



# Машинное обучение в науках о Земле

Михаил Криницкий

К.Т.Н.,  
зав. Лабораторией машинного обучения в науках о Земле МФТИ  
с.н.с. Институт океанологии РАН им. П.П. Ширшова



# Задачи классификации

Михаил Криницкий

К.Т.Н.,  
зав. Лабораторией машинного обучения в науках о Земле МФТИ  
с.н.с. Институт океанологии РАН им. П.П. Ширшова

# ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

## 1. формулировка задачи:

- какой тип (классификация, регрессия, другой)? Или переформулировать в легко решаемый тип!
- определиться, что есть объекты (события)
- определиться, что есть целевая переменная
- определить признаковое описание объектов (событий)
- определить критерии качества решения задачи (MSE, MAE, pattern correlation, etc.)

# ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

2. формулировка модели:
  - задать вид модели (линейная регрессия, дерево решений, композиционный алгоритм, нейронная сеть, etc.)
  - задать сложность модели (задается гиперпараметрами – настройками модели)
  - определиться с функцией потерь (MSE, MAE, LogLikelihood, etc.)
3. подготовка данных или генератора данных:
  - стандартизировать данные (если нужно)
  - обработать пропуски, категориальные значения, подготовить кодирование текста, применить понижение размерности данных
  - оставить часть данных для проверки качества (train-test split)
  - подготовить генератор данных с учетом стратегии скользящего контроля (cross-validation quality estimation)
4. оптимизация модели на обучающей выборке:
  - $\hat{p} = \underset{\mathbb{P}}{\operatorname{argmin}}(L(\vec{p}, \mathcal{T}))$

# ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

5. оценка модели:
  - оценить качество по метрикам, определенным на этапе **1.** на тестовой выборке
  - оценить неопределенности параметров модели (если возможно)
  - оценить неопределенности оценок целевой переменной
  - определить наличие недообучения или переобучения
  - оценить соотношение сложности модели и сложности закономерностей в данных; при неадекватной сложности модели вернуться к **п.2**
6. применение модели на вновь получаемых данных:
  - оценка распределения вновь получаемых данных: генерируются ли они из того же распределения, что и обучающая выборка?
  - предобработка новых данных идентично **п.3** с точностью до коэффициентов стандартизации и деталей способов предобработки
  - применение модели к предобработанным новым данным для получения значений целевой переменной
  - построение научных выводов

# ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

7. публикация результатов:

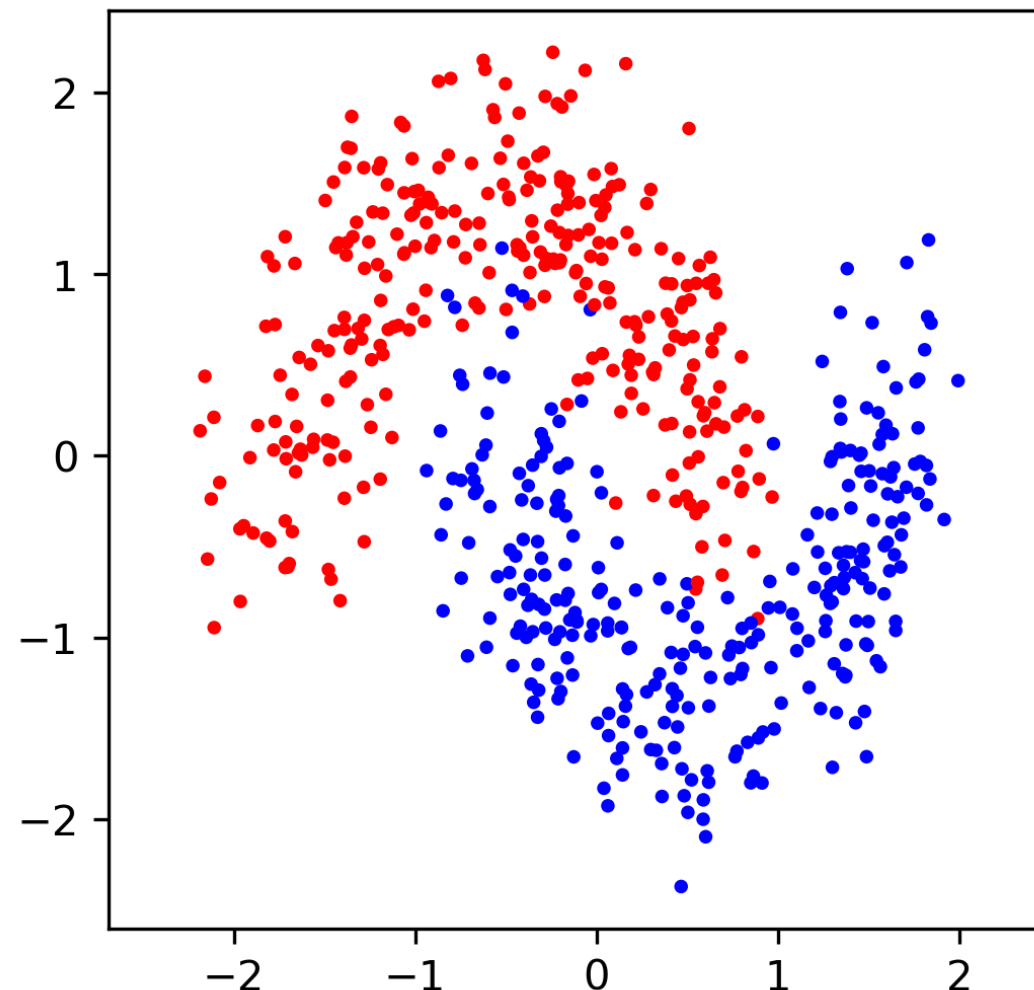
- описание результатов в виде статей, отчетов об исследованиях
- защита результатов перед лицом научного сообщества
- получение наград, признание успехов, etc.

# КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

## Формулировка задачи (в терминах машинного обучения)

- «Обучение с учителем»
  - восстановление регрессии
  - классификация

**что я хочу?** — метку класса  
**«красный или синий?»**  
(бинарная классификация)



# ЗАДАЧА КЛАССИФИКАЦИИ

**цель** – метка класса ( $y$  – категориальная переменная)

«спам / не-спам»

«мезоциклон / не-мезоциклон»

«превышает порог / не превышает»

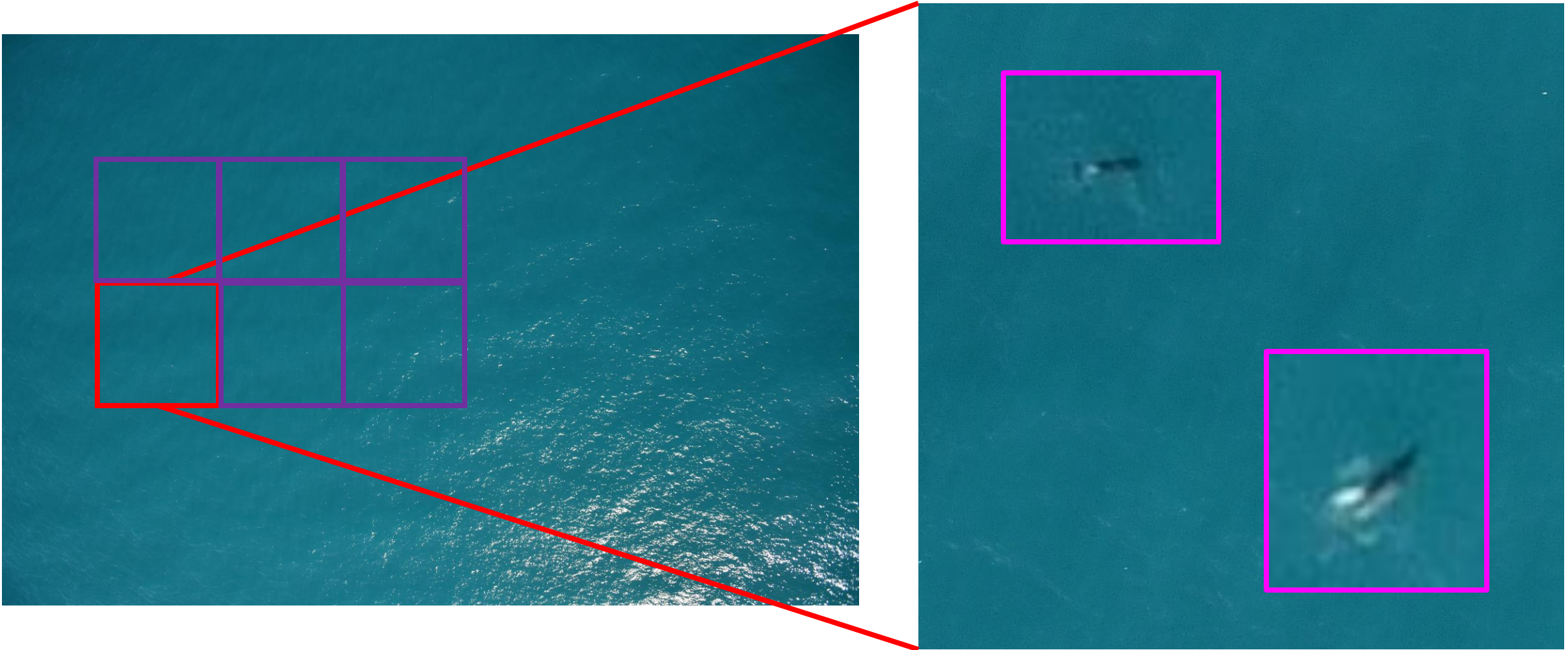
«кот / собака / лошадь»

«0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9»



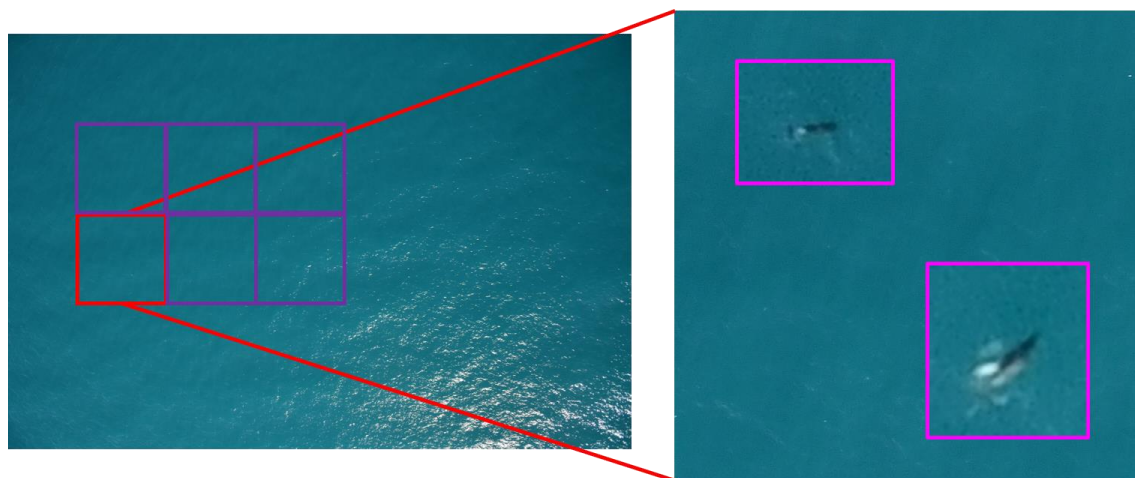
# ЗАДАЧА КЛАССИФИКАЦИИ

**Пример:** классификация участков учетных снимков поверхности океана в отношении признака наличия морских млекопитающих



# ЗАДАЧА КЛАССИФИКАЦИИ

**Пример:** классификация участков учетных снимков поверхности океана в отношении признака наличия морских млекопитающих



- Что есть **объекты** выборки?
- Что есть **признаковое описание**?
- Что есть **целевая переменная**?
- Тип целевой переменной?
- Размерность ц.п.?
- **Мера качества** решения?

# ЗАДАЧА КЛАССИФИКАЦИИ

Простейший пример:

объекты описываются действительным признаком  $x$   
целевая переменная  $y$  – бинарная, классы:  $A$ ,  $B$ ; по 1000 экземпляров каждого класса  
пусть для класса  $y = A$  значения  $x \sim \mathcal{N}(\mu_A, \sigma_A)$ , для класса  $y = B$  значения  $x \sim \mathcal{N}(\mu_B, \sigma_B)$

Базируясь на этих данных, каково должно быть  
решение (значение  $y$ ) при:

$$x = -10$$

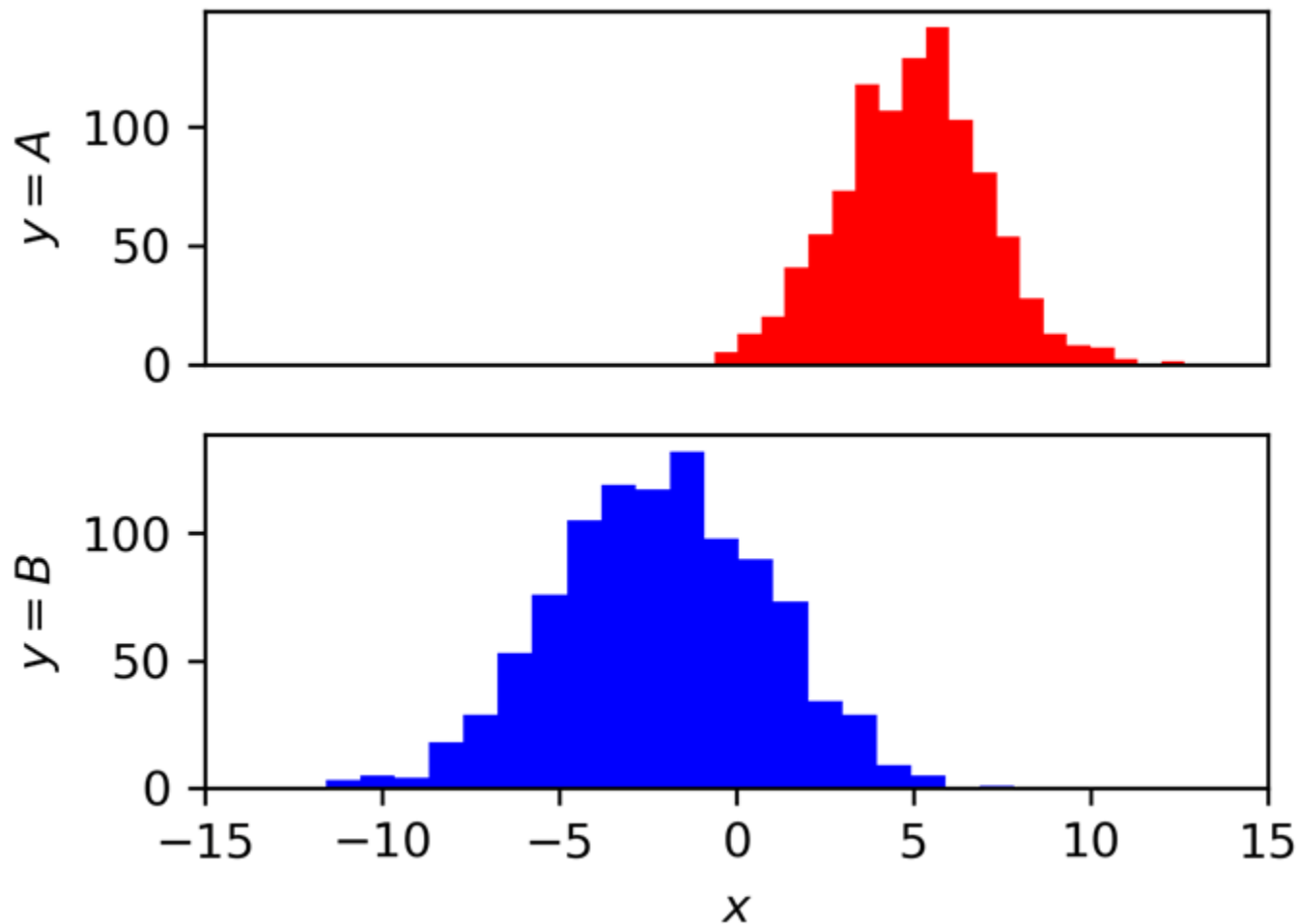
$$x = -5$$

$$x = 2$$

$$x = 5$$

$$x = 10$$

$$x = 15$$

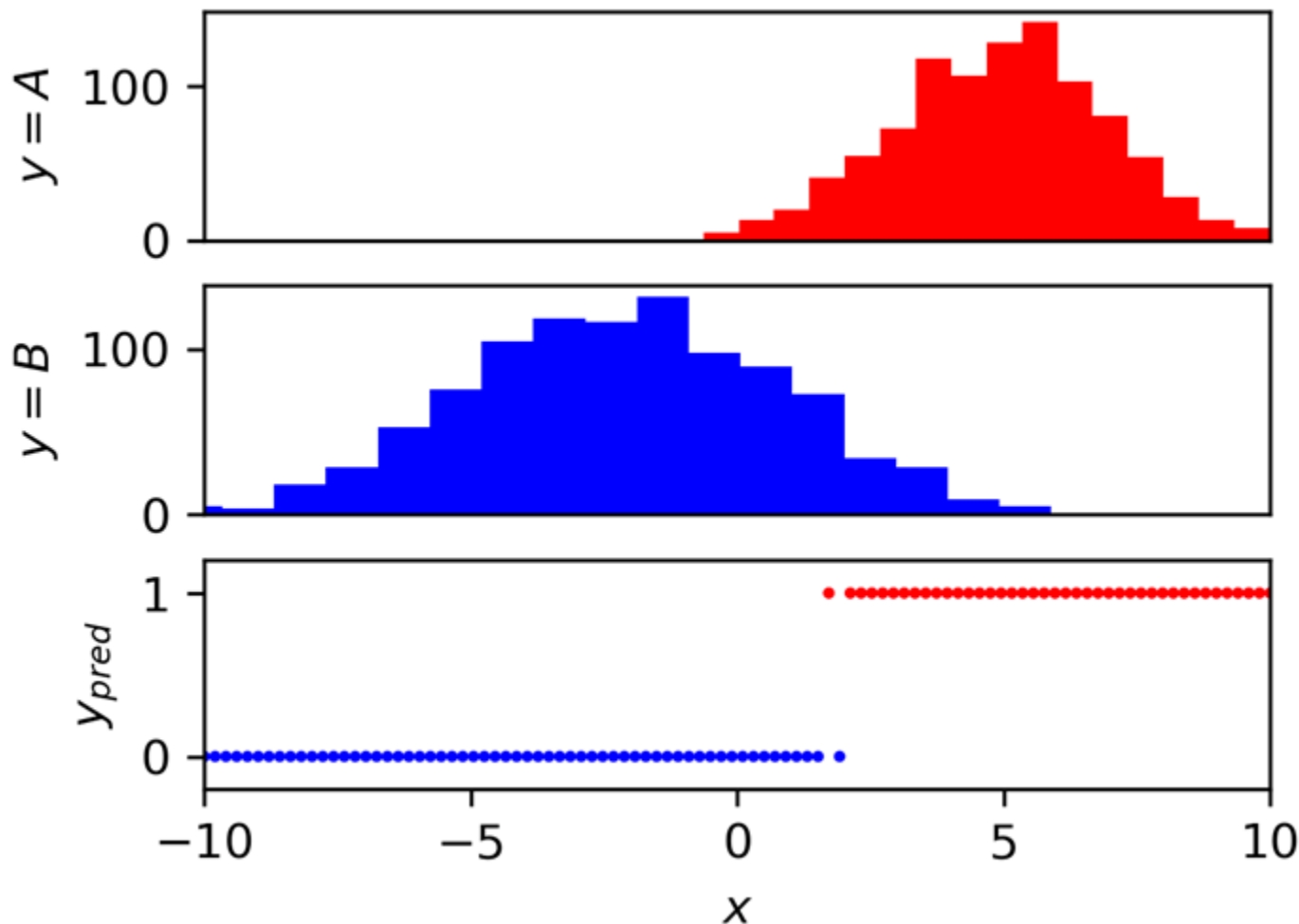


# ЗАДАЧА КЛАССИФИКАЦИИ

Простейший пример: объекты описываются действительным признаком  $x$   
целевая переменная  $y$  – бинарная, классы:  $A$ ,  $B$ ; по 1000 экземпляров каждого класса  
пусть для класса  $y = A$  значения  $x \sim \mathcal{N}(\mu_A, \sigma_A)$ , для класса  $y = B$  значения  $x \sim \mathcal{N}(\mu_B, \sigma_B)$

Подход №1: **KNN** (метод  $K$  ближайших соседей)

1. выбрать  $K$  ближайших соседей для нового объекта (! нужно определить меру близости !)
2. осреднить (можно с разными весами) целевую переменную по этим объектам («простое голосование», «majority vote» или «взвешенное голосование», «weighted vote»)
3. считать полученный результат значением целевой переменной на новом объекте



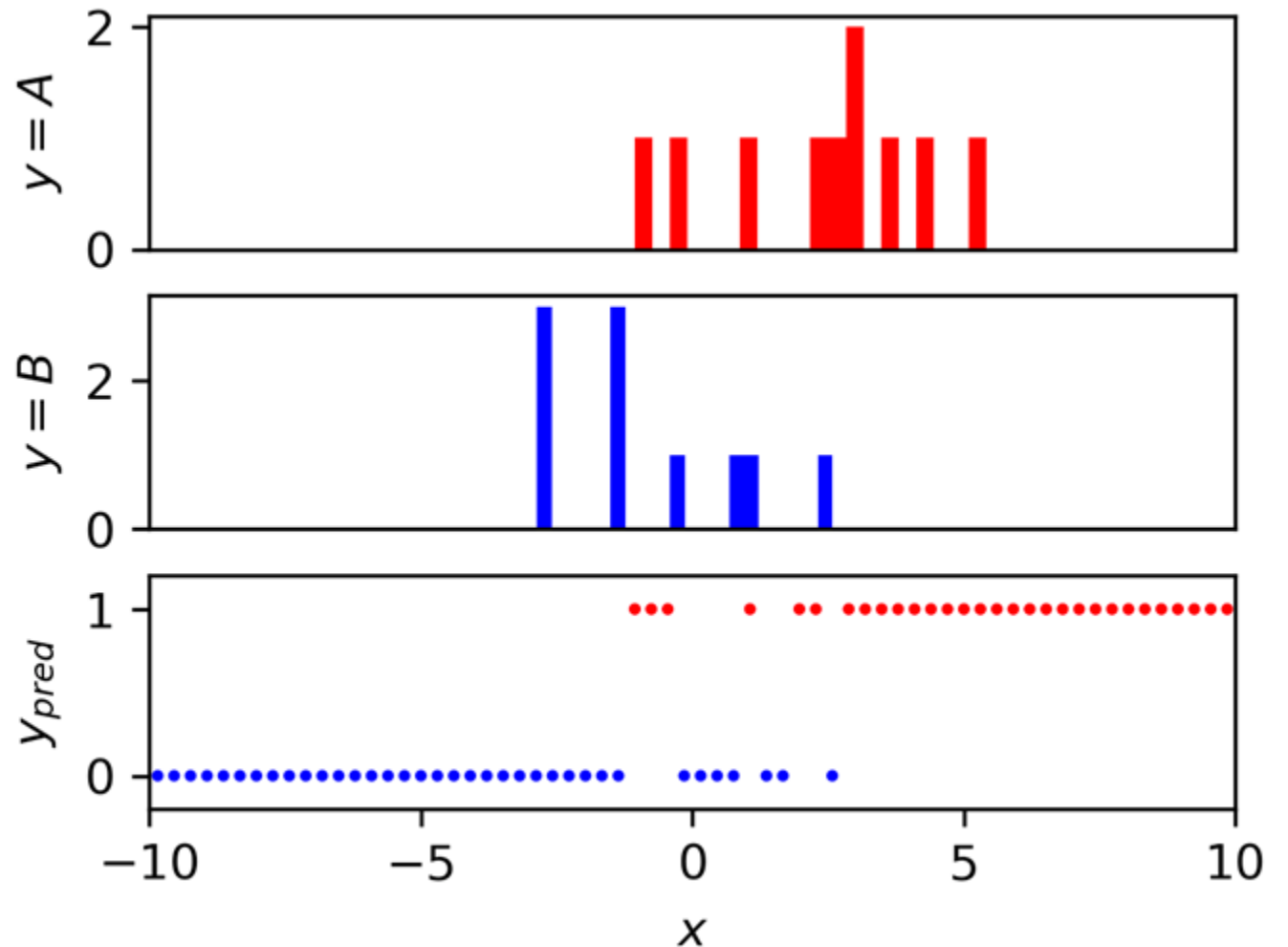
# ЗАДАЧА КЛАССИФИКАЦИИ

Подход №1: **KNN** (метод  $K$  ближайших соседей)

- простой
- быстрый
- легко настраивается. Гиперпараметр  $K$  регулирует «сложность» модели

## **А ЧТО ЕСЛИ ДАННЫХ МАЛО?..**

- требуется большое количество обучающих данных
- обучающие данные должны быть распределены достаточно плотно в исследуемой области  $x$
- не обобщает закономерности в данных



# ЗАДАЧА КЛАССИФИКАЦИИ

Подход получше – оценить **вероятность** классов **A** и **B** для объекта, описываемого значением  $x$ .

$$P(Y = k | X = x)$$

# ЗАДАЧА КЛАССИФИКАЦИИ

Подход получше – оценить **вероятность** классов  $A$  и  $B$  для объекта, описываемого значением  $x$ .

$$P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$$

Кстати, если нужно принять решение относительно значения  $Y$  при определенном значении  $x_i$ , помним, что  $P(x_i)$  – константа, которую можно не учитывать при сравнении  $P(Y = \textcolor{red}{A}|X = x_i)$  и  $P(Y = \textcolor{blue}{B}|X = x_i)$

$$P(X) = \sum_{y_i} P(X|Y = y_i)P(Y = y_i)$$

формула полной вероятности

# ЗАДАЧА КЛАССИФИКАЦИИ

Подход получше – оценить **вероятность** классов  $A$  и  $B$  для объекта, описываемого значением  $x$ .

$$P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$$

Кстати, если нужно **принять решение** относительно значения  $Y$  при определенном значении  $x_0$ , помни, что  $P(x_0)$  – константа, которую можно не учитывать при сравнении  $P(Y = A|X = x_0)$  и  $P(Y = B|X = x_0)$

ЕСЛИ нам повезло и МЫ ЗНАЕМ (или полагаем как допущение в процессе решения) распределения  $X$  для каждого из классов  $P(X|Y = A)$ ,  $P(X|Y = B)$  etc., - то можно получить **аналитическое решение!**

И это решение будет ЛУЧШИМ из всех возможных.



# ЗАДАЧА КЛАССИФИКАЦИИ

$$P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$$

ЕСЛИ нам повезло и МЫ ЗНАЕМ (или полагаем как допущение в процессе решения) распределения  $X$  для каждого из классов  $P(X|Y = A)$ ,  $P(X|Y = B)$  etc., - то можно получить **аналитическое решение!**

## previously on ML4ES

объекты описываются действительным признаком  $x$   
целевая переменная  $y$  – бинарная

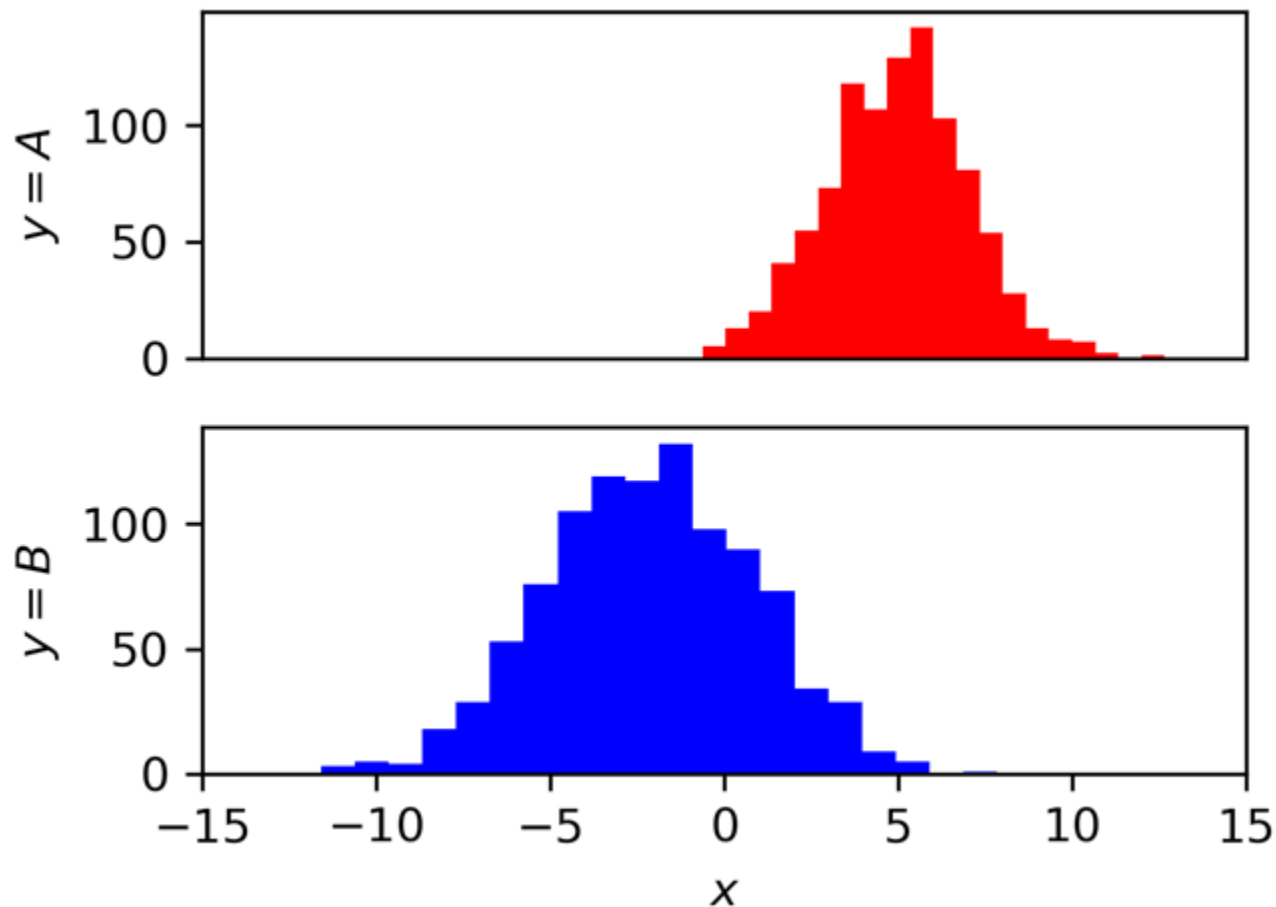
пусть для класса  $y = A$  значения  $x \sim \mathcal{N}(\mu_A, \sigma_A)$ , для  
класса  $y = B$  значения  $x \sim \mathcal{N}(\mu_B, \sigma_B)$

$$\mu_A = 5$$

$$\mu_B = -2$$

$$\sigma_A = 2$$

$$\sigma_B = 3$$



# ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

$$P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$$

ЕСЛИ нам повезло и МЫ ЗНАЕМ (или полагаем как допущение в процессе решения) распределения  $X$  для каждого из классов  $P(X|Y = \textcolor{red}{A})$ ,  $P(X|Y = \textcolor{blue}{B})$  etc., - то можно получить **аналитическое решение!**

$$P(Y = \textcolor{blue}{B}|X = x) = \frac{e^{-\frac{(x+2)^2}{2 \cdot 9}} \cdot \frac{1}{2}}{e^{-\frac{(x-5)^2}{2 \cdot 4}} \cdot \frac{1}{2} + e^{-\frac{(x+2)^2}{2 \cdot 9}} \cdot \frac{1}{2}}$$

«Байесовский классификатор»

(не путать с «naïve bayes»)

