Разбор летучки

Лекция 7

Выбор моделей

Екатерина Тузова

Задача выбора метода обучения

X - множество объектов

Y - множество классов

Обучающая выборка: $X^l = (x_i, y_i)_{i=1}^l$

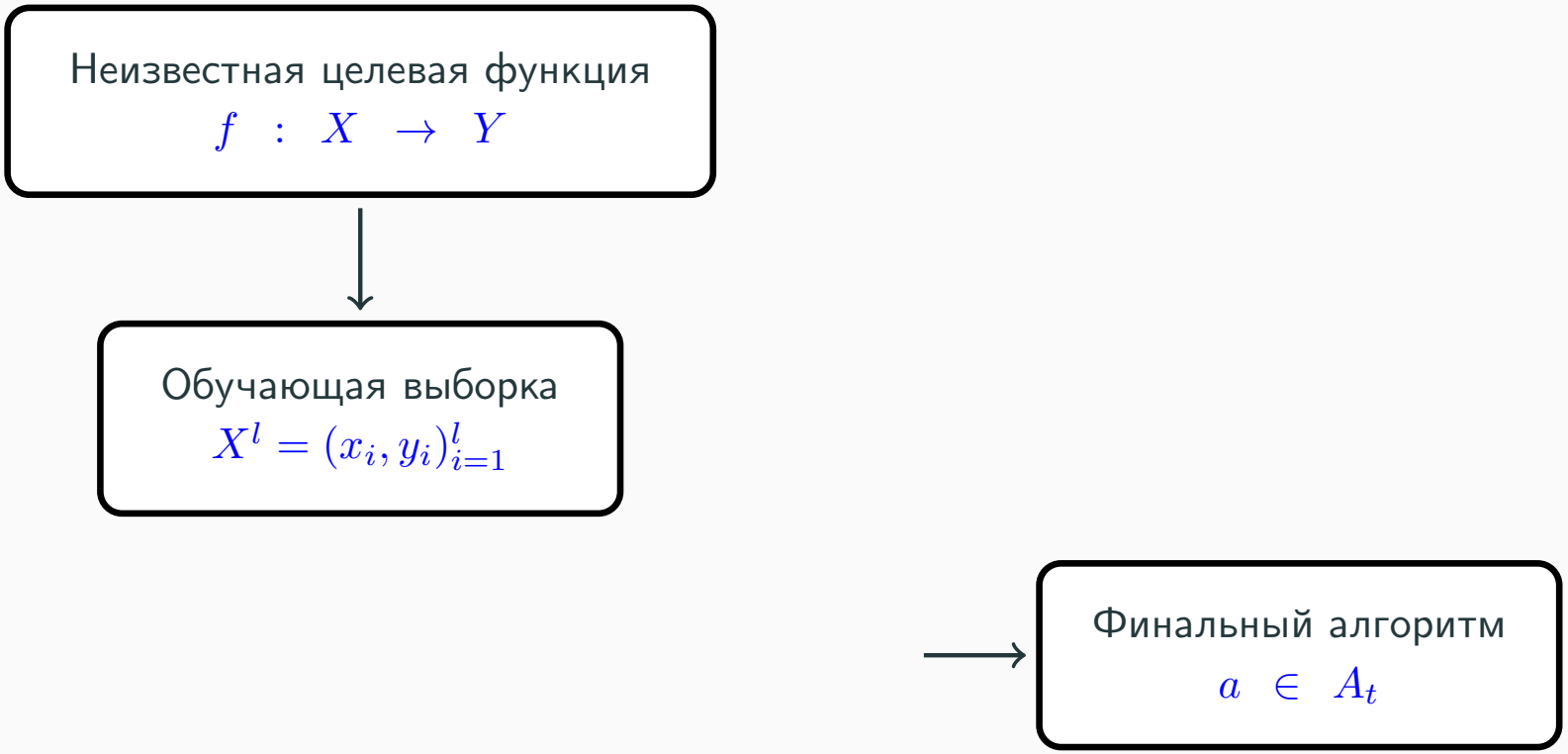
Целевая функция: $f : X \to Y$

Набор моделей алгоритмов $A_t : X \to Y$, $t \in T$

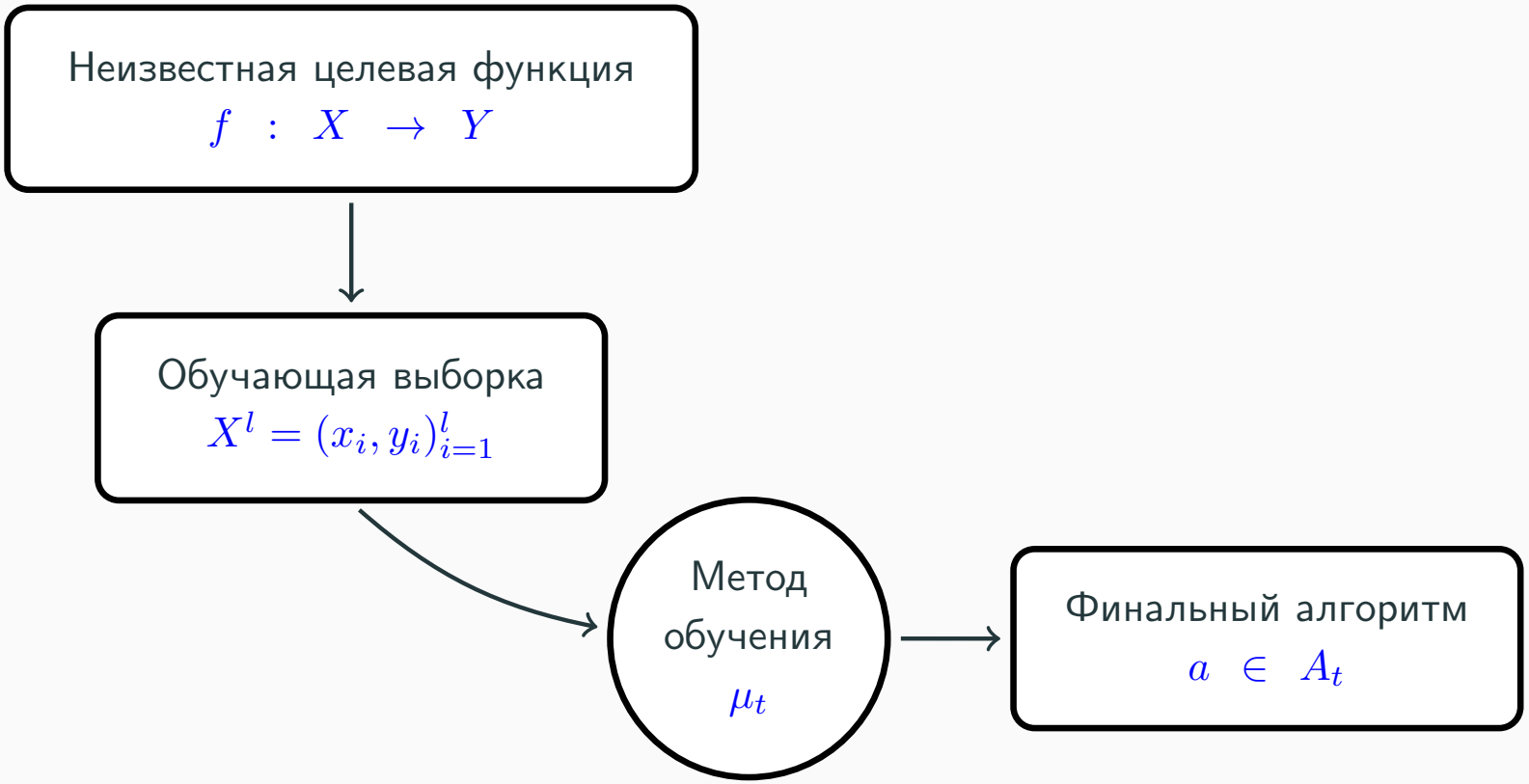
Методы обучения $\mu_t : X \times Y \to A_t$, $t \in T$

Задача: Найти алгоритм $a \in A_t$ с наилучшей обобщающей способностью

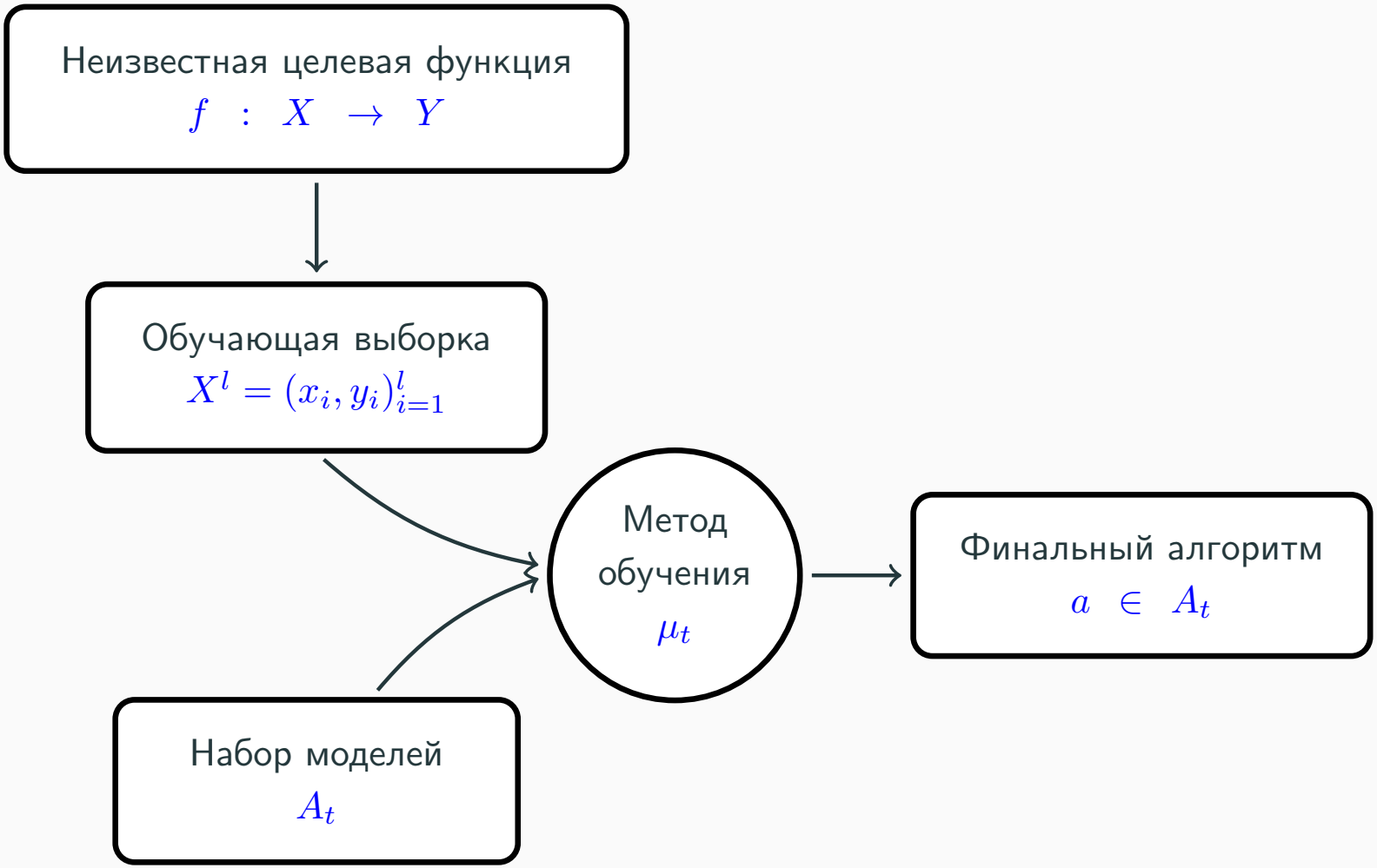
Модель


Неизвестная целевая функция
$f : X \to Y$
Обучающая выборка
$X^l = (x_i, y_i)_{i=1}^l$
Финальный алгоритм
$a \in A_t$

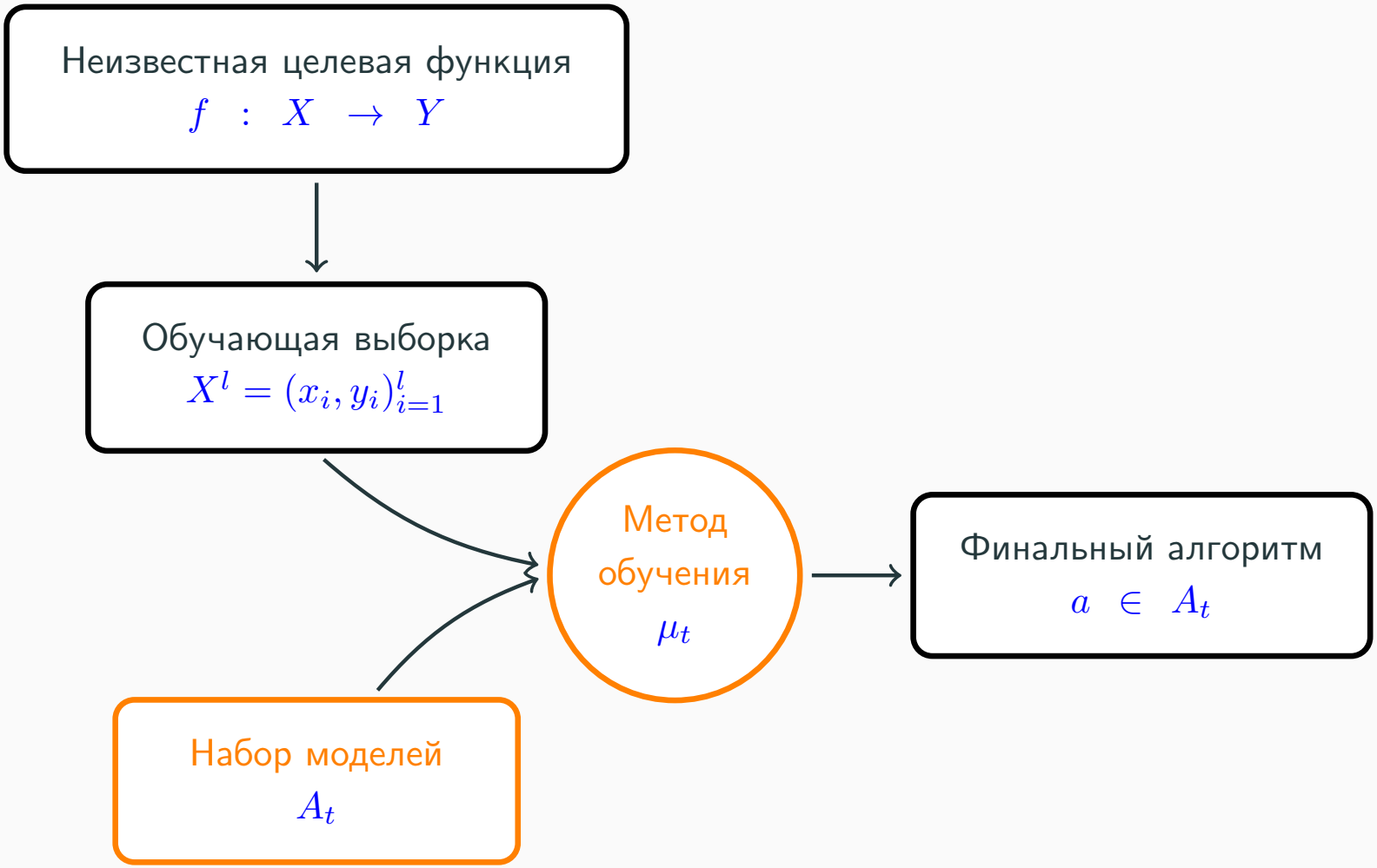
Модель


Неизвестная целевая функция
$f : X \to Y$
Обучающая выборка
$X^l = (x_i, y_i)_{i=1}^l$
Метод обучения
μ_t
Финальный алгоритм
$a \in A_t$

Модель


Неизвестная целевая функция
$f : X \to Y$
Обучающая выборка
$X^l = (x_i, y_i)_{i=1}^l$
Метод обучения
μ_t
Финальный алгоритм
$a \in A_t$
Набор моделей
A_t

Модель


Неизвестная целевая функция
$f : X \to Y$
Обучающая выборка
$X^l = (x_i, y_i)_{i=1}^l$
Метод обучения
μ_t
Финальный алгоритм
$a \in A_t$
Набор моделей
A_t

Пример

Перцептрон

Набор моделей:

$$a(\mathbf{x}) = \text{sign}(\sum_{j=1}^{n} w_j x^j - w_0)$$

Перцептрон

Набор моделей:

$$a(\mathbf{x}) = \text{sign}(\sum_{j=1}^{n} w_j x^j - w_0)$$

Перцептрон

Набор моделей:

$$a(\mathbf{x}) = \text{sign}(\sum_{j=1}^{n} w_j x^j - w_0)$$

Метод обучения:

```
1 function PERCEPTRON(X^l)
2     Инициализировать w_0, ..., w_n
3     repeat[пока w изменяются]
4         for i = 1, ..., l do
5             if a(x_i) ≠ y_i then
6                 w = w + y_i x_i
```

Задача

Научиться оценивать метод обучения и обобщающую способность алгоритма

Функция потерь

Функция потерь $\mathcal{L}(a, x_i)$ – характеризует величину ошибки алгоритма a на объекте x_i.

Если $\mathcal{L}(a, x_i) = 0$, то ответ $a(x_i)$ называется корректным.

Функция потерь

Задача классификации:
$\mathcal{L}(a, x_i) = [a(x_i) \neq y_i]$ – индикатор ошибки

Задача регрессии:
$\mathcal{L}(a, x_i) = |a(x_i) - y_i|$ – абсолютное значение ошибки
$\mathcal{L}(a, x_i) = (a(x_i) - y_i)^2$ – квадратичная ошибка

Функционал качества

Функционал качества алгоритма a на выборке X^l:

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(a, x_i)$$

Минимизация эмпирического риска:

$$\arg\min_{A_t} Q(a, X^l)$$

Внутренний функционал

$$Q_\mu(X^l) = Q(\mu(X^l), X^l)$$

Этот функционал оценивает качество обучения на выборке X^l

Внутренний функционал

$$Q_\mu(X^l) = Q(\mu(X^l), X^l)$$

Этот функционал оценивает качество обучения на выборке X^l

Какая с этим может быть проблема?

Внешний функционал

$$E_{in} = Q_\mu(X^l) = Q(\mu(X^l), X^l)$$

Внешний функционал по отложенной выборке:

$$E_{out} = Q_\mu(X^t, X^k) = Q(\mu(X^t), X^k)$$

Внешний функционал

$$E_{in} = Q_\mu(X^l) = Q(\mu(X^l), X^l)$$

Внешний функционал по отложенной выборке:

$$E_{out} = Q_\mu(X^t, X^k) = Q(\mu(X^t), X^k)$$

Какой здесь недостаток?

Внешний функционал

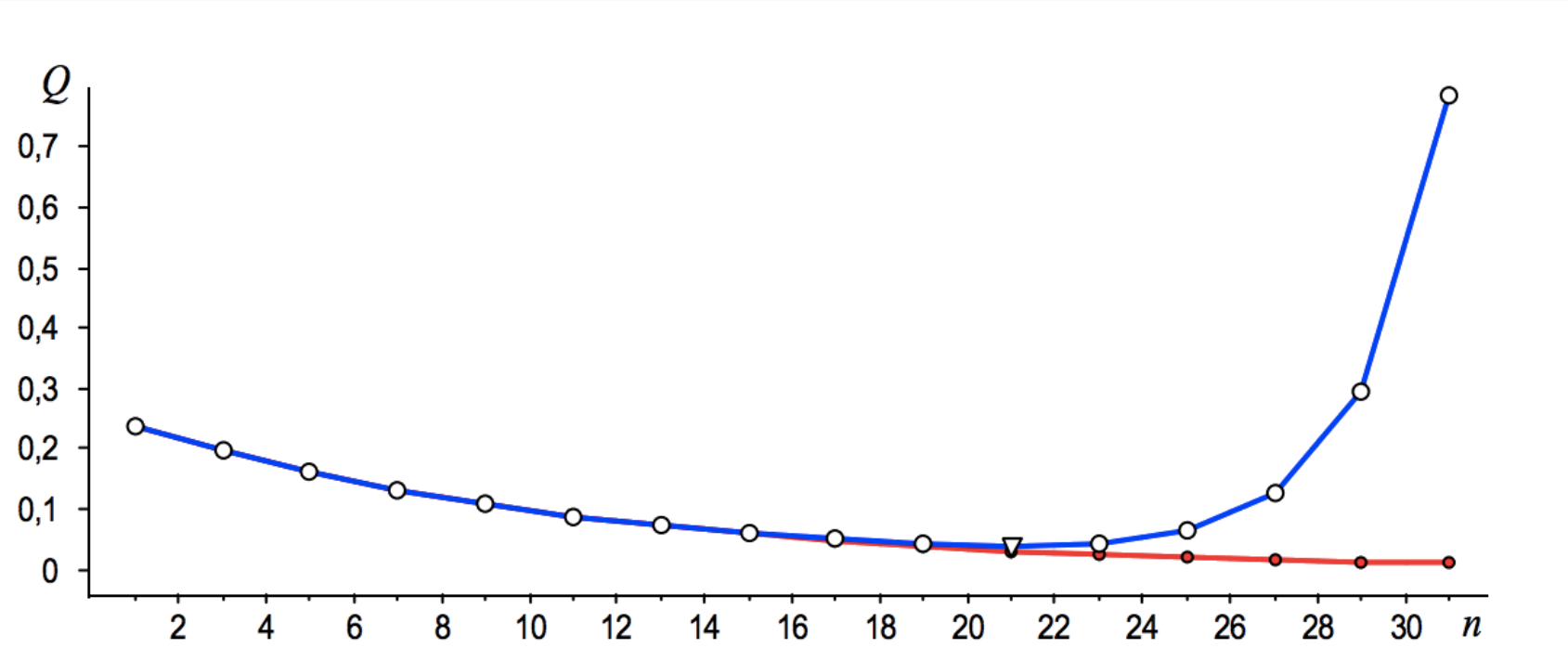
$$E_{in} = Q_\mu(X^l) = Q(\mu(X^l), X^l)$$

Внешний функционал по отложенной выборке:

$$E_{out} = Q_\mu(X^t, X^k) = Q(\mu(X^t), X^k)$$

Сильная зависимость от разбиения $X^l = X^t \sqcup X^k$

Кривая обучения



Кроссвалидация

Идея: Усреднить по всем C_l^t выборкам $X^l = X^t \sqcup X^k$

$$CCV(\mu, X^l) = \frac{1}{C_l^t} \sum_{X^t} Q_\mu(X^t, X^k)$$

Кроссвалидация

Идея: Усреднить по всем C_l^t выборкам $X^l = X^t \sqcup X^k$

$$CCV(\mu, X^l) = \frac{1}{C_l^t} \sum_{X^t} Q_\mu(X^t, X^k)$$

Во что превратится оценка при $k = 1$?

Недостатки

- Оценка вычислительно слишком сложна
- Не учитывает дисперсию X^k

k-fold кроссвалидация

Идея: Возьмём случайное разбиение $X^l = X_1 \sqcup \cdots \sqcup X_k$ на k блоков равной длины.

$$CV_k(\mu, X^l) = \frac{1}{k} \sum_{i=1}^{k} Q_\mu(X^l \setminus X_i, X_i)$$

k-fold кроссвалидация

Идея: Возьмём случайное разбиение $X^l = X_1 \sqcup \cdots \sqcup X_k$ на k блоков равной длины.

$$CV_k(\mu, X^l) = \frac{1}{k} \sum_{i=1}^{k} Q_\mu(X^l \setminus X_i, X_i)$$

Недостатки:

- Оценка зависит от разбиения на блоки
- Каждый объект только один раз участвует в контроле

Повторяющийся k-fold CV

Идея: Выборка разбивается t раз случайным образом на k блоков

$$CV_{tk}(\mu, X^l) = \frac{1}{t}\sum_{j=1}^{t}\frac{1}{k}\sum_{i=1}^{k} Q_\mu(X^l \setminus X_{ji}, X_{ji})$$

Повторяющийся k-fold CV

Идея: Выборка разбивается t раз случайным образом на k блоков

$$CV_{tk}(\mu, X^l) = \frac{1}{t}\sum_{j=1}^{t}\frac{1}{k}\sum_{i=1}^{k} Q_\mu(X^l \setminus X_{ji}, X_{ji})$$

+ Выбором t можно улучшать точность оценки
+ Каждый объект участвует в контроле t раз

Критерий непротиворечивости моделей

Идея: Если модель верна, то алгоритмы, настроенные по разным частям данных, не должны противоречить друг другу.

Аналитический подход

Как использовать аналитические оценки

1. Получить верхнюю оценку вероятности переобучения R_ε

$$R_\varepsilon(\mu, X^l) = P[E_{out} - E_{in} \geq \varepsilon] \leq \eta(\varepsilon, A)$$

2. Тогда для любых X^l, A, μ и η с вероятностью не менее $1 - \eta$ справедливо

$$E_{out} \leq E_{in} + \varepsilon(\eta, A)$$

3. Будем оптимизивать

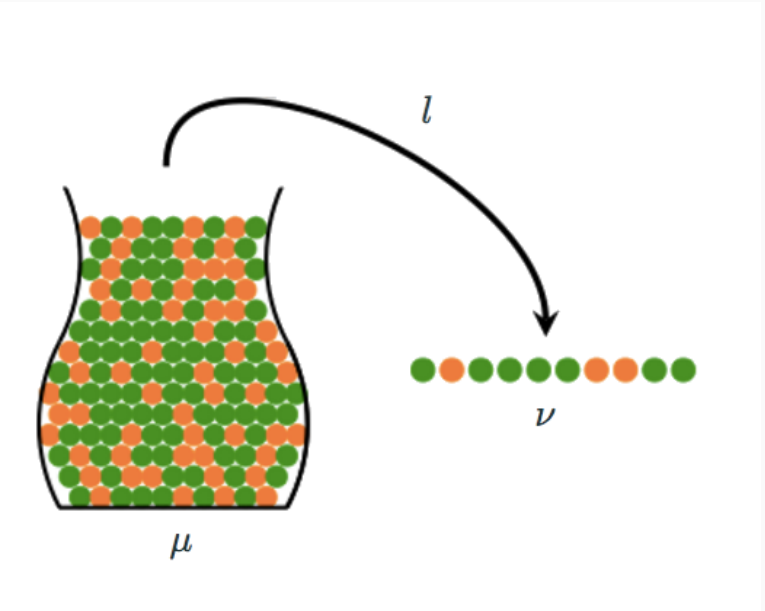
$$E_{in} + \varepsilon(\eta, A) \to \min_{\mu}$$

Регуляризация

Регуляризатор ε — аддитивная добавка к внутреннему критерию, обычно штраф за сложность модели A.

Неравенство Бернштейна-Хёфдинга

$$P[|\nu - \mu| < \varepsilon] \leq 2e^{-2\varepsilon^2 l}$$



ν – доля оранжевых шаров в выборке размера l

μ – истинная доля оранжевых шаров в данных

Какое отношение это имеет к нашим гипотезам?

Каждый шар это объект x из пространства X.
Неизвестная целевая функция f.

Оранжевый шар – модель h верна $(h(x) = f(x))$
Зелёный шар – модель h не верна $(h(x) \neq f(x))$

По выборке X^l можем оценить ν – долю объектов, на которых верна модель.

E_{in} и E_{out}

$E_{in}(h) = \nu$ - доля объектов в выборке X^l, на которых h верна
$E_{out}(h) = \mu$ - доля объектов во всём множестве X, на которых h верна

$$P[|E_{in}(h) - E_{out}(h)| < \varepsilon] \leq 2e^{-2\varepsilon^2 l}$$

Неравенство выполняется для каждой модели.

Важная деталь

Какова вероятность, что модель $a \in \mathcal{H}$, наилучшим образом приближающая f по выборке, наилучшим образом приближает f на всём множестве?

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз?

Простая аналогия

С какой вероятностью монета, подброшенная 10 раз, выпадет одной и той же стороной все 10 раз?

0.001

Простая аналогия

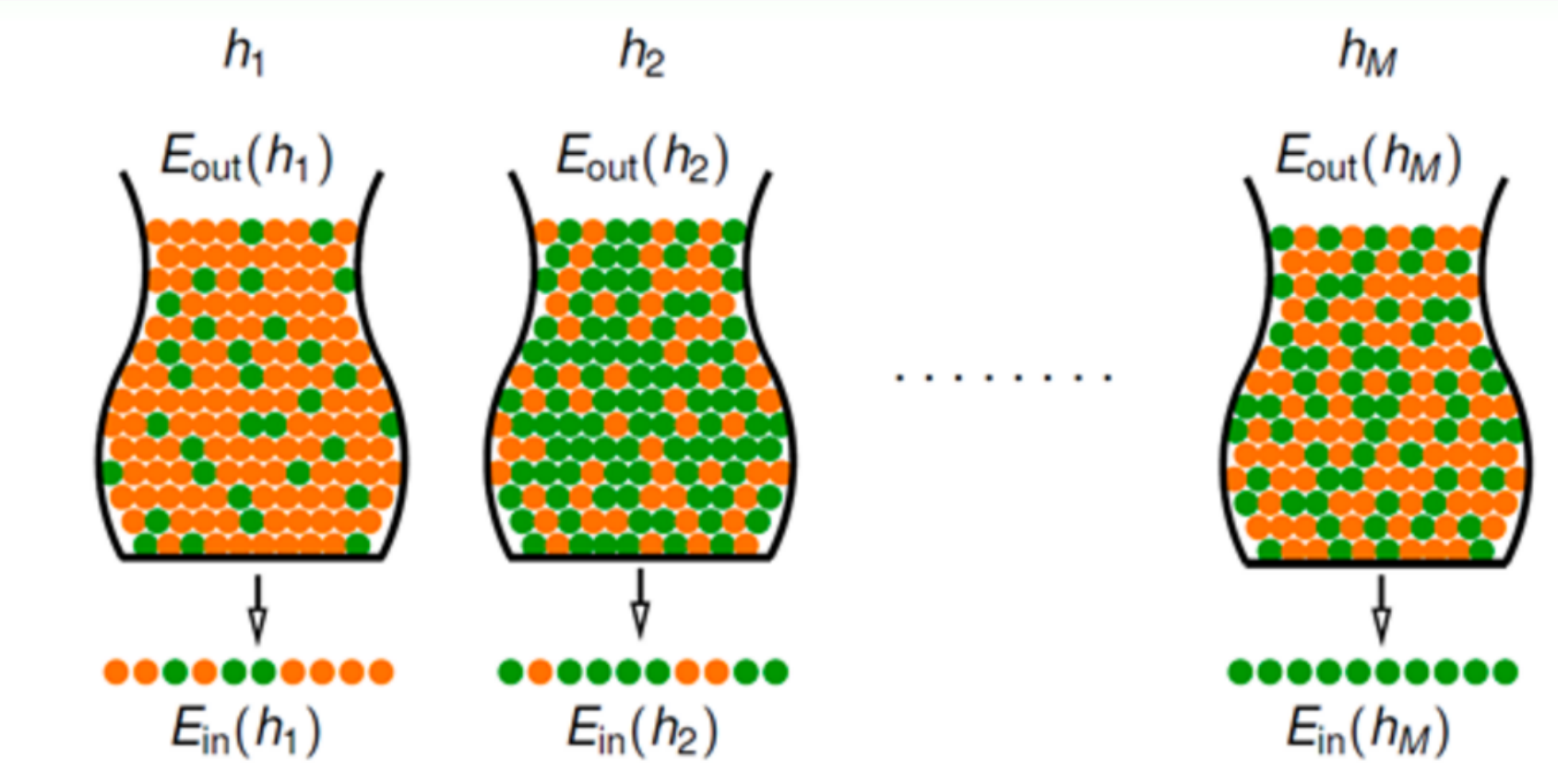
С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз?

Простая аналогия

С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз?

0.63

K нашей задаче



На h_M модели наблюдаем переобучение.

Решение

$$P[|E_{in}(a) - E_{out}(a)| < \varepsilon] \leq \sum_{m=1}^{M} P[|E_{in}(h) - E_{out}(h)| < \varepsilon]$$
$$\leq 2Me^{-2\varepsilon^2 l}$$

Полная и поточечная ошибка

$$E_{in}(h, X^l) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(h, x_i)$$

$$E_{out} = \mathbb{E}_x \mathcal{L}(h, x)$$

Обучение в терминах вероятности

В идеале хотим достичь $E_{out}(a) = 0$

Пытаемся достичь $E_{in}(a) \approx E_{out}(a)$ и $E_{in}(a) = 0$

Обучение в терминах вероятности

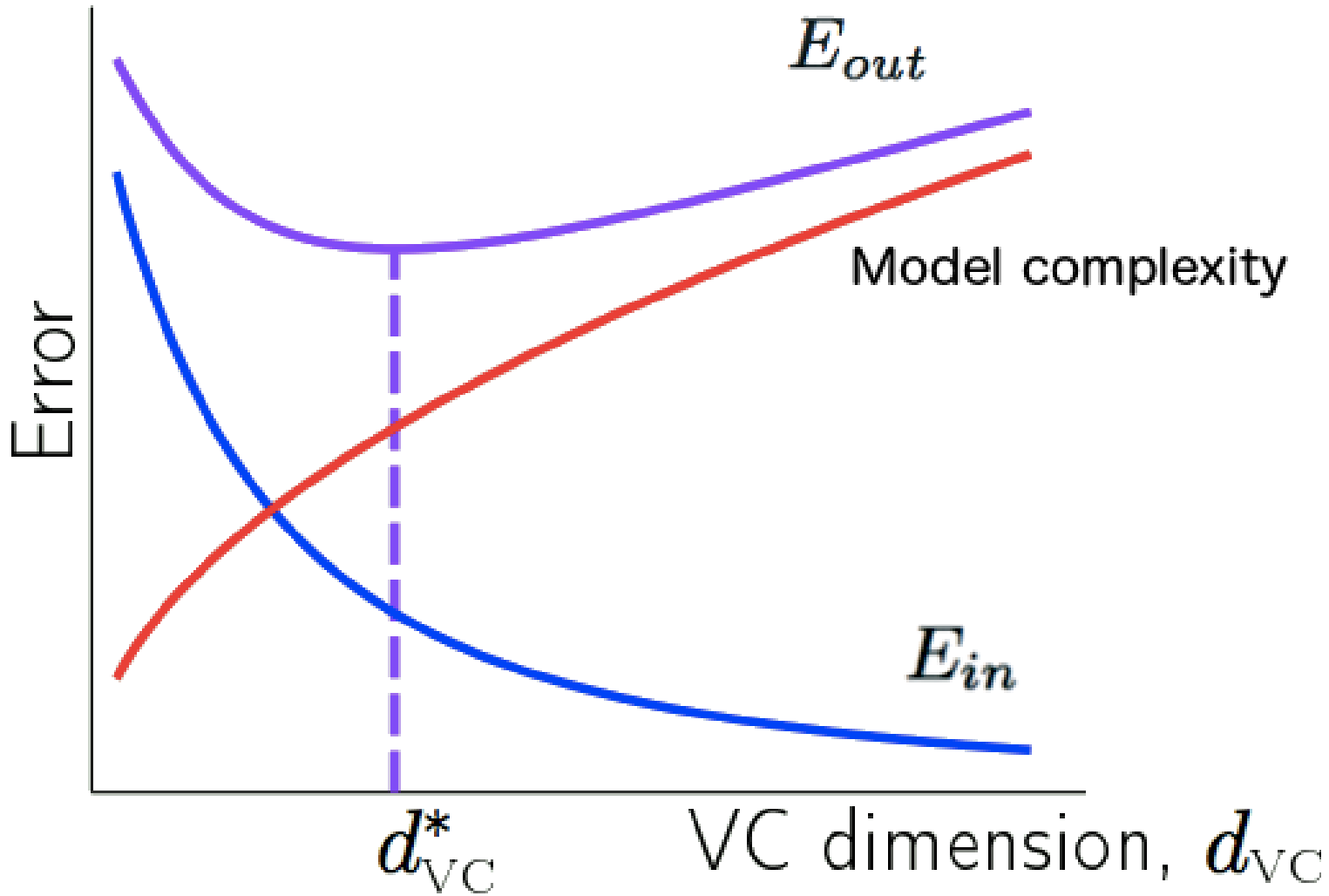
В идеале хотим достичь $E_{out}(a) = 0$

Пытаемся достичь $E_{in}(a) \approx E_{out}(a)$ и $E_{in}(a) = 0$

Два вопроса:

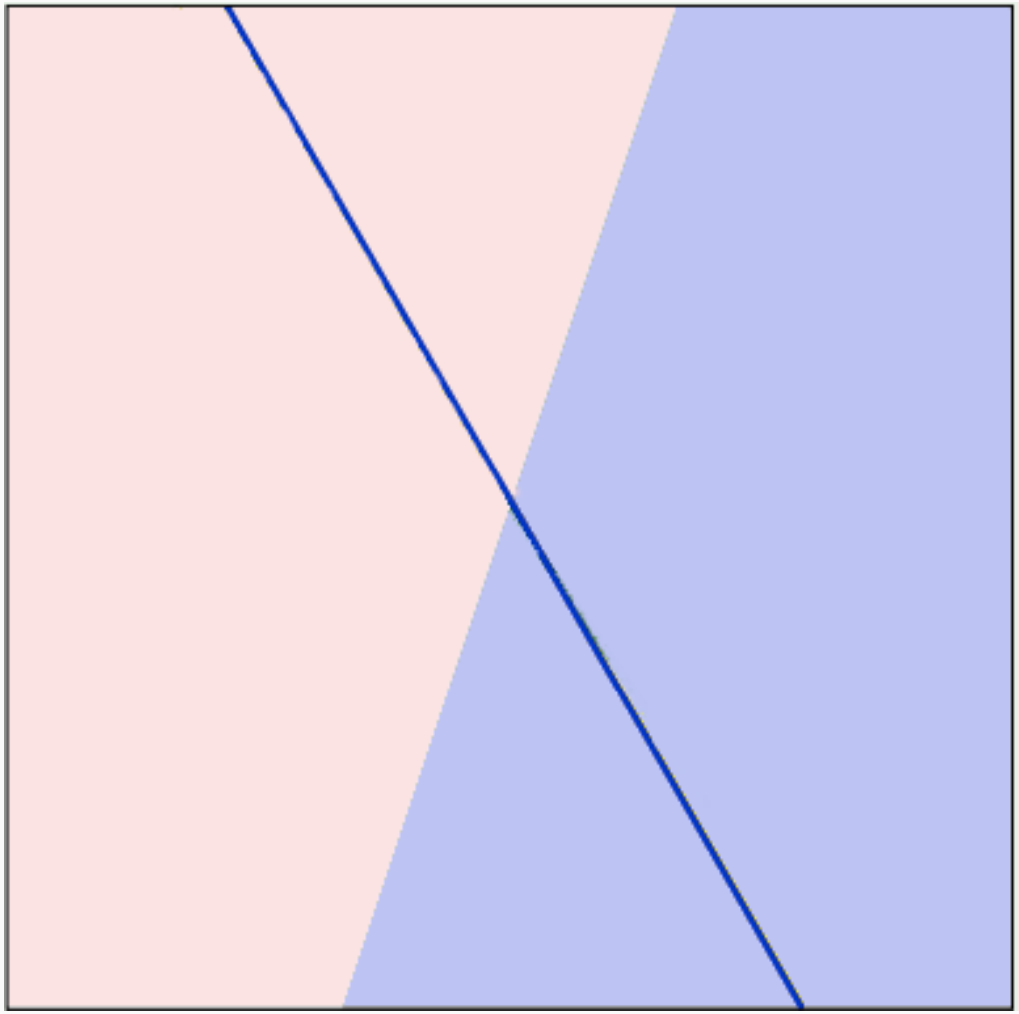
1. Можно ли судить о E_{out} по E_{in}?
2. Как уменьшить E_{in}?

Сложность модели

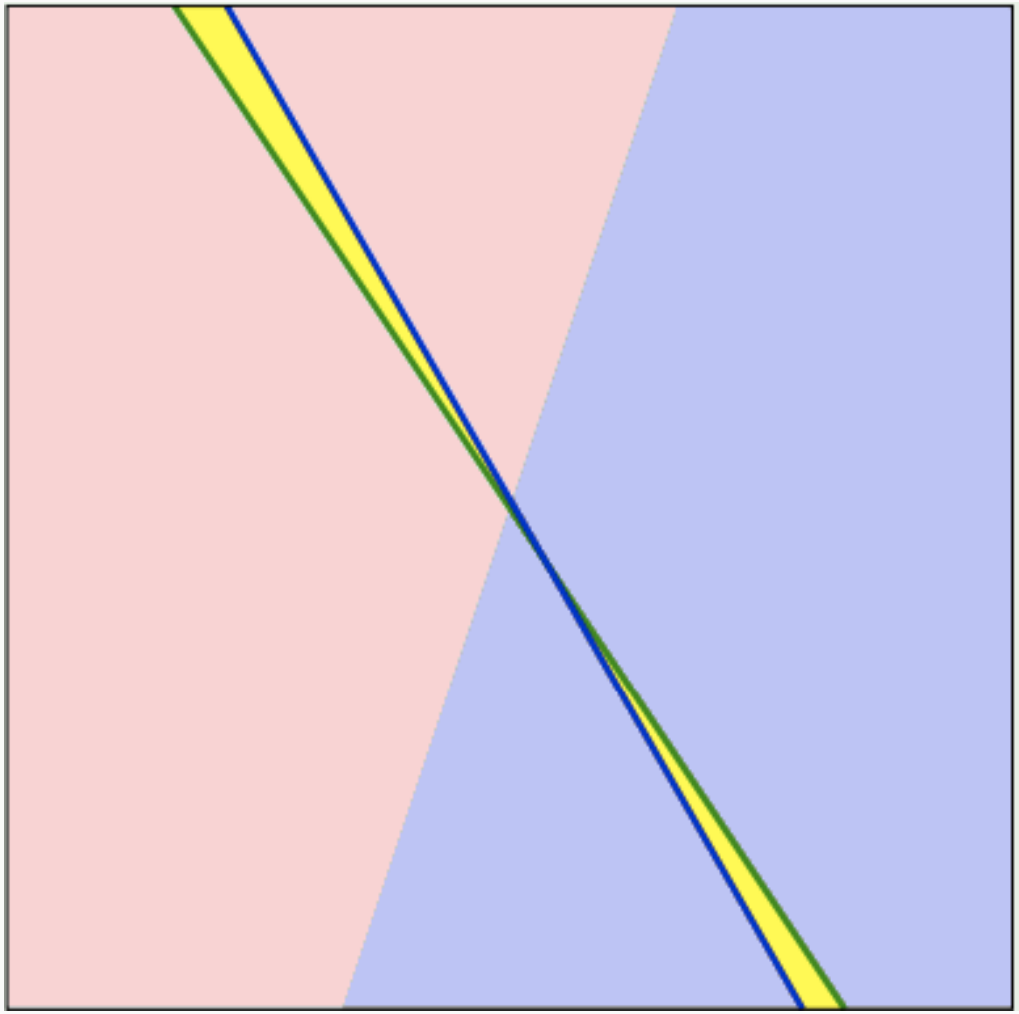


Какое количество гипотез в нашем пространстве $\mathcal{H}$?

Близость моделей



Близость моделей



Близость моделей

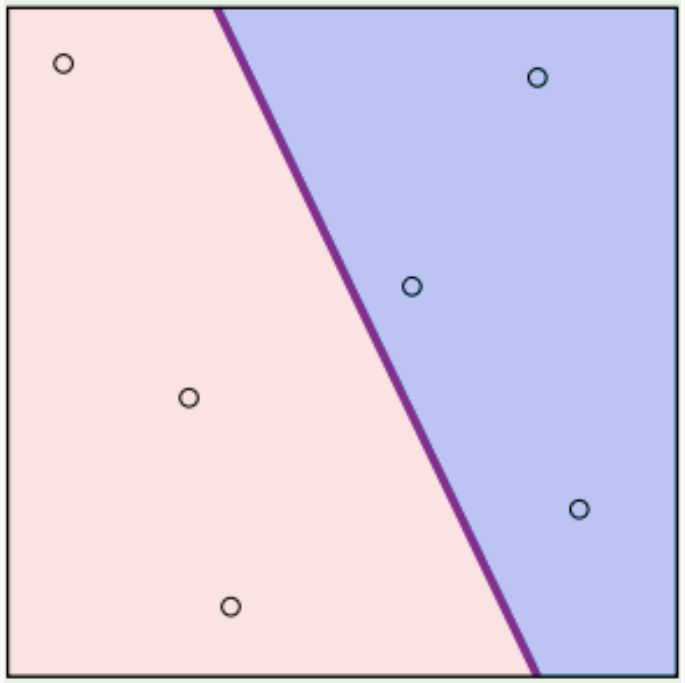
$\triangle E_{out} =$ площадь жёлтой области

$\triangle E_{in} =$ изменение меток объектов жёлтой области в выборке

$|E_{in}(h_1) - E_{out}(h1)| \approx |E_{in}(h_2) - E_{out}(h2)|$

Вопрос

Обучающая выборка $x_1, \ldots, x_l$ и набор бинарных значений меток $y_1, \ldots, y_l$.



Сколько вариантов $y_1, \ldots, y_l$?

Дихотомии

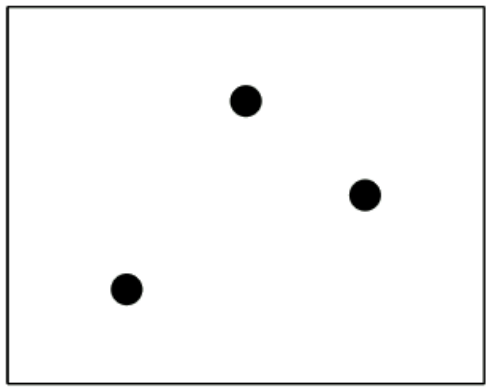
$$P(|E_{in}(a) - E_{out}(a)| > \varepsilon) \leq 2Me^{-2\varepsilon^2 l}$$

Наш набор моделей $\mathcal{H}$ может породить $|\mathcal{H}(x_1, \ldots, x_l)|$ дихотомий.

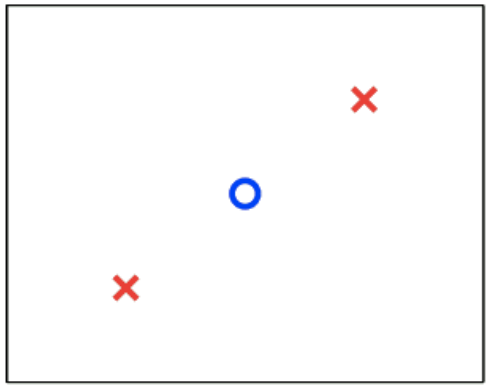
$$|\mathcal{H}(x_1, \ldots, x_l)| \leq 2^l$$

При этом сам набор моделей может быть бесконечным.

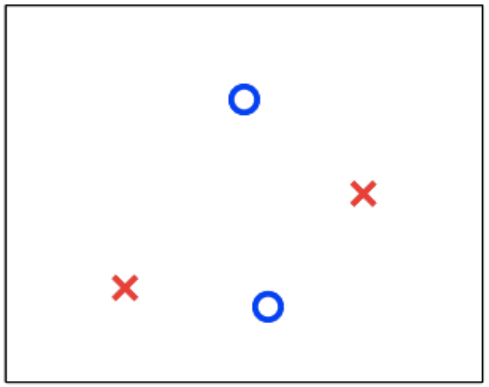
Функция роста $m_{\mathcal{H}(l)}$



$l = 3$



$l = 3$



$l = 4$

$$m_{\mathcal{H}(l)} = \max_{x_1,\ldots,x_l} |\mathcal{H}(x_1, \ldots, x_l)|$$

$$m_{\mathcal{H}(l)} \leq 2^l$$

Функция роста $m_{\mathcal{H}(l)}$

Гипотезы:

1. Луч на множестве точек: $m_{\mathcal{H}}(l) = l + 1$
2. Интервал на множестве точек: $m_{\mathcal{H}}(l) = C_{l+1}^2 + 1$
3. Выпуклое множество на множестве точек: $m_{\mathcal{H}}(l) = 2^l$

Точка разрыва

Если для некоторого k выполняется $m_{\mathcal{H}}(k) < 2^k$, то k называется точкой разрыва.

Точка разрыва

Наличие точки разрыва означает наличие полиномиального ограничения на функцию роста $m_{\mathcal{H}}(l)$

Пример

	# of rows	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{l-1}$	$\mathbf{x}_l$
S_1	α	$+1$	$+1$	...	$+1$	$+1$
		-1	$+1$	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	-1	-1
		-1	$+1$	...	-1	$+1$
S_2						
S_2^+	β	$+1$	-1	...	$+1$	$+1$
		-1	-1	...	$+1$	$+1$
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	$+1$
		-1	-1	...	-1	$+1$
S_2^-	β	$+1$	-1	...	$+1$	-1
		-1	-1	...	$+1$	-1
		⋮	⋮	⋮	⋮	⋮
		$+1$	-1	...	$+1$	-1
		-1	-1	...	-1	-1

Разделим на 2 ситуации относительно x_l:

либо есть только один вариант (+1 или -1), либо оба (+1 и -1)

Число дихотомий

$B(l, k)$ – максимальное число дихотомий для выборки размера l при наличии точки разрыва k.

- $B(l, k) = \alpha + 2\beta$
- $\alpha + \beta \leq B(l-1, k)$, т.к. $\alpha + \beta$ – число дихотомий для $l - 1$
- $\beta \leq B(l-1, k-1)$

$$\Longrightarrow B(l, k) \leq B(l-1, k) + B(l-1, k-1)$$

Число дихотомий

$$B(l,k) \leq \sum_{i=0}^{k-1} C_l^i$$

Доказывается по индукции.

Индукционный шаг:

$$\sum_{i=0}^{k-1} C_l^i = \sum_{i=0}^{k-1} C_{l-1}^i + \sum_{i=0}^{k-2} C_l^i$$

Полиномиальное ограничение

$$m_{\mathcal{H}(l)} \leq \sum_{i=0}^{k-1} C_l^i$$

Доказывается по индукции.

Индукционный шаг:

$$\sum_{i=0}^{k-1} C_l^i = \sum_{i=0}^{k-1} C_{l-1}^i + \sum_{i=0}^{k-2} C_l^i$$

Неравенство Вапника-Червоненкиса

$$P(E_{in}(a) - E_{out}(a) > \varepsilon) \leq 4m_{\mathcal{H}}(2l)e^{-\frac{1}{8}\varepsilon^2 l}$$

Размерность d_{VC} = наибольшее l, для которого $m_{\mathcal{H}}(l) = 2^l$.
$d_{VC} = k - 1$.

Вопросы?

Что почитать по этой лекции

- Professor Yaser Abu-Mostafa MOOC
- Tom Mitchell "Machine Learning" Chapter 4

На следующей лекции

- Задача максимизации зазора - аналог классификатора с регуляризацией
- Двойственная задача
- Что такое опорный вектор
- Регуляризация
- Решение для неразделимых выборок
- Kernel trick