Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н., с.н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

Задачи классификации

Михаил Криницкий

к.т.н., с.н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

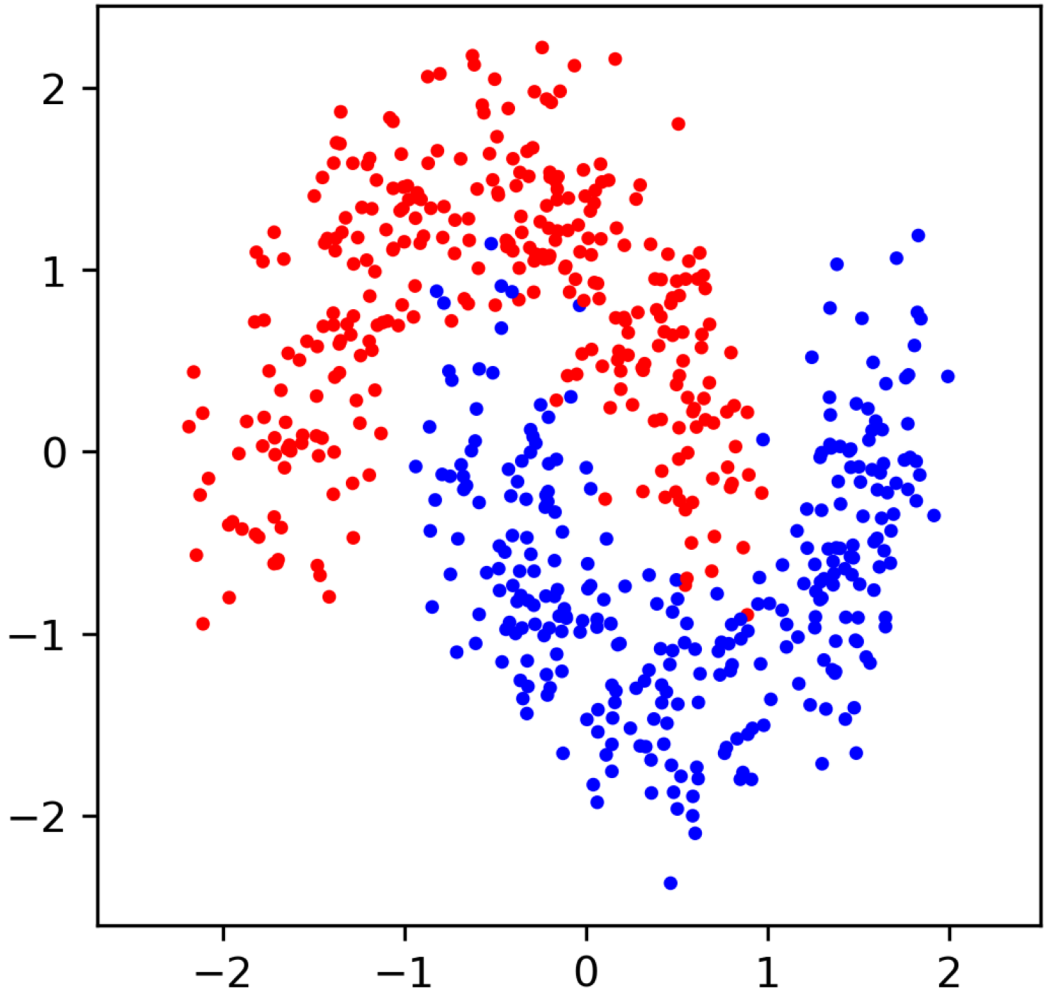
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»
  - восстановление регрессии
  - классификация

что я хочу? – метку класса

«красный или синий?»

(бинарная классификация)



ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

цель – метка класса (y – категориальная переменная)

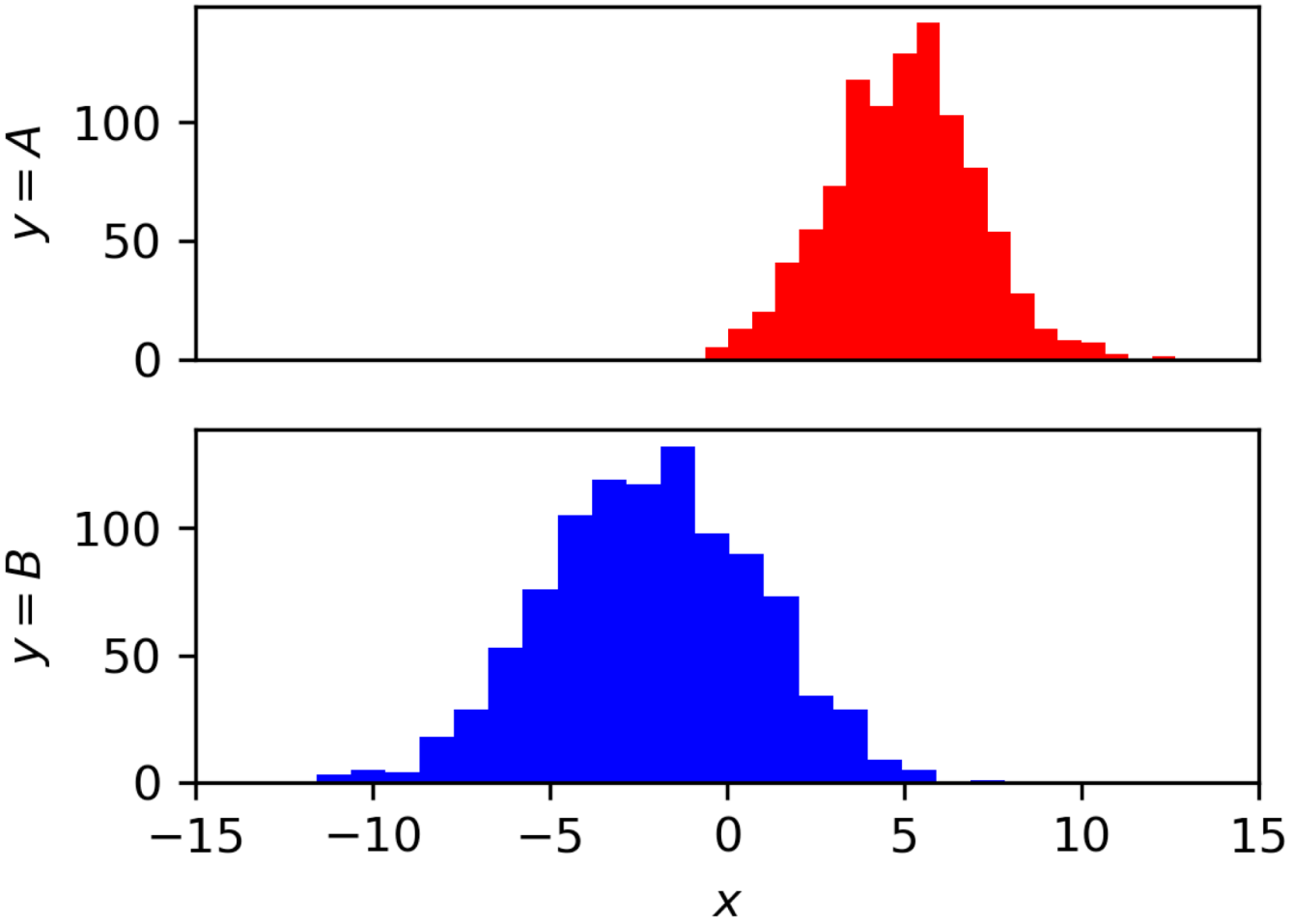
«спам / не-спам»	y – категориальная, бинарная
«мезоциклон / не-мезоциклон»	y – категориальная, бинарная
«кот / собака / лошадь»	y – категориальная, 3 класса
«0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9»	y – категориальная, 10 классов
«есть дельфин / нет дельфина»	y – категориальная, бинарная

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

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

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

$x = -10$
$x = -5$
$x = 2$
$x = 5$
$x = 10$
$x = 15$

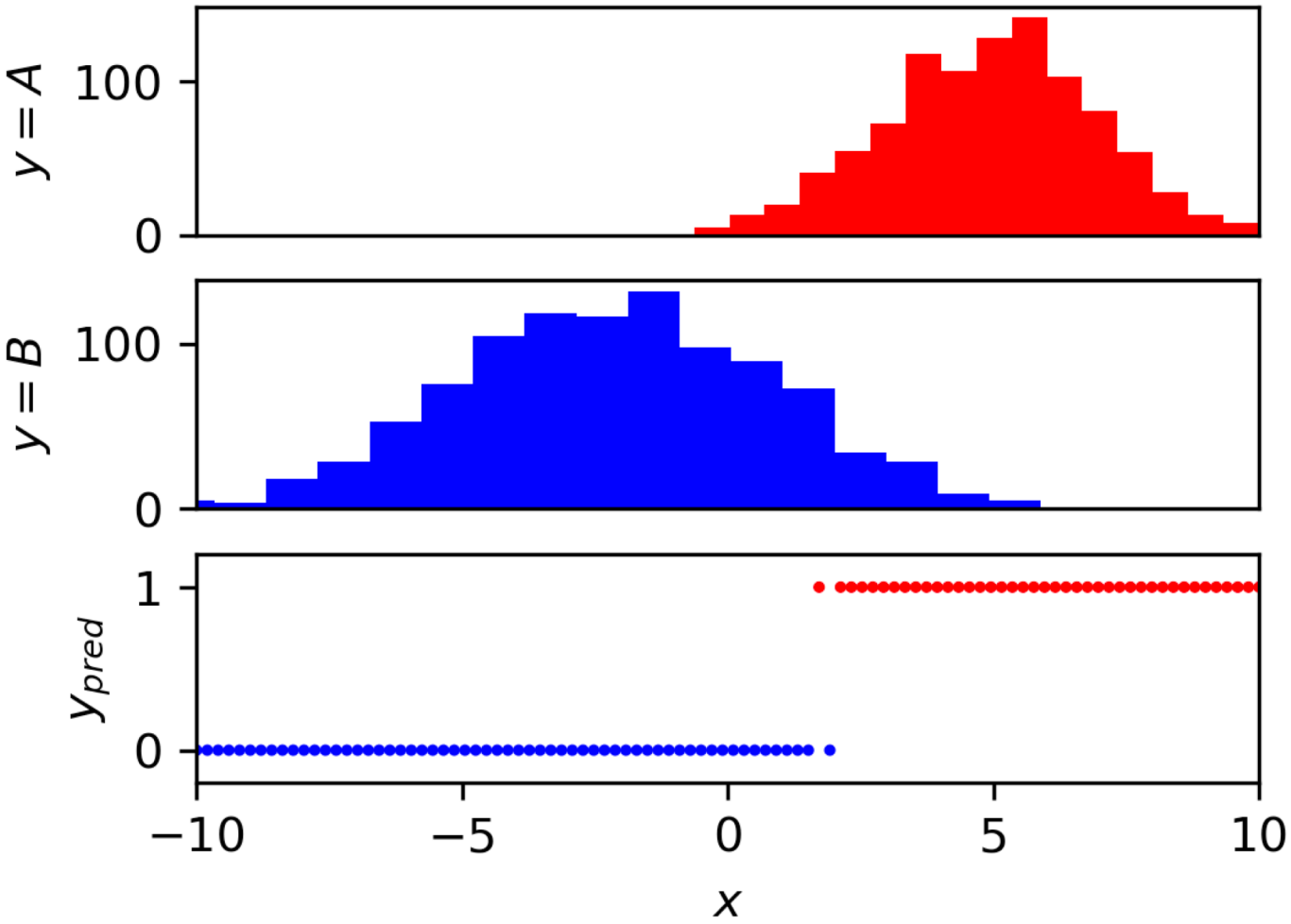


ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

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

Подход №1: **KNN** (метод $\boldsymbol{K}$ ближайших соседей)

1. выбрать $\boldsymbol{K}$ 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



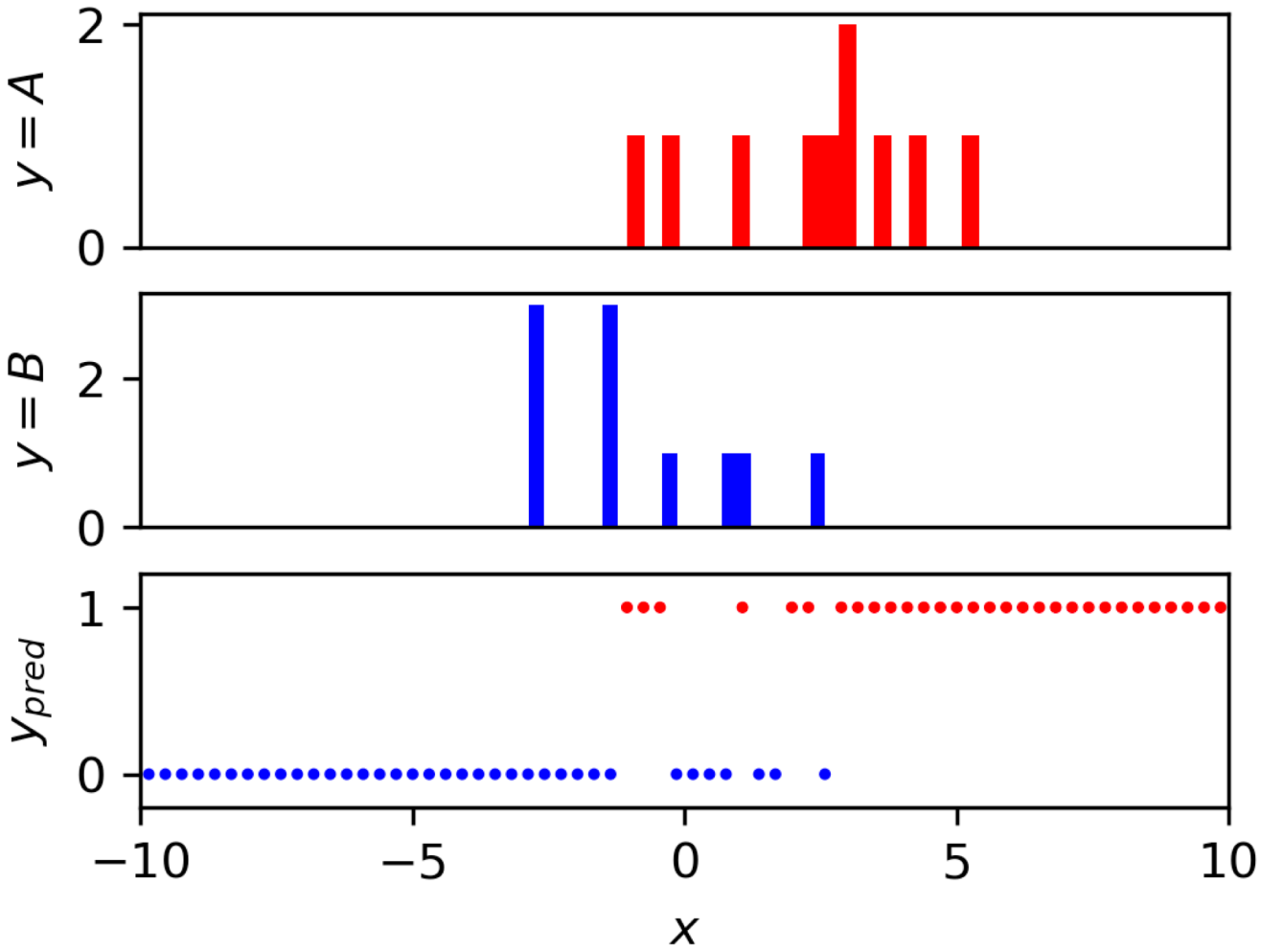
ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

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



ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

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

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

ОБУЧЕНИЕ С УЧИТЕЛЕМ: 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, помни, что $P(x_i)$ – константа, которую можно не учитывать при сравнении $P(Y = \boldsymbol{A}|X = x_i)$ и $P(Y = \boldsymbol{B}|X = x_i)$

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

формула полной вероятности

ОБУЧЕНИЕ С УЧИТЕЛЕМ: 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 $P(X|Y = \boldsymbol{A})$, $P(X|Y = \boldsymbol{B})$ etc., - то можно получить **аналитическое решение!**

И это решение будет ЛУЧШИМ из всех возможных.

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 $P(X|Y = \boldsymbol{A})$, $P(X|Y = \boldsymbol{B})$ etc., - то можно получить **аналитическое решение!**

previously on ML4ES

объекты описываются действительным признаком $\boldsymbol{x}$

целевая переменная $\boldsymbol{y}$ – бинарная

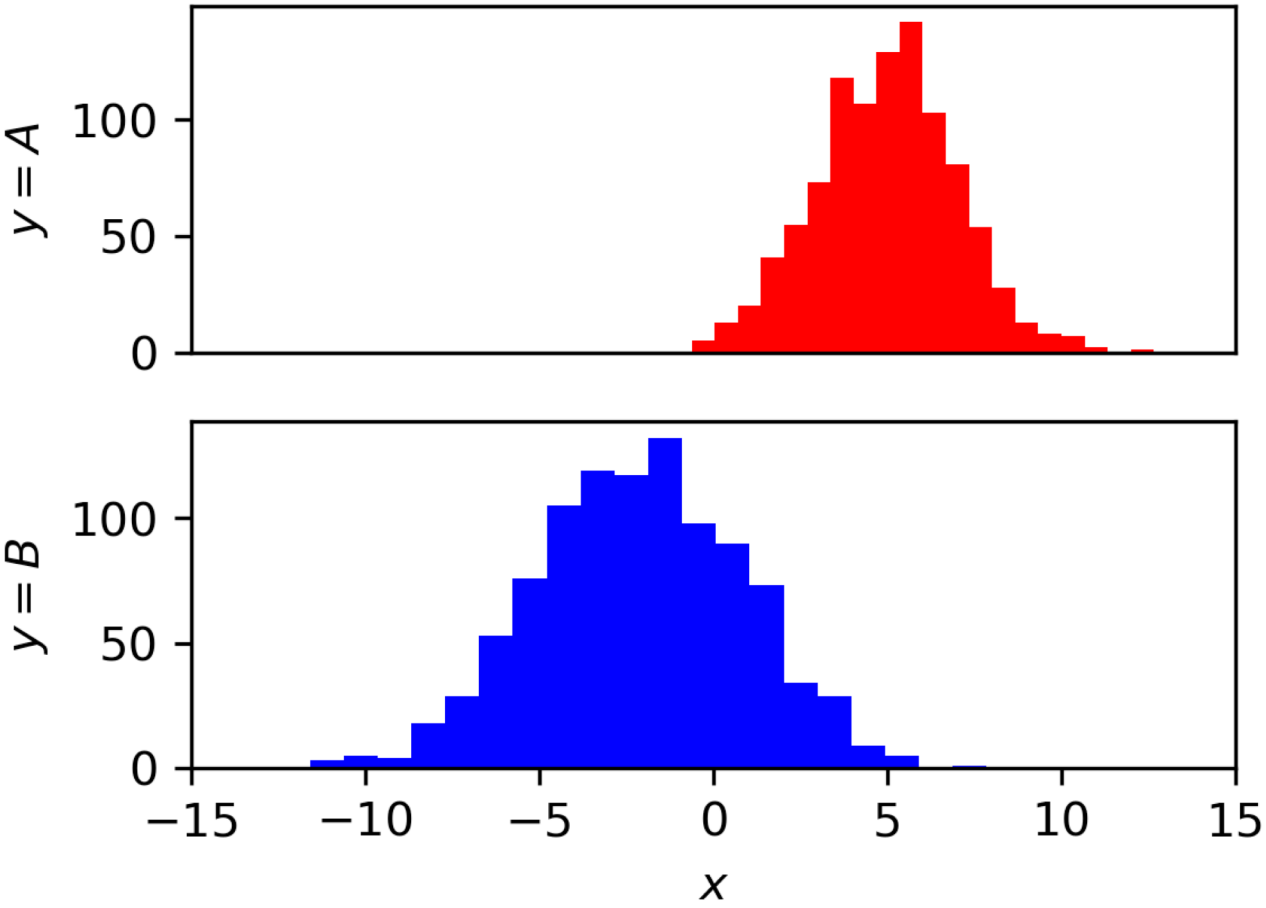
пусть для класса $y = \boldsymbol{A}$ значения $x \sim \mathcal{N}(\mu_A, \sigma_A)$, для класса $y = \boldsymbol{B}$ значения $x \sim \mathcal{N}(\mu_B, \sigma_B)$

$\mu_A = 5$

$\mu_B = -2$

$\sigma_A = 2$

$\sigma_B = 3$



ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

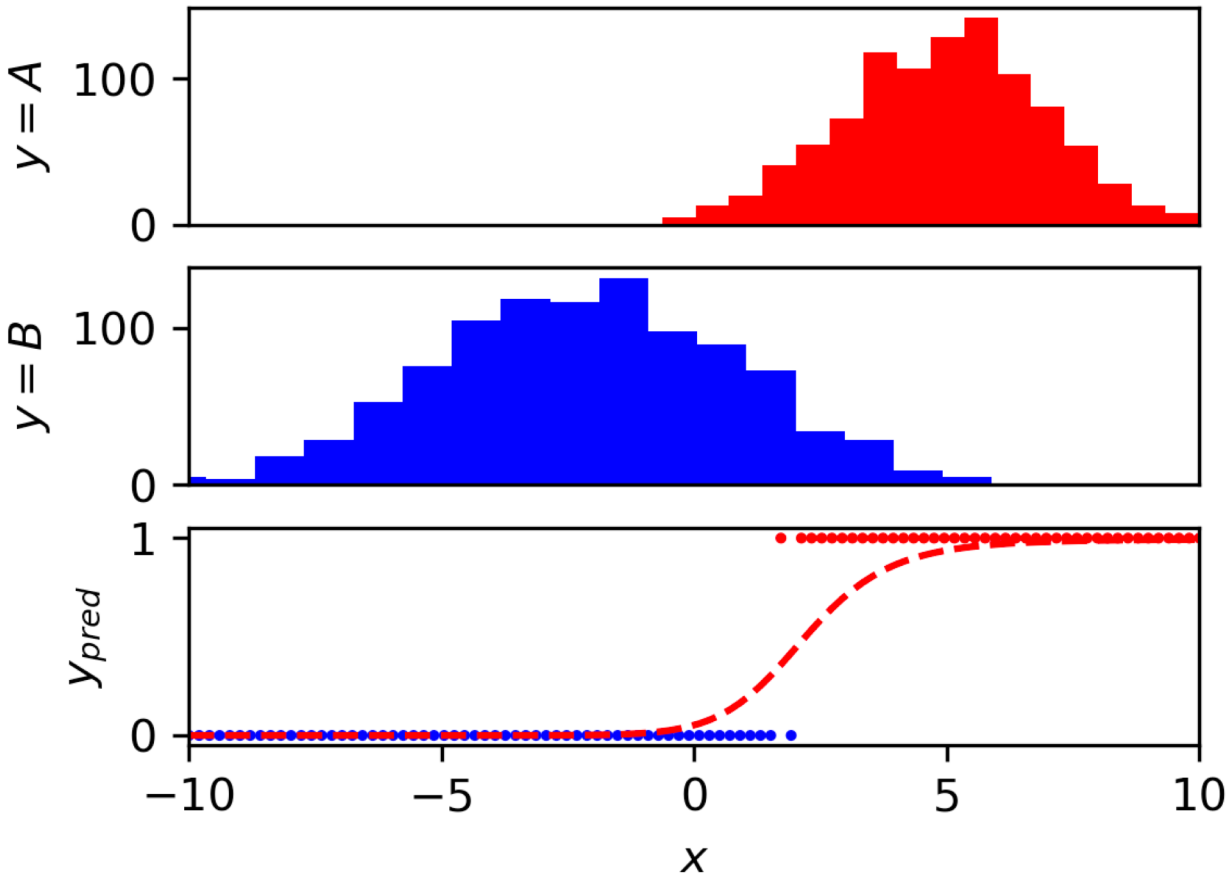
$$P(Y|X) = \frac{P(X|Y)\ \ P(Y)}{P(X)}$$

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

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

«Байесовский классификатор»

(не путать с «naïve bayes»)



ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

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

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

ЕСЛИ нам повезло и мы знаем распределения X для каждого из классов $P(X|Y = \boldsymbol{A})$, $P(X|Y = \boldsymbol{B})$ etc.,
то можно получить **аналитическое решение!**

И это решение будет ЛУЧШИМ из всех возможных.

А ЧТО ЕСЛИ НАМ НЕ ПОВЕЗЛО???

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

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

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

ЕСЛИ нам повезло и мы знаем распределения X для каждого из классов $P(X|Y = \boldsymbol{A})$, $P(X|Y = \boldsymbol{B})$ etc.,
то можно получить **аналитическое решение!**

И это решение будет ЛУЧШИМ из всех возможных.

А ЧТО ЕСЛИ НАМ НЕ ПОВЕЗЛО???

Надо как-то оценить распределения $P(X|Y = \boldsymbol{A})$, $P(X|Y = \boldsymbol{B})$ руководствуясь данными, которые у нас на руках

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Naïve bayes classifier, «наивный байесовский классификатор»

$$P(Y|X) \propto P(X|Y)\ P(Y)$$

можно забыть про $P(X = x)$

$$P(X|Y) = P\left(x_1, x_2, x_3 \dots x_p \middle| Y\right) = P\left(x_2, x_3 \dots x_p \middle| Y, x_1\right) * P(x_1|Y) = \cdots$$

$$= P(x_1|Y) * P(x_2|Y, x_1) * P(x_3|Y, x_1, x_2) * \cdots * P(x_p|Y, x_1, \dots, x_{p-1})$$

наивное предположение: переменные x_i условно независимы:

$$P(x_k|Y, x_1, x_2, \dots, x_{k-1}) = P(x_k|Y)$$

Остается оценить распределения $P(x_k|Y)$ для всех k независимо друг от друга – и можно подставлять в оценку вероятности $P(Y|X)$. Оценка распределения может базироваться на предположении об их нормальности, - тогда нужно оценить выборочное среднее (оценка параметра μ) и выборочную дисперсию (оценка параметра σ^2).

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Naïve bayes classifier, «наивный байесовский классификатор»