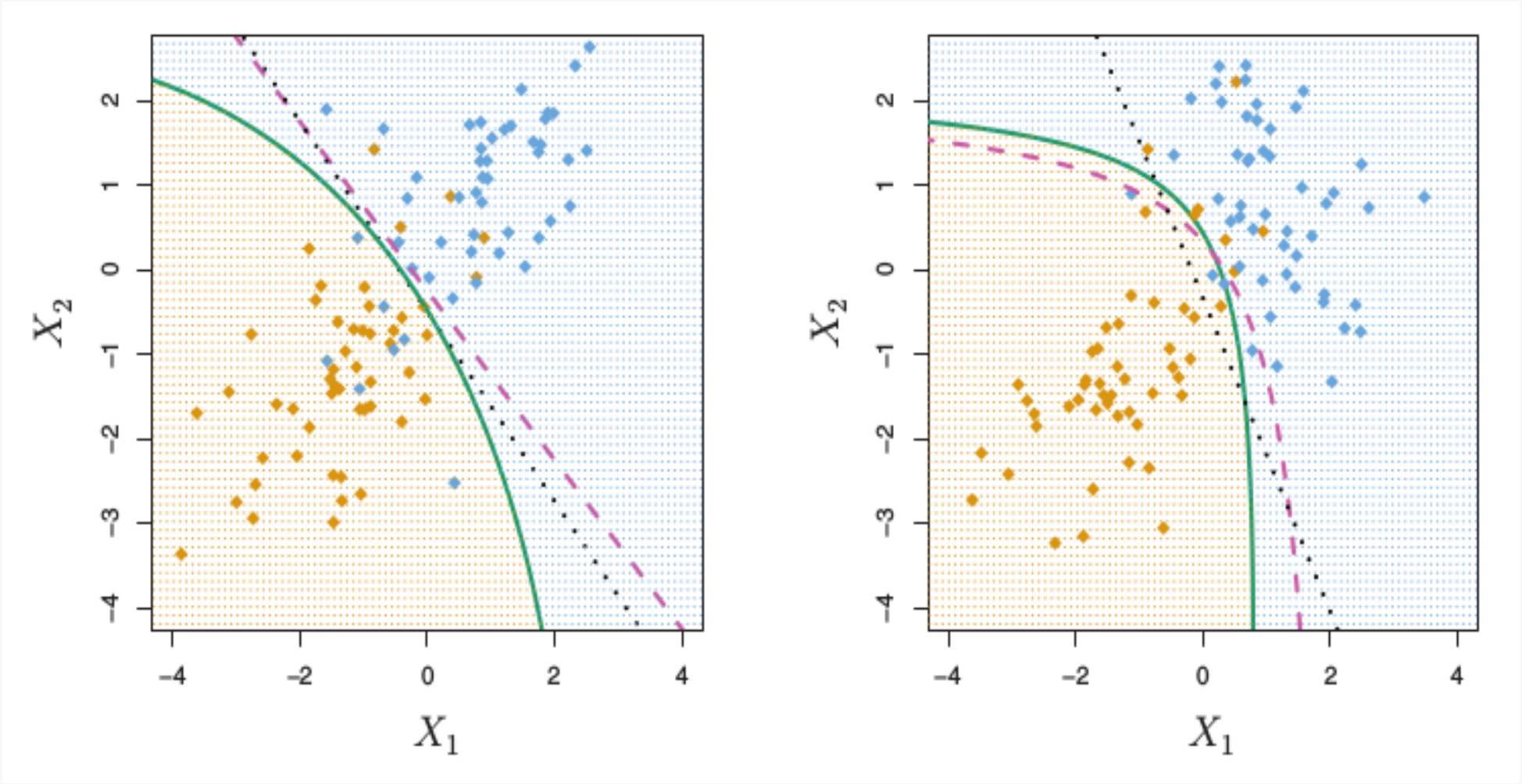
$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2}\log|\Sigma_k| +$$
$$+\log\pi_k \to \max_k$$

LDA или QDA?

LDA оценивает единую Σ, число параметров для оценки: $p(p+1)/2$.

LDA оценивает свою Σ_k для каждого класса, число параметров для оценки: $Kp(p+1)/2$.

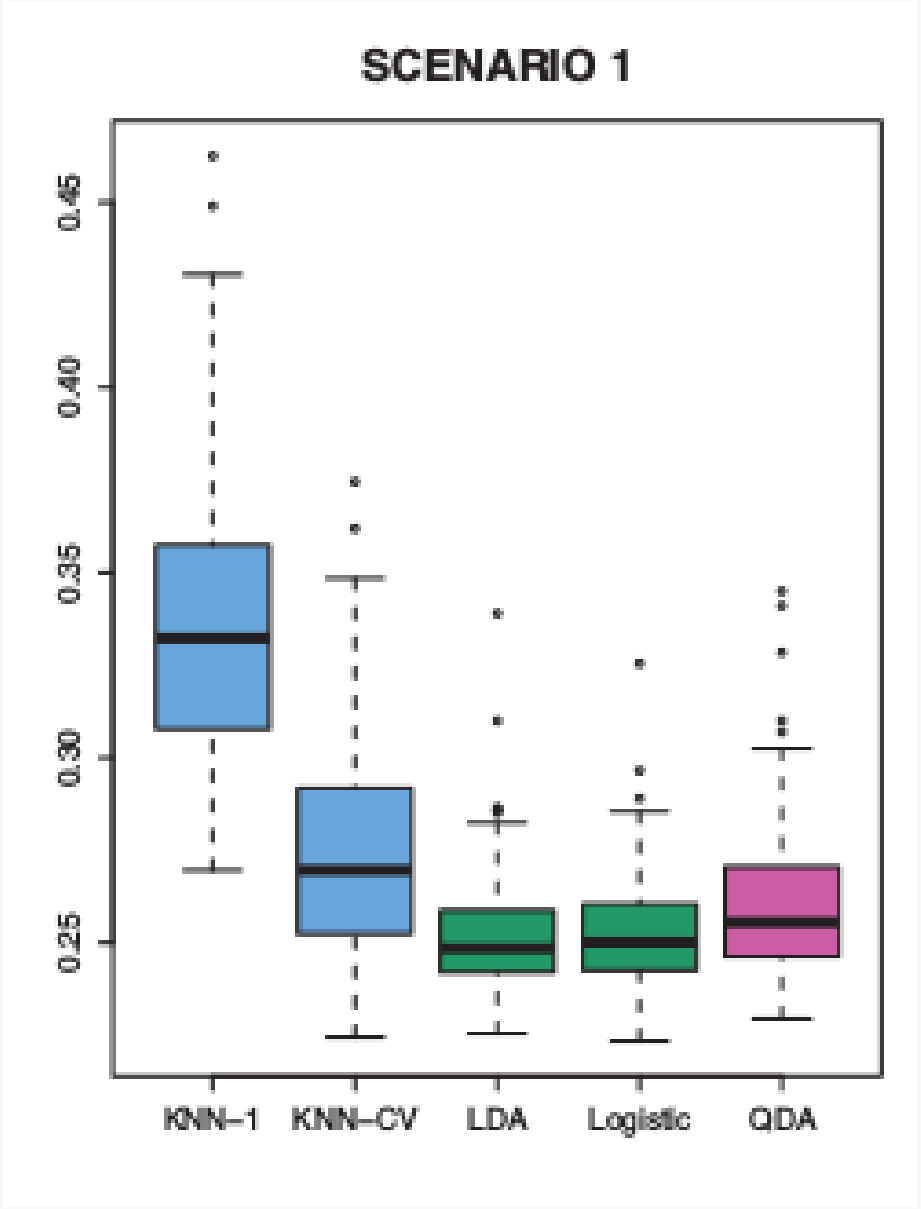
- LDA требует меньше вычислений.
- Гибкость QDA выше.
- Дисперсия LDA меньше, чем QDA.
- Если в действительности ковариационные матрицы классов разные, у LDA выше смещение.



Зелёная линия – граница по QDA, чёрный пунктир – LDA,

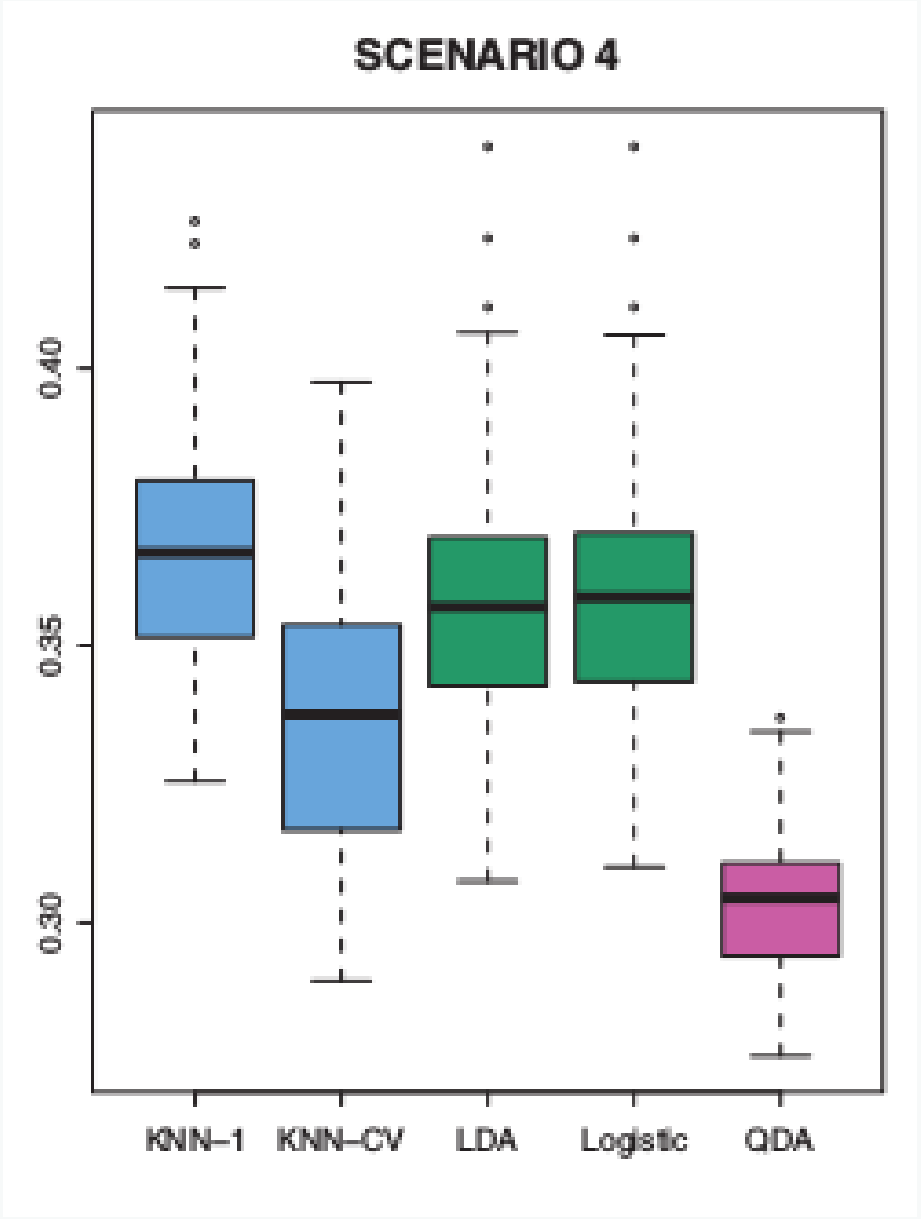
Фиолетовый пунктир – байесовский классификатор (эталон)

Слева: $\Sigma_1 = \Sigma_2$**. Справа:** $\Sigma_1 \neq \Sigma_2$**.**



Сценарий 1:
В каждом классе наблюдения распределены по нормальному закону; разные средние; классы независимы друг от друга.

$p = 2$, 100 случайных выборок, кросс-валидация.
На графике – разброс частот ошибок для разных методов



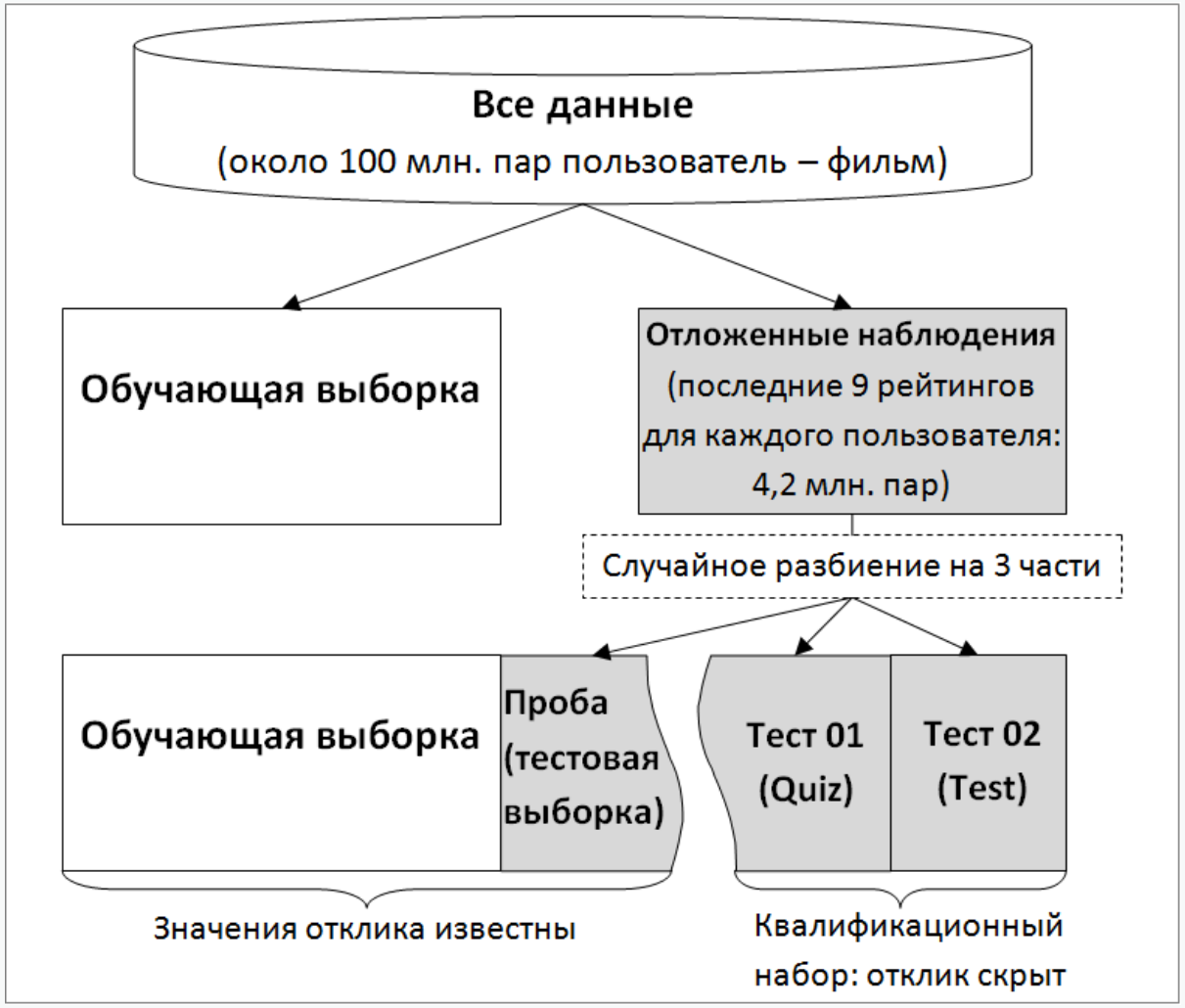
Сценарий 4:
Предикторы распределены по нормальному закону с разными корреляционными матрицами для каждого класса.

$p = 2$, 100 случайных выборок, кросс-валидация.
На графике -- разброс частот ошибок для разных методов

План лекции

- Логистическая регрессия
- Линейный дискриминантный анализ
- Порог отсечения классов и ROC-кривая
- Квадратичный дискриминантный анализ

- Методы создания повторных выборок



Соревнование Netflix на лучший алгоритм коллаборативной фильтрации. Источник: [5]

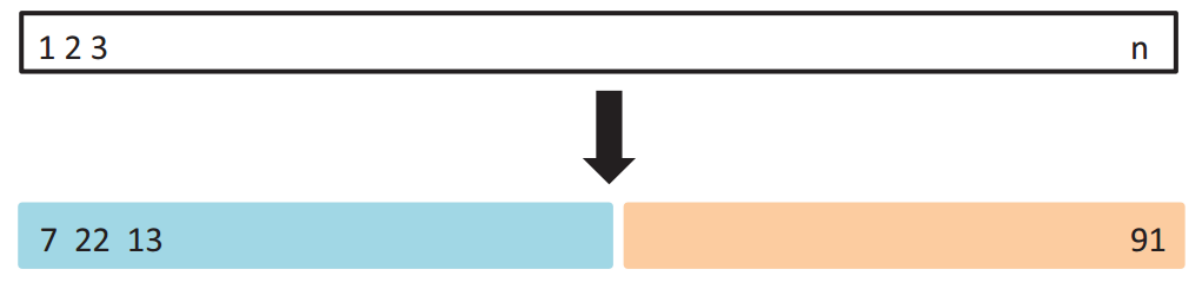
Ключевые идеи

- Цель любой модели – подстроиться под данные, поэтому оценка точности на обучающей выборке слишком оптимистична
- Более правдоподобная оценка точности должна проводиться на независимом от обучающей выборки наборе данных – тестовой выборке

Проблемы:

- Как оценить оценки точность и гибкость *при построении модели*?
- Как оценить качество модели, если подходящей тестовой выборки вообще нет?

Метод проверочной выборки

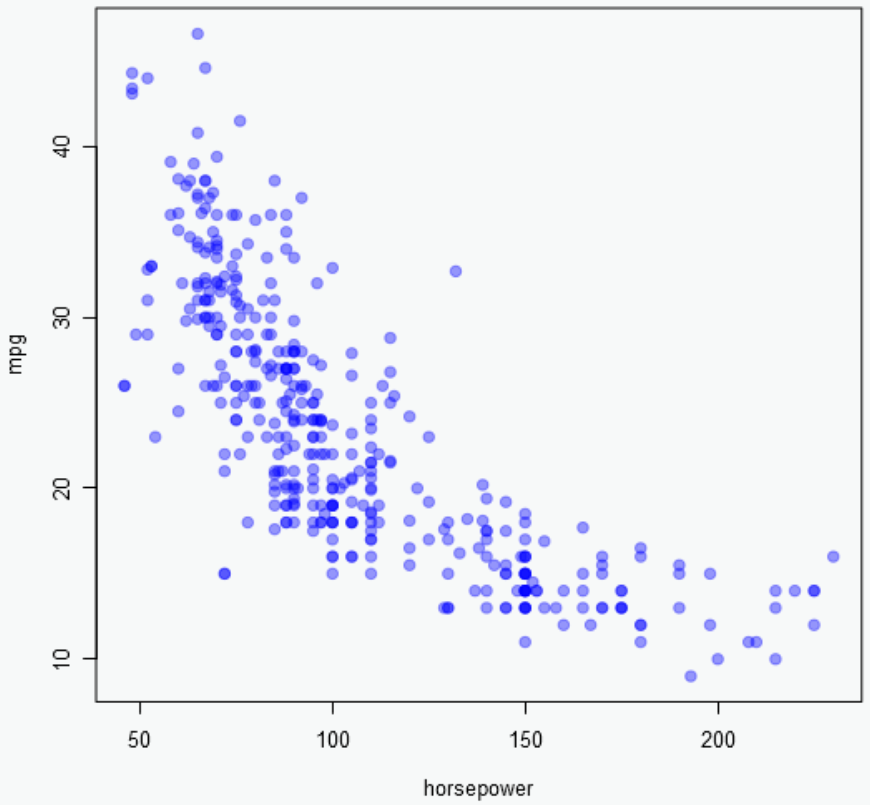


Преимущество: **прост в использовании**

Недостатки:

- MSE на проверочной выборке может оказаться очень вариабельной
- На небольших выборках MSE на проверочной выборке оказывается переоцененной

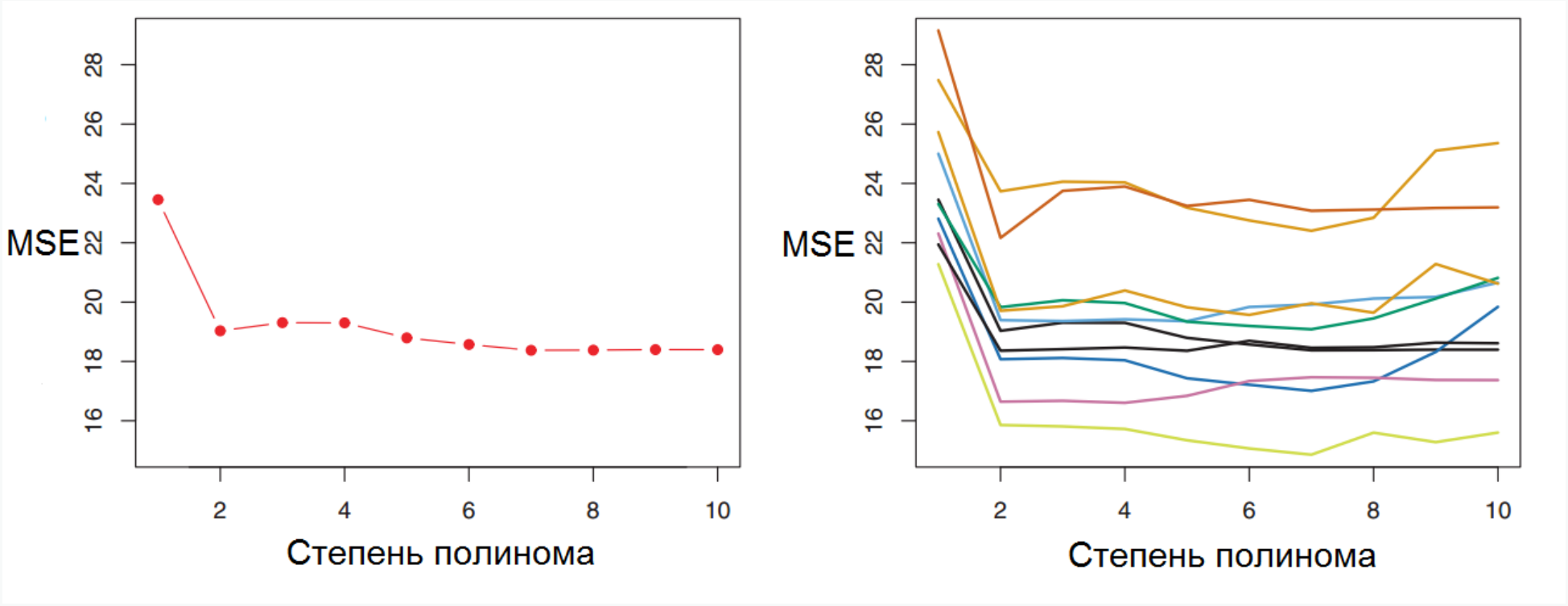
Пример на данных `Auto {ISLR}`



Связь `mpg` и `horsepower` нелинейна. Сравнить модели методом проверочной выборки:

1. $$\text{mpg} = \alpha_0 + \alpha_1 \cdot \text{horsepower}$$
2. $$\text{mpg} = \alpha_0 + \alpha_1 \cdot \text{horsepower} + \\ + \alpha_2 \cdot \text{horsepower}^2$$

Пример на данных `Auto {ISLR}`



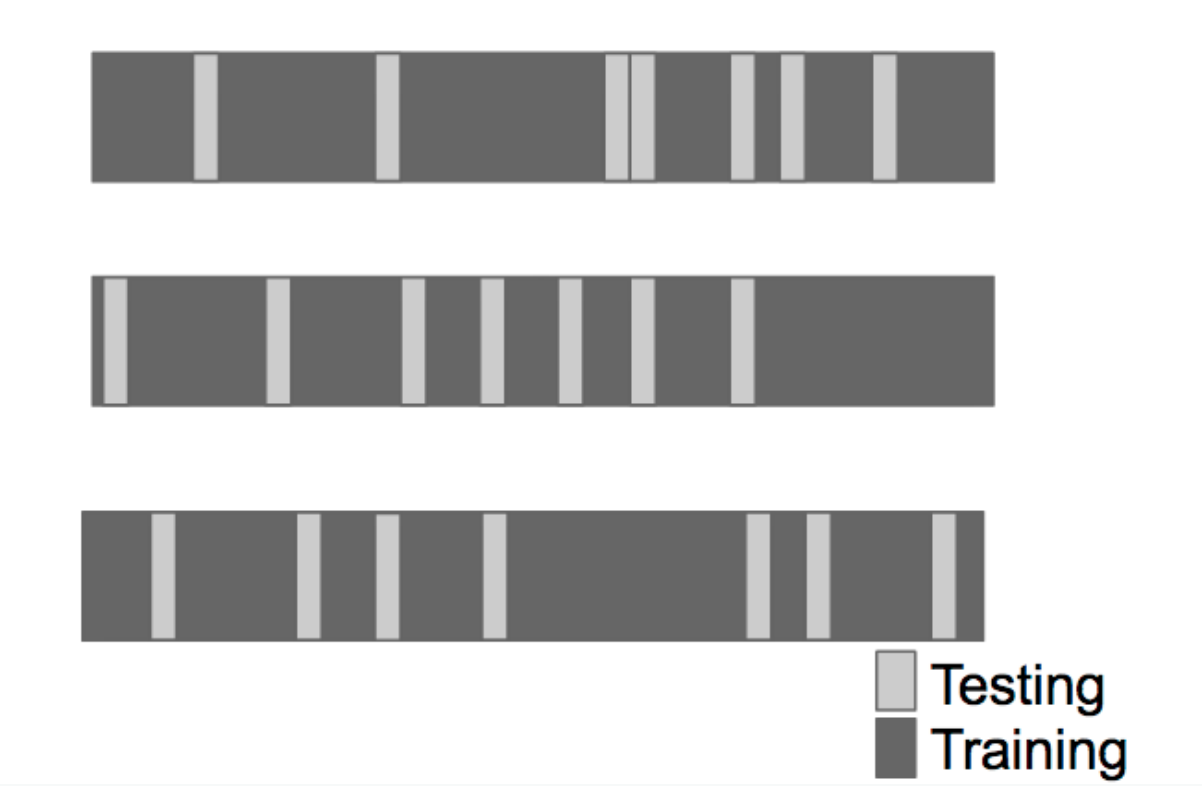
Слева: однократное разбиение на обучающую и проверочную выборки
Справа: метод проверочной выборки применён 10 раз

Алгоритм перекрёстной проверки

1. Берём обучающую выборку. $i = 1$.
2. Разбиваем на две части: обучающую- i и тестовую- i
3. Строим модель на обучающей- i
4. Оцениваем точность на тестовой- i
5. $i + 1$.
6. Повторяем 2...5 (много) раз.
7. Усредняем оценки точности на тестовых- i выборках.

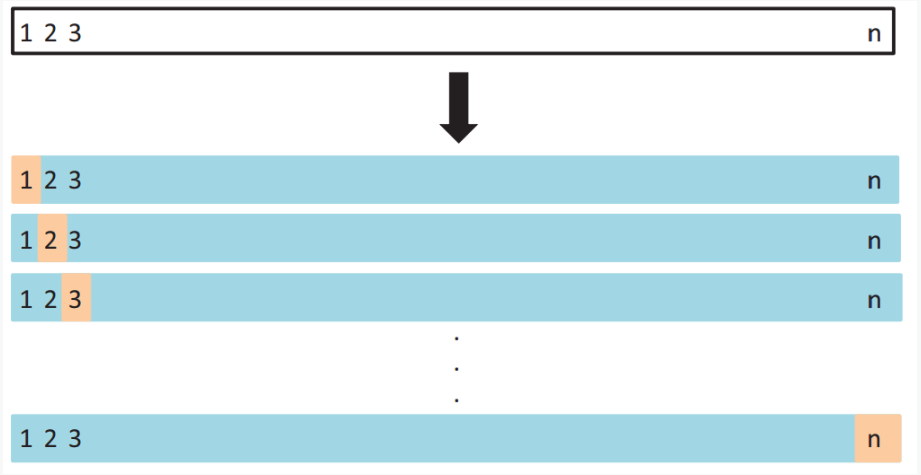
Как действовать на шаге 2?

Случайные бесповторные подвыборки



Повторение метода проверочной выборки n раз, со всеми недостатками.

Перекрёстная проверка по отдельным наблюдениям (LOOCV*)



$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

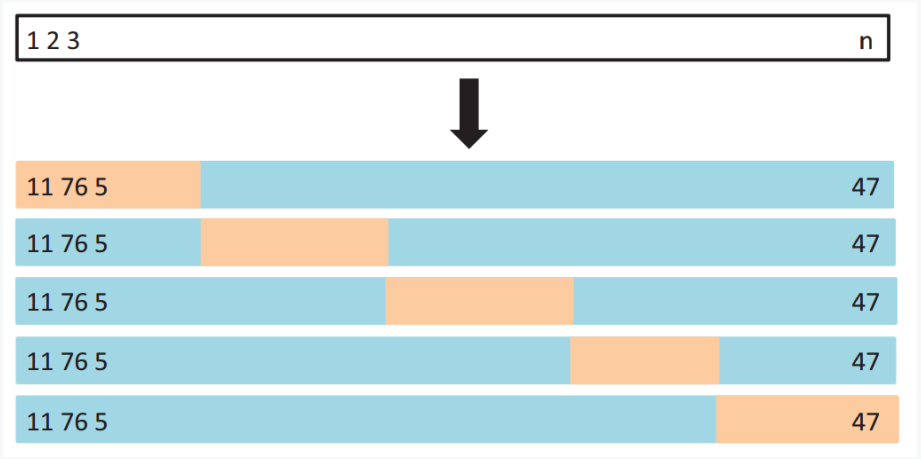
*leave-one-out cross validation

- Метод пытается преодолеть недостатки метода проверочной выборки.
- Ресурсоёмкий: модель нужно прогнать n раз.
- Для линейной и полиномиальной регрессии *с параметрами, оцененными по МНК*, можно использовать экономную модификацию:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

где $\hat{y}_i$ – модельное значение отклика, $h_i \in \left[\frac{1}{n}, 1\right]$ – показатель разбалансировки (Leverage) для i-го наблюдения.

k-кратная перекрёстная проверка*

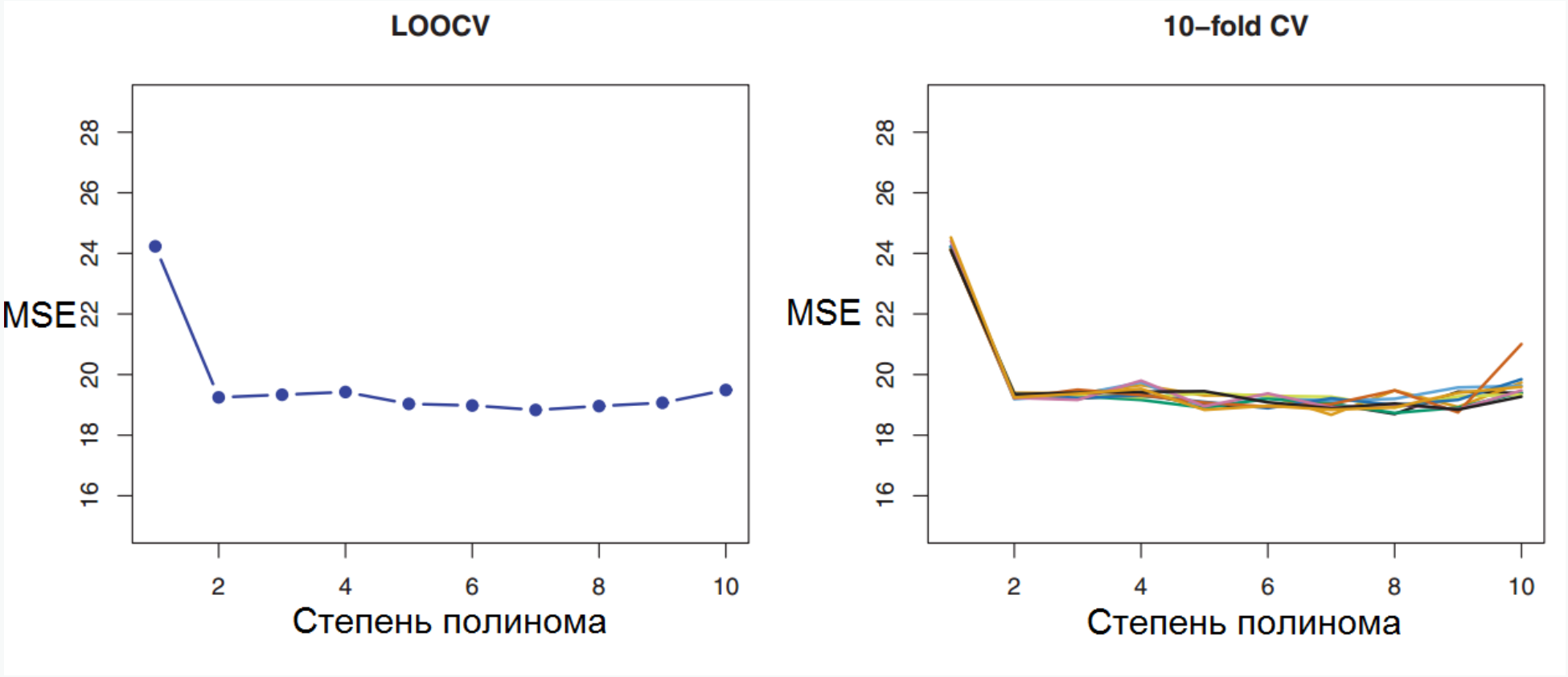


$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

*k-fold validation

- LOOCV обладает более высокой дисперсией, чем k-fold, из-за взаимной корреляции оценок MSE при LOOCV
- Смещение у k-fold выше, чем у LOOCV, и ниже, чем у метода проверочной выборки
- Установлено, что $k = 5$ и $k = 10$ дают разумный компромисс между смещением и дисперсией
- Ресурсоёмкость у k-fold средняя

Пример на данных `Auto {ISLR}`



Слева: кривая ошибок LOOCV Справа: 10-кратная перекрёстная проверка повторена 9 раз, каждый раз новое разбиение

О чём важно помнить

- оценку точности сильно искажают нетипичные наблюдения
- важно обеспечить репрезентативность данных в блоках скользящего контроля
- если в данных присутствует фактор времени, нужно следить за тем, чтобы признаки, доступные в будущем, не использовались для предсказания прошлого
- чем выше число блоков, тем точнее оценка ошибки и тем затратнее скользящий контроль

Источники

1. *Джеймс Г.*, *Уиттон Д.*, *Хасти Т.*, *Тибширани Р.* Введение в статистическое обучение с примерами на языке R. Пер. с англ. С.Э. Мастицкого – М.: ДМК Пресс, **2016** – 450 с.

2. Данные `Default`, `Auto` из пакета `ISLR`.

3. *Shireen Elhabian*, *Aly A. Farag* A Tutorial on Data Reduction. Linear Discriminant Analysis (LDA). URL: http://www.sci.utah.edu/~shireen/pdfs/tutorials/Elhabian_LDA09.pdf

4. *W.N. Venables*, *B.D. Ripley* Modern Applied Statistics with S. **2002**. URL: https://www.researchgate.net/publication/224817420_Modern_Applied_Statistics_With_S

5. *Jeffrey Leek*. Материалы курса «Practical Machine Learning» Университета Джонса Хопкинса на портале coursera.org, доступные в репозитории: github.com/jtleek/modules/tree/master/08_PredictionAndMachineLearning

6. *Хенрик Бринк*, *Джозеф Ричардс*, *Марк Феверолф* Машинное обучение. Спб.: Питер, **2018**. -- 336 с.