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 $\mu : (X \times Y)^l \to A_t$, $t \in T$

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

Модель

Неизвестная целевая функция

$f : X \to Y$

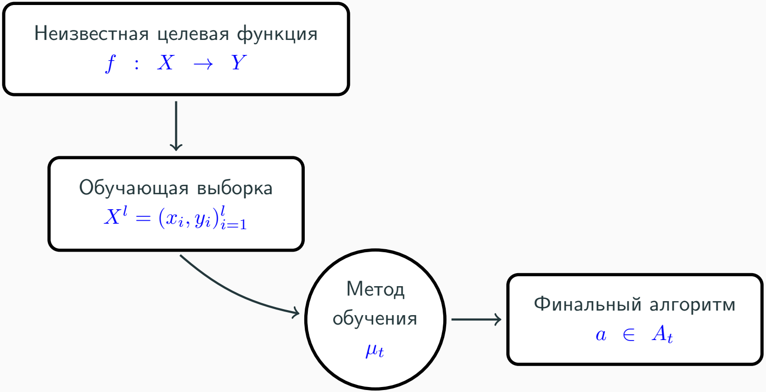
Обучающая выборка

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

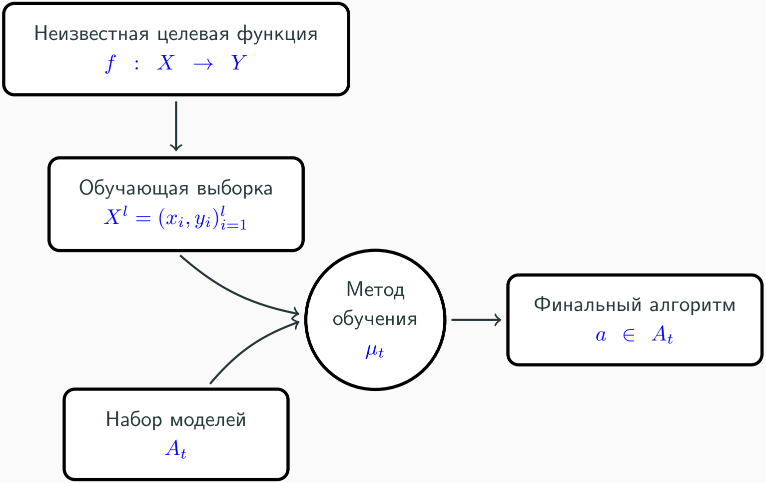
Финальный алгоритм

$a \in A_t$

Модель

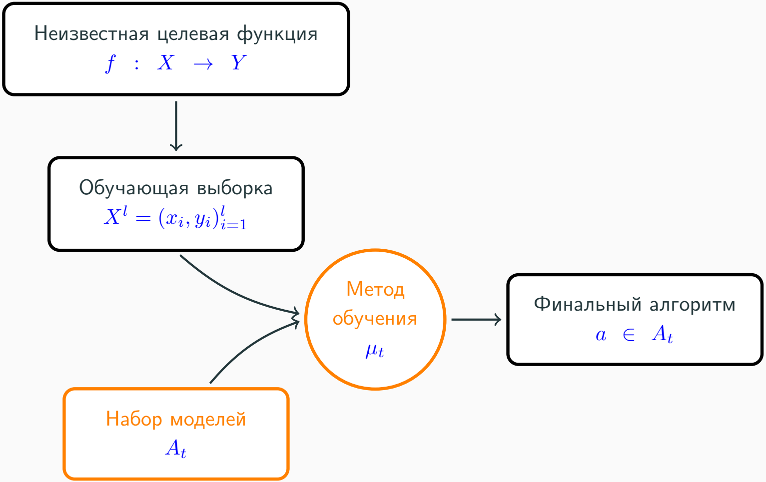

Неизвестная целевая функция
$f : X \to Y$
Обучающая выборка
$X^l = (x_i, y_i)_{i=1}^l$
Метод обучения
μ_t
Финальный алгоритм
$a \in A_t$

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:

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

Перцептрон

Набор моделей:

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

Перцептрон

Набор моделей:

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

Метод обучения:

function $\textsc{Perceptron}(X^l)$
    Инициализировать $w_0, \ldots, w_n$
    repeat[пока $\mathbf{w}$ изменяются]
        for $i = 1, \ldots, l$ **do**
            if $a(x_i) \neq y_i$ **then**
                $\mathbf{w} = \mathbf{w} + y_i \mathbf{x_i}$

Задача

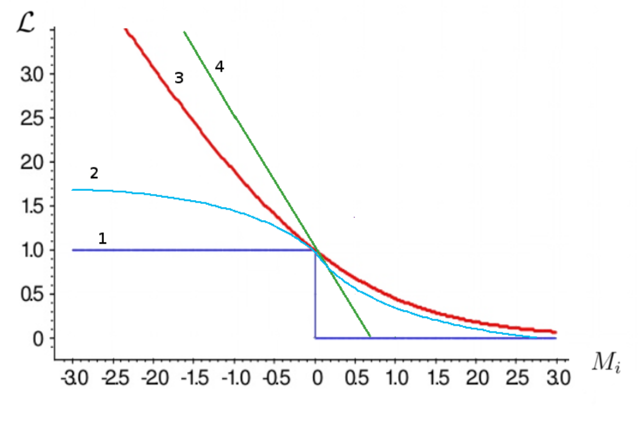
Научиться оценивать метод обучения и обобщающую способность алгоритма

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.

Примеры $\mathcal{L}$



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:

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

функционал средних потерь, Empirical Risk

Внутренний функционал

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

Этот функционал оценивает качество обучения на выборке X^l

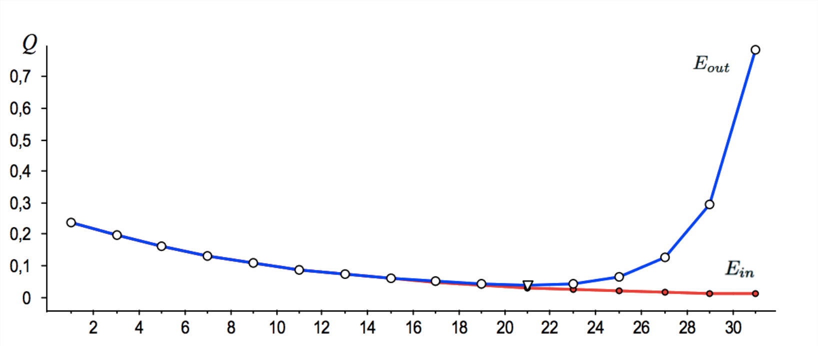
Внутренний функционал

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

Этот функционал оценивает качество обучения на выборке X^l

Какая с этим может быть проблема?

Сложность модели



Почему случается переобучение?

Проблема переобучения

- Слишком мало объектов
- Слишком много признаков
- Линейная зависимость признаков

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 $X^l = X^t \sqcup X^k$

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 $X^l = X_1 \sqcup \dots \sqcup X_k$ 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 $X^l = X_1 \sqcup \dots \sqcup X_k$ 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

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

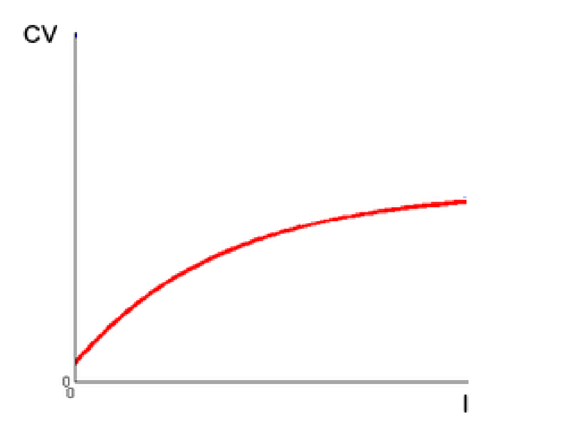
Повторяющийся k-fold CV

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

Кривая обучения



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

$$P[E_{out} - E_{in} > \varepsilon] \leq \eta(\varepsilon)$$

2. Тогда с вероятностью не менее $1 - \eta$ справедливо

$$E_{out} < E_{in} + \varepsilon(\eta)$$

3. Будем оптимизивать

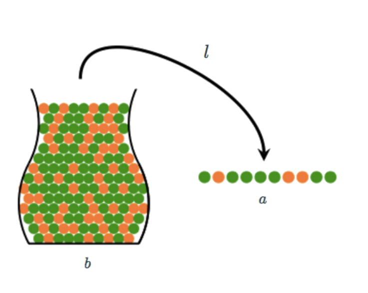
$$E_{in} + \varepsilon(\eta) \to \min_{\mu}$$

Регуляризация

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

Неравенство Бернштейна-Хёфдинга

$$P[|a - b| > \varepsilon] \leq 2e^{-2\varepsilon^2 l}$$



a – доля оранжевых шаров в выборке размера l

b – истинная доля оранжевых шаров

Какое отношение это имеет к нашим моделям?

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

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

По выборке X^l можем оценить долю объектов, на которых модель ошибается.

E_{in} и E_{out}

$E_{in}(h) = a$ - доля объектов в выборке X^l, на которых h ошибается
$E_{out}(h) = b$ - доля объектов во всём множестве X, на которых h ошибается

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

Неравенство выполняется для каждой модели.

Важная деталь

Какова вероятность, что модель $a \in A$, 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

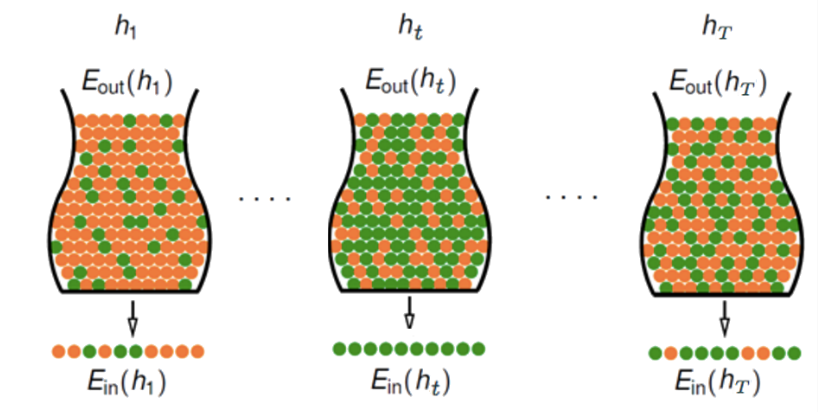
С какой вероятностью одна из 1000 монет, каждая из которых подброшена 10 раз, выпадет одной и той же стороной все 10 раз?

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

К нашей задаче



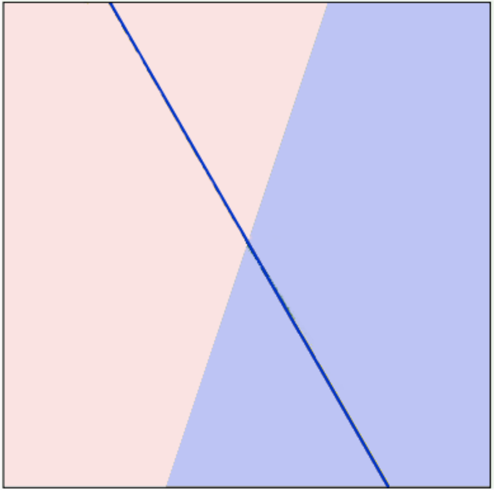
На h_t модели наблюдаем переобучение.

Решение

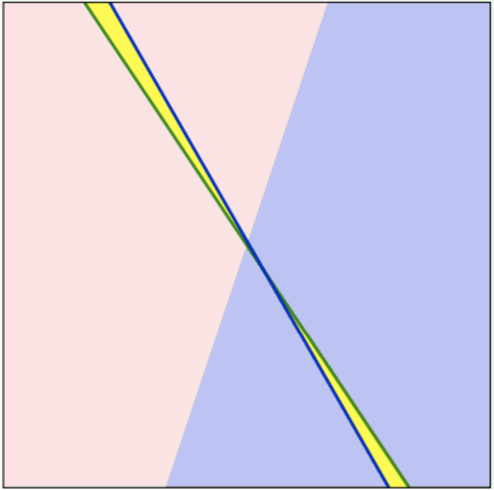
$$
\begin{aligned}
P[|E_{in}(a) - E_{out}(a)| > \varepsilon] &\leq \sum_{t=1}^{T} P[|E_{in}(h) - E_{out}(h)| > \varepsilon] \\
&\leq 2Te^{-2\varepsilon^2 l}
\end{aligned}
$$

Какое количество моделей в нашем пространстве A?

Близость моделей



Близость моделей



Близость моделей

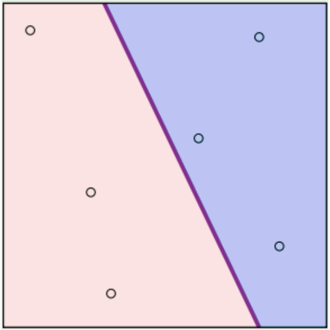
$\triangle E_{out} =$ площадь жёлтой области

$\triangle E_{in} =$ 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

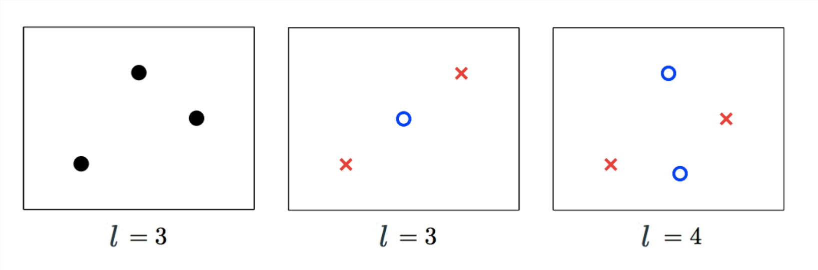
$$P(|E_{in}(a) - E_{out}(a)| > \varepsilon) \leq 2Te^{-2\varepsilon^2 l}$$

Наш набор моделей A может породить $|A(x_1, \ldots, x_l)|$ дихотомий.

$$|A(x_1, \ldots, x_l)| \leq 2^l$$

При этом сам набор моделей может быть бесконечным.

Функция роста $m_A(l)$



$$m_A(l) = \max_{x_1,\ldots,x_l} |A(x_1,\ldots,x_l)| \leq 2^l$$

Точка разрыва

Если для некоторого k выполняется $m_A(k) < 2^k$, то k называется точкой разрыва.

точка поломки, break point

Точка разрыва

Наличие точки разрыва означает наличие полиномиального ограничения на функцию роста $m_A(l)$

Пример

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

- $B(l,k) = \alpha + 2\beta$
- $\alpha + \beta \le B(l-1,k)$
- $\beta \le B(l-1,k-1)$

$$\implies B(l,k) \le B(l-1,k) + B(l-1,k-1)$$

Оценка $\alpha + \beta$

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

Оценка β

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

Число дихотомий

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

Доказывается по индукции:

$B(l,1) = 1, \;\; B(1,k) = 2$

Индукционный шаг: $\sum\limits_{i=0}^{k-1} C_l^i = \sum\limits_{i=0}^{k-1} C_{l-1}^i + \sum\limits_{i=0}^{k-2} C_l^i$

Полиномиальное ограничение

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

Финальное неравенство

$$P(E_{in}(a) - E_{out}(a) > \varepsilon) \leq 2l^{k-1}e^{-2\varepsilon^2 l}$$

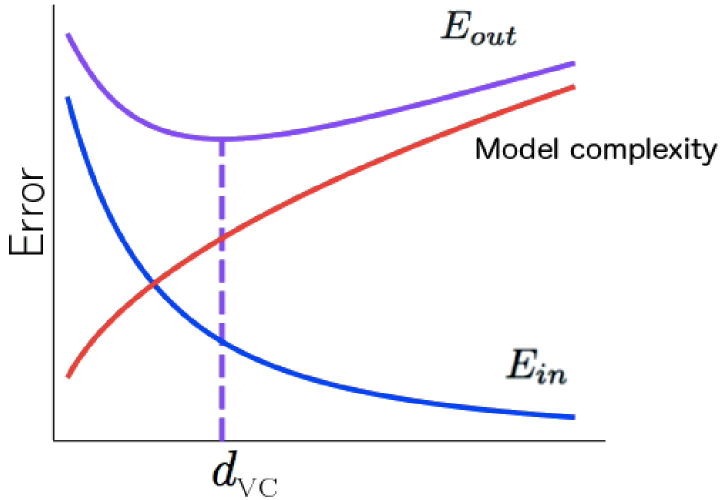
Неравенство Вапника-Червоненкиса

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

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

Для линейных классификаторов $d_{VC} = \#$ number of features + 1.

Сложность модели



Пример регуляризации в градиентном спуске

Функционал с регуляризацией:

$Q_\tau = Q + \frac{\tau}{2}\|\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}}$

Градиент:

$\triangledown Q_\tau = \triangledown Q + \tau\mathbf{w}$

Градиентный шаг:

$\mathbf{w} = \mathbf{w}(1 - \alpha\tau) - \alpha \triangledown Q(\mathbf{w})$

Вопросы?

Что почитать по этой лекции

- Professor Yaser Abu-Mostafa MOOC
- Tom Mitchell "Machine Learning" Chapter 4
- Hastie, T., Tibshirani R. "The Elements of Statistical Learning" Chapter 7.9

На следующей лекции

- Многослойная нейронная сеть
- Нелинейное преобразование
- Быстрое вычисление градиента
- Алгоритм обратного распространения ошибки
- Оптимизация структуры сети
- Сверточные нейросети