МАШИННОЕ ОБУЧЕНИЕ И ОБУЧАЕМОСТЬ: СРАВНИТЕЛЬНЫЙ ОБЗОР

В. И. Донской *
vidonskoy@mail.ru

Аннотация

В статье проведен сравнительный анализ различных определений обучаемости, рассмотрены необходимые и достаточные условия обучаемости, указаны границы применимости VC теории. Дополнительно выявленные фундаментальные положения дают объяснение практически наблюдаемой обучаемости при использовании некоторых алгоритмов и моделей обучения, несмотря на кажущееся противоречие с VC теорией: в действительности этого противоречия нет.

MACHINE LEARNING AND LEARNABILITY: COMPARATIVE SURVEY

V. I. Donskoy *
vidonskoy@mail.ru

Annotation

The comparative analysis of different definitions of learnability is presented in the article, the necessary and sufficient conditions of learnability are considered, the scopes of applicability of the VC theory are indicated. The additionally exposed fundamental results give explanation to the practically looked learnability when some algorithms and machine learning models are used, in spite of seeming contradiction with the VC theory: this contradiction is not present in actual fact.

Keywords: *PAC Learning, VC Dimension, Learnability, Stability, Necessary and Sufficient Conditions*

---

* *Vladimir I. Donskoy,*
Prof., Dr. Physical-Mathematical Sciences: Computer Science Theory
Tavrida National University named after V. I. Vernadsky
Simferopol, Ukraine

Донской В. И. Машинное обучение и обучаемость: сравнительный обзор // Intellectual Archive. – 2012. – №933 – 19 с.
http://intellectualarchive.com/getfile.php?file=6kEaOEViwnV&orig_file=Donskoy%20Mach%20Learning%20Survey.pdf

1. Введение. Основные понятия машинного обучения

Неформально машинное обучение можно представить как процесс нахождения неизвестного решающего правила (или неизвестной целевой функции) по некоторой начальной информации, которая обычно не является полной. Говорят, что значения аргументов искомой функции в совокупности являются описанием объекта в некоторой проблемной области или даже проще – допустимым объектом. Если целевая функция принимает только два значения, то её называют классифицирующей или классификатором. Уточняя процесс обучения нужно определить следующее:

- информацию о множестве (допустимых) объектов;
- о каком неизвестном решающем правиле или функции идёт речь;
- что предоставляется в качестве начальной информации;
- в каком классе решающих правил будет отыскиваться решение;
- какие дополнительные свойства множества допустимых объектов и функций должны быть учтены;
- как будет осуществляться обучение (естественно предполагать, что используется конечный компьютер или шире – вычислимые функции, но в теории машинного обучения построения зачастую выходят за рамки указанных классов); иначе говоря, определить алгоритм обучения как отображение начальной информации в некоторое множество решающих правил;
- как оценивать качество обучения;
- как определять, существует ли возможность достижения требуемого качества обучения при перечисленных условиях (имеет ли место обучаемость).

Если машинное обучение предполагает нахождение характеристических функций множеств, то такое обучение обычно называют обучением распознаванию. Собственно распознавание состоит в применении найденных в процессе обучения решающих правил для определения принадлежности рассматриваемым множествам объектов, не содержавшихся в начальной информации.

Уточнения неформальной постановки приводит к большому числу специфических задач машинного обучения и распознавания. Попытка представить классификацию таких задач была предпринята в статьях [4]. Важно отметить, что получить уточнения задачи машинного обучения по всем перечисленным выше пунктам удаётся не всегда.

Далее будет *рассматриваться задача машинного обучения распознаванию по прецедентам* (примерам) в соответствии с принципом эмпирической индукции (обобщения) в следующей постановке. Множество допустимых объектов X, называемое *признаковым пространством*, состоит из векторов (или *точек признакового пространства*) $\tilde{x} = (x_1,\ldots,x_n)$, значения координат которых в совокупности представляют описания объектов. Предпо-

лагается, что *на множестве* X *существует вероятностное распределение* P. Вид этого распределения будет полагаться неизвестным. Неизвестная, но *существующая* (*целевая*) *функция* $\varphi : \mathrm{X} \to \{0,1\}$ принадлежит некоторому *семейству* Φ, которое также является неизвестным. Требуется, используя начальную информацию – обучающую выборку длины l, извлечь из выбранного заранее класса решающих правил Ψ такую функцию $\psi : \mathrm{X} \to \{0,1\}$, которая как можно более точно приближает неизвестную целевую функции φ. Качество найденной в процессе обучения функции ψ в рассматриваемом случае можно представить как вероятностную меру несовпадения целевой функции φ с найденной в результате обучения функцией ψ. Проще говоря – как *вероятность ошибки функции* ψ, которая может быть выражена при помощи интеграла Лебега при условии измеримости соответствующих функций:

$$Err_{\psi} = \int_{\tilde{x} \in \mathrm{X}} |\psi(\tilde{x}) - \varphi(\tilde{x})| dP(\tilde{x}).$$

Чем меньше вероятность ошибки Err_{ψ} выбранного при обучении решающего правила ψ, тем лучше результат обучения. Но величину Err_{ψ} определить точно невозможно, поскольку неизвестна целевая функция φ и в подавляющем большинстве случаев неизвестна вероятностная мера P. Поэтому в статистической теории обучения используются подходящие оценки вероятности Err_{ψ} снизу и сверху.

Обучающая выборка $X_l = \{(\tilde{x}_j, \alpha_j)\}_{j=1}^{l}$ состоит из *примеров* – пар «точка – значение неизвестной функции в этой точке»: $\alpha_j = \varphi(\tilde{x}_j)$. Точки, входящие в выборку, извлекаются из множества X случайно и независимо в соответствии с распределением P. Естественно потребовать, чтобы с ростом длины обучающей выборки (с увеличением числа обучающих примеров) величина Err_{ψ} стремилась к нулю. В общих чертах это характеризует *обучаемость*, как возможность достижения нужной точности извлекаемой в процессе обучения решающей функции ψ.

Понятие обучаемости возможно точно определить не единственным способом, и это приводит к существенным различиям в постановке задачи и построении моделей обучения. Если $\Pr(Err_{\psi} > \varepsilon) < \delta$, где $\delta = \delta(l, \varepsilon)$, то величину ε называют точностью, а $(1 - \delta)$ – надежностью оценки выбранного решающего правила ψ.

Процесс машинного обучения может быть упрощенно представлен схемой (рис.1), в соответствии с которой следует обратить внимание на следующие обстоятельства.

Выборка может быть извлечена различными способами, и это должно уточняться – должна быть определена *схема извлечения выборки*.

Результат обучения – решающая функция ψ – может быть извлечена из семейства Ψ различными методами. Понятие *метода* или *алгоритма обучения* является центральным, поскольку главным образом именно его выбор определяет: будет ли иметь место обучаемость. Алгоритм обучения управляет процессом выбора решения ψ, используя обучающую выборку. С точки зрения постановки задачи, предполагая компьютерную реализацию, целесообразно говорить именно об *алгоритме обучения*. А с точки зрения центральной роли этого алгоритма в схеме машинного обучения, представляется возможным применение термина «*метод обучения*» [3]. Далее будет использоваться термин «алгоритм обучения».

Любой алгоритм обучения A представляет собой отображение множества всех допустимых обучающих выборок во множество $\mathrm{Im}\, A \subseteq \Psi$ – образ отображения A.

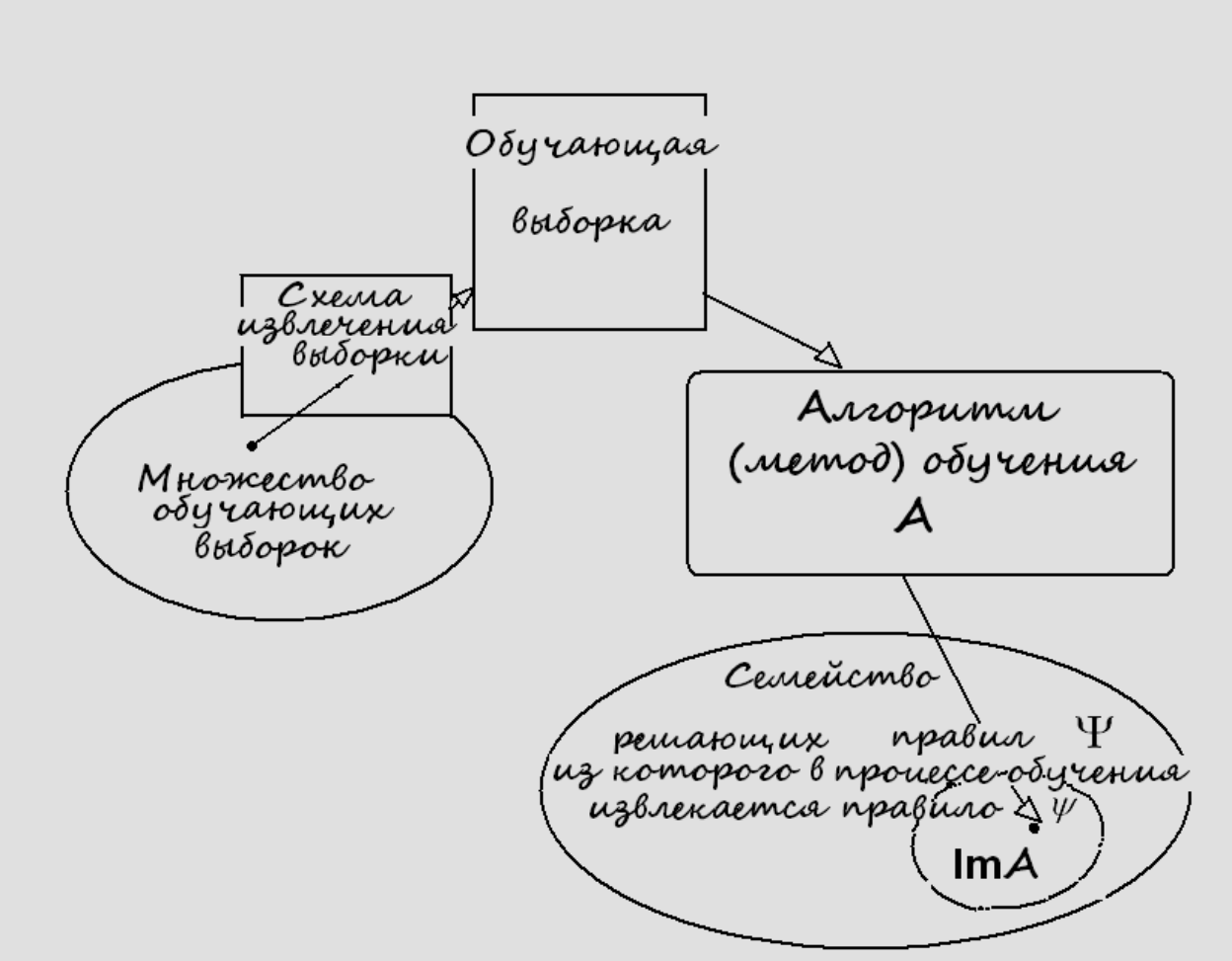

Рис.1. Схематическое представление процесса обучения

Будем называть приведенное выше уточнение задачи машинного обучения *функциональным*. В большинстве современных научных работ, посвященных машинному обучению, даётся другое – теоретико-множественное уточнение.

Концептами называют собственные подмножества X. Классом концептов называют семейство $H \subseteq 2^{\mathrm{X}}$ концептов. В дальнейшем полагается, что семейства концептов состоят из борелевских множеств. Задание классифицирующей функции $\varphi : \mathrm{X} \to \{0,1\}$ взаимно-однозначно определяет концепт h_φ как множество $h_\varphi = Dom_1(\varphi) = \{\tilde{x} \in \mathrm{X} : \varphi(\tilde{x}) = 1\}$. Множество, на котором функция φ принимает значение 0, является дополнением концепта h_φ во множестве X. *Примером* (*обучающим примером*) концепта $h \in H$ называют пару $(\tilde{x}, \alpha)$, где $\alpha = 1$, если $\tilde{x} \in h$, и $\alpha = 0$, если $\tilde{x} \notin h$. *Выборка* – это множество примеров некоторого концепта. Длина выборки – это число содержащихся в ней примеров.

Если класс концептов H является перечислимым $(h_1, h_2, ...)$, то его можно представлять перечислением конечных бинарных строк $s(h_1), s(h_2), ...$, определенным образом описывающих входящие в класс концепты. Такой подход позволяет рассматривать сложность концепта как длину кратчайшей описывающей его строки и использовать другие (не кратчайшие) строки-описания с целью оценивания сложности концепта.

Некоторый выделенный концепт $g_\varphi \in G$ называют целевым, а соответствующую ему функцию φ – целевой. Целевая функция полагается неизвестной и принадлежащей некоторому классу $\Phi = \Phi(G)$.

Обучающий алгоритм A использует выборку X_l длины $l = l_A(\varepsilon, \delta)$ в соответствии с вероятностным распределением P на X и вычисляет концепт-гипотезу $h_A = h_A(S) \in H$ по этой обучающей выборке.

Таким образом, имеет место следующее соответствие (табл. 1):

Табл. 1. Классифицирующие функции и концепты

Неизвестная заранее целевая классифицирующая функция φ	Неизвестный целевой концепт – множество $g_\varphi = Dom_1(\varphi)$
Неизвестный класс функций Φ, которому принадлежит функция φ	Неизвестный класс концептов G, содержащий целевой концепт g_f; $G = \bigcup_{\{\varphi : \varphi \in \Phi\}} Dom_1(\varphi)$
Решающая функция ψ – результат обучения	Результирующий концепт – множество $h_\psi = Dom_1(\psi)$
Известный, заранее выбранный класс функций Ψ, из которого в процессе обучения извлекается функция ψ	Известный, заранее выбранный класс концептов H, содержащий результирующий концепт h_ψ; $H = \bigcup_{\{\psi : \psi \in \Psi\}} Dom_1(\psi)$

Из приведенной таблицы видно, что использование концептов приводит к постановке задачи обучения на теоретико-множественной основе, которая эквивалентна постановке этой же задачи на функциональном подходе. Оба подхода имеют свои преимущества, и в силу эквивалентности представленных в таблице теоретико-множественных и функциональных описаний их можно и нужно использовать по мере проявления нужных преимуществ.

2. Обучаемость

Говоря неформально, понятие обучаемости необходимо для того, чтобы иметь возможность находить ответ на вопрос: удастся ли при некоторых заданных алгоритмах обучения и семействах функций, из которых извлекается решающее правило, достигнуть приближения этого правила к неизвестной целевой функции с нужной точностью? Т. е. можно ли в результате обучения получить достаточно точную аппроксимацию неизвестной целевой функции?

Для понятия обучаемости существует ряд различных определений.

Определение 1 (***PAC***-learning***, Probably Approximately Correct-learning***).

1) Будем говорить, что класс концептов $G \subset 2^X$ является PAC-*обучаемым* (или (ε,δ)-*обучаемым*) с использованием класса концептов $H \subset 2^X$, если найдется (обучающий) алгоритм A, который *при любом вероятностном распределении* P на X, *при любом целевом концепте* $g \in G$, для любых $\varepsilon,\delta : 0 < \varepsilon,\delta < \frac{1}{2}$, вычисляет по обучающей выборке X_l, извлеченной в соответствии с распределением P на X, концепт-гипотезу h_A, и при этом существует функция $l = l(\varepsilon,\delta)$, которая определяет длину обучающей выборки, обеспечивающую выполнение неравенства

$$\Pr\{P(h_A \Delta g) \le \varepsilon\} \ge 1-\delta,$$

где $h_A \Delta g = (h_A \setminus g) \cup (g \setminus h_A)$, а $\Pr\{Z\}$ – вероятность того, что событие Z – истинно.

Классы концептов H и G, вообще говоря, могут совпадать. В этом случае будем называть алгоритм обучения A *собственным или согласованным с целевым концептом*: $A(X_l) \in G$. Вариант модели *PAC* обучаемости, когда целевой концепт g заведомо содержится в используемом для обучения классе концептов H, называют *реализуемой PAC* моделью (*The Realizable PAC Model*) или *правильной* PAC-обучаемостью [15].

2) *Полиномиальная* PAC *обучаемость* (***RBPAC*** *– Resource Bounded PAC*) при всех перечисленных в первой части определения условиях дополнительно требует, чтобы алгоритм A обеспечивал (ε,δ)-обучение (выполнялся) за число *шагов, ограниченное полиномом* от $1/\varepsilon$, $1/\delta$, числа n переменных-признаков, длины описания $s(H)$ класса концептов H, и также

использовал длину обучающей выборки, ограниченную полиномом от всех указанных величин.

Наименьшее число примеров, обеспечивающее полиномиальную PAC обучаемость называют *сложностью выборки* относительно алгоритма обучения A. □

Важно обратить внимание на то, что в определении *PAC*-обучаемости не оговариваются никакие (кроме сложностных в $RBPAC$) свойства алгоритма обучения. Может применяться любой удовлетворяющий определению алгоритм A. Но при этом область его значений как алгоритмического отображения точно не оговаривается: возможно, что она совпадает с классом концептов H, но не исключается, что она существенно ′уже класса H. При этом распределение вероятностей P на X может быть любым. В силу такой широкой трактовки понятия *PAC* обучаемости, необходимым и достаточным условием для её достижения является конечность VC размерности класса, из которого извлекается концепт.

Теорема 1 [6]. Класс концептов H является *PAC* обучаемым тогда и только тогда, когда $VCD(H) < \infty$.

Сложностные свойства алгоритма обучения, фигурирующие в $RBPAC$ модели, предназначены для гарантии эффективной (полиномиальной) реализуемости обучения. Многие авторы научных работ в области машинного обучения не уделяют внимания сложности обучающих алгоритмов, ограничиваясь только требованием их сходимости. Для $RBPAC$ обучаемости предыдущая теорема верна при условии полиномиальной сложности алгоритма обучения.

Алгоритм обучения (и решающее правило) называют *согласованными* (с обучающей выборкой), если решающее правило правильно классифицирует все примеры обучающей выборки. Если κ – число примеров, неправильно классифицируемых выбранным при обучении решающим правилом, а l – длина обучающей выборки, то величину $\nu_{emp} = \dfrac{\kappa}{l}$ называют эмпирической частотой ошибок. Согласованные алгоритмы обеспечивают выбор решающих правил, имеющих $\nu_{emp} = 0$. Будем говорить, что алгоритм обучения *частично согласован* с обучающей выборкой, если $\kappa > 0$. Тогда он согласован с некоторой подвыборкой длины $l - \kappa$.

Определение 2 (A*gnostic PAC-learning*). Пусть P – вероятностное распределение (неизвестное) на $X \times \{0,1\}$ и $g : X \to \{0,1\}$ – заранее неизвестная (целевая) функция. Пусть $H = \{h : X \to \{0,1\}\}$ – класс гипотез; $A(X_l) = h$ – гипотеза, которую, используя обучающую выборку $X_l = (\tilde{x}_j, \alpha_j)_{j=1}^{l}$, извлекает из H обучающий алгоритм A; $\alpha_j = g(\tilde{x}_j)$. Ошибка гипотезы h согласно мере P есть $Err(h) = P\{(\tilde{x}, \alpha) : h(\tilde{x}) \neq \alpha\}$. Эмпирическая ошиб-

ка гипотезы h есть $Err_l(h) = \frac{1}{l}\left|\{(\tilde{x},\alpha) \in X_l : h(\tilde{x}) \neq \alpha\}\right|$. Говорят, что имеет место *agnostic PAC обучаемость*, если для любых положительных $\varepsilon, \delta < 1$, *для любого распределения* P на $\mathrm{X} \times \{0,1\}$ можно указать такое значение $l = l(A, \varepsilon, \delta, H)$, что для любой случайно извлеченной в соответствии с P^l обучающей выборкой X_l длины l имеет место неравенство

$$\Pr\{Err(A(X_l)) - \inf_{h \in H} Err_l(h) \leq \varepsilon\} \geq 1 - \delta. \ \square$$

В определении A*gnostic PAC learning* не фигурирует класс, в котором содержится целевой концепт. Распределение вероятностей полагается произвольным и предполагается использование принципа минимизации эмпирического риска ($\inf_{h \in H} Err_l(h)$). По сравнению с *PAC* обучением, модель A*gnostic PAC learning* шире, но и для неё остаётся справедливым необходимое и достаточное условие обучаемости – конечность класса, в котором заведомо содержится образ алгоритма обучения ($\mathrm{Im}\, A$).

GSL обучаемость, определяемая далее, практически является A*gnostic PAC* обучаемостью – «едва заметным» её расширением в случае, когда верхняя грань семейства всевозможных вероятностных распределений не является достижимой.

Определение 3 (*Обобщенная статистическая обучаемость, GSL* [25]). При условиях, сформулированных в определении, *статистическая обучаемость* имеет место, если для любого $\varepsilon > 0$ можно указать такое значение длины обучающей выборки $l = l(A, \varepsilon, \delta, H)$, что

$$\sup_{P \in \mathcal{P}} \Pr\{Err(A(X_l)) - \inf_{h \in H} Err_l(h) \leq \varepsilon\} \geq 1 - \delta,$$

где $\mathcal{P}$ – всевозможные вероятностные распределения на $\mathrm{X} \times \{0,1\}$.

Рассмотрим ещё ряд определений обучаемости, встречающихся в научной литературе.

Определение 4 [25]. Будем говорить, что при обучении имеет место *равномерная сходимость независимо от распределений* (***DFUC***), если

$$\sup_{P \in \mathcal{P}} \int_{X_l \in \mathrm{X}^l} \{\sup_{h \in H} |Err(h) - Err_l(h)|\} dP^l \to 0 \ \text{ при } \ l \to \infty.$$

Определение 5 [21]. H называется ε-равномерным классом Гливенко-Кантелли, если

$$\sup_{P \in \mathcal{P}} \Pr\{\sup_{m \geq l} \sup_{h \in H} |Err(h) - Err_l(h)| > \varepsilon\} = 0.$$

Теорема 2 [21]. Пусть H – класс функций из X в $\{0,1\}$. Тогда H является равномерным классом Гливенко-Кантелли (***uGC***), если и только если $VCD(H) < \infty$.

Определение 6 [1,2]. (Двусторонняя) р*авномерная сходимость по Вапнику* (***VUC***) имеет место при обучении в классе решающих правил H, если для любого положительного $\varepsilon < 1$

$$\lim_{l\to\infty} \Pr\{\sup_{h\in H} | Err(h) - Err_l(h) | > \varepsilon\} = 0.$$

В этом определении независимость от вероятностного распределения не указана. Речь идет о некотором имеющемся распределении на $X \times \{0,1\}$, в соответствии с которым происходит случайное и независимое извлечение примеров в обучающую выборку. Однако полученное В. Н. Вапником достаточное условие равномерной сходимости – конечность $VCD(H)$ – не зависит от свойств распределения. Также независимым от свойств распределения является необходимое и достаточное условие равномерной сходимости [27, с. 57] для любой вероятностной меры.

$$\lim_{l\to\infty} \frac{G^H(l)}{l} = 0, \qquad (1)$$

где $G^H(l) = \ln \sup_{(\tilde{x}_1,\ldots,\tilde{x}_l)\in X^l} N^H(\tilde{x}_1,\ldots,\tilde{x}_l)$ – функция роста семейства H, а $N^H(\tilde{x}_1,\ldots,\tilde{x}_l)$ – число способов разбиения выборки на два класса гипотезами семейства H. *Если условие* (1) *не выполняется, то найдётся вероятностная мера на* $X \times \{0,1\}$, *для которой равномерная сходимость по Вапнику не будет иметь места* [27, с. 72].

При выполнении достаточного условия равномерной сходимости по классу гипотез H – ограниченности $VCD(H)$ – выбор любой гипотезы $h \in H$, минимизирующей эмпирический риск, с ростом длины обучающей выборки будет *гарантировать* со сколь угодно большой вероятностью $1-\delta$ сколь угодно малое отклонение вероятности ошибки выбранной гипотезы h от её эмпирической ошибки на обучающей выборке. Причем ограниченность $VCD(H)$ гарантирует равномерную сходимость при любом вероятностном распределении P на $X \times \{0,1\}$. Конечность $VCD(H)$ перестаёт быть необходимым условием, если не требовать выполнения равномерной сходимости *для любых распределений*. Так, в работе [23] рассматривается обучаемость в случае неатомических (диффузных) вероятностных мер, и такое сужение условий приводит к некоторому *новому определению модулярной* VC *размерности* $VC(H \bmod \omega_1)$, которая, вообще говоря, может быть конечной при $VCD(H) = \infty$.

Одним из подходов к получению оценок ошибок алгоритмов обучения (эмпирического обобщения) является оценивание их устойчивости. *Под устойчивыми обучающими алгоритмами понимаются такие, которые извлекают гипотезы, незначительно изменяющиеся при малом изменении обучающей выборки*. Получаемые при таком подходе оценки оказываются неза-

висимыми от VC размерности используемого пространства гипотез [8,9], а скорее зависят от того, *как алгоритм обучения осуществляет поиск* в этом пространстве, и поэтому можно рассчитывать на обучаемость в случае, когда пространство гипотез имеет бесконечную VC размерность. Но при этом следует оговаривать, о каком определении обучаемости идет речь.

Введение в определение обучаемости дополнительных свойств алгоритма обучения влечёт сужение этого определения, выделяет частный случай из множества ситуаций, когда алгоритм обучения является произвольным, и может ослабить необходимые и достаточные условия обучаемости.

Подход на основе устойчивости обучающих алгоритмов требует введения некоторых окрестностей для выборки (в пространстве обучающих выборок) и для выбираемой гипотезы (в пространстве гипотез). В этом плане он близок к подходу, основанному на оценке подмножества используемых гипотез, которое в силу свойств выбранного алгоритма обучения может оказаться гораздо более узким по сравнению со всем пространством гипотез.

Естественно считать малым изменением заданной выборки удаление из неё ровно одного примера (или замену в ней ровно одного примера на произвольный другой пример). Всевозможные такие удаления образуют своеобразную окрестность выборки. Её называют *Loo* окрестностью (*Leave-one-out*). Обучение в окрестности данной выборки приводит к отбору алгоритмом обучения, вообще говоря, различных гипотез, близость которых можно оценивать, сравнивая частоты ошибок этих гипотез на выборке.

Пусть α – истинное значение целевой функции в точке $\tilde{x}$, а $h = A(X_l)$ – выбранная обучающим алгоритмом A по выборке $(\tilde{x}_j, \alpha_j)_{j=1}^{l}$ длины l решающая функция. Определим функцию потерь (ошибку):

$$\lambda(h, \tilde{x}) = \begin{cases} 0, & h(\tilde{x}) = \alpha; \\ 1, & h(\tilde{x}) \neq \alpha. \end{cases}$$

Обозначим X_l^j обучающую выборку, из которой удалён пример $(\tilde{x}_j, \alpha_j)_{j=1}^{l}$, и $A(X_l^i)$ – найденное обучающим алгоритмом A по этой укороченной на единицу выборке X_l^j решающее правило h.

Определение 7. *Loo*-ошибкой называется усреднённая по всем примерам обучающей выборки $(\tilde{x}_j, \alpha_j)_{j=1}^{l}$ величина функции потерь

$$\frac{1}{l}\sum_{j=1}^{l} \lambda(A(X_l^j), \tilde{x}_j) \ .$$

Определение 8 [20]. Алгоритм обучения A называется CV_{Loo} устойчивым (*Leave-one-out Cross-Validation*) независимо от распределения, если *для любой вероятностной меры*, для любой длины выборки l, для любой точки $\tilde{x}_j$ найдутся такие положительные $\varepsilon(l), \delta(l) < 1$, что

$$\forall j \in \{1,...,l\} \quad \Pr\{\left|\lambda(A(X_l^j),\tilde{x}_j) - \lambda(A(X_l),\tilde{x}_j)\right| \le \varepsilon(l)\} \ge 1-\delta(l),$$

где $\varepsilon(l) \to 0$ и $\delta(l) \to 0$ при $l \to \infty$.

Согласно определению, CV_{Loo} устойчивость предполагает сколь угодно близкую точность обучающего алгоритма в *Loo* окрестности обучающей выборки с ростом её длины l для каждого из l вариантов извлечения одного примера.

Определение 9 [20]. Алгоритм обучения A называется $ELoo_{err}$ устойчивым независимо от распределения, если для любой вероятностной меры найдутся такие положительные $\varepsilon, \delta < 1$, что

$$\Pr\{|\, Err_l(A(X_l)) - \frac{1}{l}\sum_{j=1}^{l} \lambda(A(X_l^j),\tilde{x}_j) \,| \le \varepsilon\} \ge 1-\delta.$$

В отличии от определения, $ELoo_{err}$ устойчивость предполагает сколь угодно близкую точность обучающего алгоритма и средней точности по *Loo* окрестности.

Определение 10. Алгоритм обучения A называется *LOO* устойчивым, если он является CV_{Loo} и $ELoo_{err}$ устойчивым (по каждому малому «отклонению» и в среднем по малой окрестности).

Теорема 3 [20]. CV_{Loo} устойчивость алгоритма обучения A является необходимым и достаточным условием равномерной сходимости частоты ошибок решающего правила $h(A)$, выбранного из заданного семейства H при обучении методом минимизации эмпирического риска, к вероятности его ошибки.

Доказательство этой теоремы приведено в [20].

Различные определения обучаемости, приведенные выше, некоторым образом связывались с семействами гипотез. Но говорить об обучаемости можно и в более общей постановке как о возможности эмпирического обобщения.

Определение. *Универсальное эмпирическое обобщение* (*universal generalization*) *имеет место, если для любой выбранной алгоритмом обучения гипотезы частота ошибки этой гипотезы на обучающей выборке сходится по вероятности к её математическому ожиданию при неограниченном росте длины обучающей выборки независимо от вероятностного распределения*, то есть

$$\forall \varepsilon > 0 \quad \Pr\{\left|Err(A(X_l)) - Err_l(A(X_l))\right| \ge \varepsilon\} \to 0 \text{ при } l \to \infty$$

для любой гипотезы $A(X_l)$ и для любой меры P^l.

Теорема 4 [20]. *LOO* устойчивость алгоритма обучения при условии ограниченности функции потерь является достаточным условием для обеспечения универсального эмпирического обобщения.

Доказательство. Оценим математическое ожидание квадрата отклонения математического ожидания ошибки решающего правила (гипотезы) $h = A(X_l)$, выбранной *LOO* устойчивым алгоритмом обучения A, от эмпирической ошибки этой гипотезы. И распределение P^l, и семейство H, которому принадлежит гипотеза h, полагаются любыми.

$$\mathbf{E}_l(Err(A(X_l)) - Err_l(A(X_l)))^2 = \mathbf{E}_l(Err(h) - Err_l(h))^2 =$$

$$= \mathbf{E}_l\left(Err(h) - \frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j), \tilde{x}_j) + \frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j), \tilde{x}_j) - Err_l(h) \right)^2$$

$$\le 2\mathbf{E}_l\left(Err(h) - \frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j), \tilde{x}_j) \right)^2$$

$$+ 2\mathbf{E}_l\left(\frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j), \tilde{x}_j) - Err_l(h) \right)^2 .$$

При переходе к неравенству использовано тождество $(a+b)^2 \le 2a^2 + 2b^2$. Оценим второе слагаемое

$$2\mathbf{E}_l\left(\frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j), \tilde{x}_j) - Err_l(A(X_l)) \right)^2$$

$$= 2\mathbf{E}_l\left(\frac{1}{l}\sum_{i=1}^{l}\lambda(A(X_l), \tilde{x}_j) - \frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j), \tilde{x}_j) \right)^2$$

$$= 2\mathbf{E}_l \frac{1}{l^2}\left| \sum_{j=1}^{l}[\lambda(A(X_l), \tilde{x}_j) - \lambda(A(X_l^j), \tilde{x}_j)] \right|^2$$

$$\le 2M\mathbf{E}_l \frac{1}{l}\left| \sum_{i=1}^{l}[\lambda(A(X_l), \tilde{x}_j) - \lambda(A(X_l^j), \tilde{x}_j)] \right|$$

(здесь использовано условие ограниченности функции потерь, в силу которого $\left| \sum_{j=1}^{l}[\lambda(A(X_l), \tilde{x}_j) - \lambda(A(X_l^j), \tilde{x}_j)] \right| \le M \cdot l$, где M – константа)

$$\le 2M\mathbf{E}_l \frac{1}{l}\sum_{j=1}^{l}\left| \lambda(A(X_l), \tilde{x}_j) - \lambda(A(X_l^j), \tilde{x}_j) \right|$$

$$= 2M \frac{1}{l}\sum_{j=1}^{l}\mathbf{E}_l\left| \lambda(A(X_l), \tilde{x}_j) - \lambda(A(X_l^j), \tilde{x}_j) \right|$$

$$= 2M\mathbf{E}_l\left| \lambda(A(X_l), \tilde{x}_j) - \lambda(A(X_l^j), \tilde{x}_j) \right|$$

Окончательно получаем неравенство

$$\mathbf{E}_l(Err(A(X_l)) - Err_l(A(X_l)))^2$$
$$\le 2\mathbf{E}_l\left(Err(h) - \frac{1}{l}\sum_{j=1}^{l}\lambda(A(X_l^j),\tilde{x}_j)\right)^2$$
$$+ 2M\mathbf{E}_l\left|\lambda(A(X_l),\tilde{x}_j) - \lambda(A(X_l^j),\tilde{x}_j)\right|,$$

в правой части которого содержатся два слагаемых. Первое слагаемое соответствует определению $ELoo_{err}$ устойчивости, а второе – CV_{Loo} устойчивости. Если оба эти слагаемые при $l \to \infty$ одновременно стремятся к нулю, то, согласно определению, имеет место LOO устойчивость, что влечёт эмпирическое обобщение, поскольку сумма указанных слагаемых является верхней оценкой вероятности математического ожидания ошибки выбранной гипотезы от её эмпирической ошибки. □

Существуют и другие походы к определению устойчивости алгоритмов обучения.

Определение 11. Пусть в обучающей выборке $X_l = \{(\tilde{x}_1,\alpha_1),...,(\tilde{x}_j,\alpha_j),...,(\tilde{x}_l,\alpha_l)\}$ произведена замена ровно одного примера $(\tilde{x}_i,\alpha_i)$ на некоторый другой пример $(\tilde{x},\alpha)$. Будем обозначать полученную после такой замены выборку $\overleftarrow{X}_l^i$ и говорить, что $\overleftarrow{X}_l^i$ получена из X_l по правилу ***RO*** (*Replace one*).

Определение 12.

1. Обучающий алгоритм A называется *равномерно **RO** устойчивым* на уровне $\varepsilon_{stable}(l)$, если для всех возможных $\overleftarrow{X}_l^i$ и любого замещающего примера $(\tilde{x},\alpha)$

$$\frac{1}{l}\sum_{i=1}^{l}\left|\kappa(A(\overleftarrow{X}_l^i);(\tilde{x},\alpha)) - \kappa(A(X_l);(\tilde{x},\alpha))\right| \le \varepsilon_{stable}(l),$$

где $\kappa(\cdot)$ – число ошибок гипотезы, извлеченной обучающим алгоритмом A при некоторой заданной обучающей выборке.

2. Обучающий алгоритм A называется ***RO*** *устойчивым в среднем* на уровне $\varepsilon_{stable}(l)$, если

$$\left|\frac{1}{l}\sum_{i=1}^{l}\int_{X_l\in X^l}(\kappa(A(\overleftarrow{X}_l^i);(\tilde{x},\alpha)) - \kappa(A(X_l);(\tilde{x},\alpha)))dP^l(X_l)\right| \le \varepsilon_{stable}(l).$$

3. *Универсальной **RO*** устойчивостью в среднем называется ***RO*** устойчивость в среднем для любого вероятностного распределения P.

Определение 13. Алгоритм обучения A называется ***AERM*** *правилом* (*Asymptotic Risk Minimizer*), если

$$\int_{X_l\in X^l}(Err(A(X_l)) - \inf_{h\in H} Err_l(h))dP^l \le \varepsilon_{erm}(l),$$

и называется *универсальным* ***AERM*** *правилом*, если ***AERM*** имеет место для любого вероятностного распределения P. В этом случае говорят, что имеет место *универсальная* ***AERM*** *устойчивость.*

Определения устойчивости алгоритмов обучения, основанные на замене одного из примеров обучающей выборки некоторым другим примером (***RO***), достаточно схожи с определениями ***LOO***. Их различие проявляется в некоторых результах обучения при помощи соответствующих алгоритмов [25].

Теорема 4 [25, с. 33]. *При использовании* ***AERM*** *правила универсальная* ***RO*** *устойчивость в среднем является необходимым и достаточным условием для обеспечения универсального эмпирического обобщения.*

Примеры устойчивых алгоритмов представлены в ряде научных работ. А. Елисеевым показана устойчивость алгоритма построения линейной регрессии [9,10] с использованием правила ***RO*** согласно следующему определению.

Определение 14 [9]. Обучающий алгоритм A называется β-устойчивым относительно неотрицательной вещественной функции потерь λ, если

$$\forall X_l \forall X_l^{i,\tilde{u}} \in \mathrm{X}^l, \forall \tilde{x} \in \mathrm{X} \quad |\lambda(A(X_l),\tilde{x}) - \lambda(A(X_l^{i,\tilde{u}}),\tilde{x})| \le \beta,$$

где $X_l^{i,\tilde{u}}$ – выборка, полученная из выборки X_l путём замены в ней $i-$го примера на некоторый другой пример $\tilde{u}$ (правило ***RO***).

Теорема 5 [9]. Пусть A есть β-устойчивый обучающий алгоритм, функция потерь удовлетворяет условию $0 \le \lambda(A(X_l),\tilde{x}) \le M$ для любой обучающей выборки X_l и любого $\tilde{x} \in \mathrm{X}$. Тогда для любых $\varepsilon > 0$ и $l \ge 1$ имеет место неравенство

$$P^l\{|\, Err_l(A(X_l)) - Err(A(X_l))\,| > \varepsilon + 2\beta M\} \le \exp\left(-\frac{2l\varepsilon^2}{(4l\beta + M)^2}\right),$$

и с вероятностью $1-\delta$, где $\delta \to 0$ при $l \to \infty$,

$$Err(A(X_l)) \le Err_l(A(X_l)) + 2\beta + (4l\beta + M)\sqrt{\frac{\ln \delta^{-1}}{2l}}.$$

Из последней теоремы видно, что *обучаемость может иметь место независимо от ёмкости класса гипотез* H*, которому принадлежат полученные* β*-устойчивым алгоритмом обучения решающие правила* $A(X_l)$.

Доказательство этой теоремы основано на следующей теореме Мак-Дьярмида:

Теорема 6 [19]. Пусть X_l – произвольная выборка, а $X_l^{i,\tilde{u}}$ – выборка, полученная из X_l по правилу ***RO***. Пусть $F : \mathrm{X}^l \to \mathrm{R}$ – любая измеримая функция и найдутся константы c_i, $i = 1,...,m$, такие что

$$\sup_{X_l \in \mathrm{X}^l, \tilde{u} \in \mathrm{X}} |F(X_l) - F(X_l^{i,\tilde{u}})| \le c_i.$$

Тогда

$$P^l\{|F(X_l) - \mathbf{E}_l[F(X_l)]| > \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{l} c_i^2}\right).$$

Буск*е* и Елисеев показали $\beta-$устойчивость тихоновской регуляризации при построении регрессии. Им же принадлежит результат об устойчивости *SVM – Support Vector Machine* [9].

Для метода потенциальных функций устойчивость установлена в работе [11]. В работе Р. Рифкина [24] показана устойчивость баггинга. Это результат не представляется неожиданным, поскольку можно было предположить, что использование совокупности решающих правил с усреднением должно повлечь устойчивость решений. Не рассматривая подробно устойчивость баггинга, отметим только, что в упомянутой работе Рифкина используется несколько отличающееся от $\beta-$устойчивости определение $\alpha-$устойчивости, применяемое для случая, когда решающие правила не являются бинарными, а принимают вещественные значения.

Определение 15 [24]. Обучающий алгоритм A называется α-устойчивым, если

$$\forall X_l \forall X_l^{i,\tilde{u}} \in \mathrm{X}^l, \forall \tilde{x} \in \mathrm{X} \quad |A(X_l)(\tilde{x}) - A(X_l^{i,\tilde{u}})(\tilde{x})| \le \alpha,$$

где $X_l^{i,\tilde{u}}$ – выборка, полученная из выборки X_l путём замены в ней $i-$го примера на некоторый другой пример $\tilde{u}$.

Определение α-устойчивости, в котором оцениваются построенные алгоритмом обучения решающие правила (функция риска не фигурирует), оказалось более удобным для выполнения операций усреднения при использовании машинного обучения для построения регрессии.

3. Сравнение моделей и условия обучаемости

Различные определения обучаемости и устойчивости сведены ниже в таблицу 2 для их сравнительного анализа. Из таблицы видно, что в зависимости от определения обучаемости может быть явно указано или нет, в каком семействе (G) содержится целевой концепт, и из какого семейства (H) из-

влекается гипотеза. Например, в определении *PAC* обучения эти два семейства содержатся. А в определении *Realizable PAC* обучения даже предполагается, что $G = H$.

Табл. 2. Определения и модели обучаемости

Определение обучаемости	*В каком семействе содержится целевой концепт*	*Из какого семейства извлекается гипотеза*	*Для некоторой фиксированной или для любой вероятностной меры*	*Требования к алгоритму обучения*	*Другое*
PAC	*G*	*H*	для любой	–	необходимое и достаточное условие – $VCD(H)<\infty$
Realizable PAC	*H*	*H*	для любой	–	необходимое и достаточное условие – $VCD(H)<\infty$
Poly PAC	*G*	*H*	для любой	$A \in PTIME$	
Realizable Poly PAC	*H*	*H*	для любой	$A \in PTIME$	
Agnostic PAC	–	*H*	некоторая фиксированная	–	достаточное условие – $VCD(H)<\infty$
Равномерная сходимость по Вапнику (*VUC*)	–	*H*	некоторая фиксированная	–	достаточное условие – $VCD(H)<\infty$
Равномерный класс Гливенко-Кантелли	–	*H*	Для любой равномерно	–	необходимое и достаточное условие – $VCD(H)<\infty$
LOO устойчивость	–	–		устойчивость в малой окрестности выборки	
Универсальная *PO* устойчивость	–	–		устойчивость в малой окрестности выборки	
Универсальное эмпирическое обобщение	–	–	для любой	–	*LOO* устойчивость – достаточное условие; универсальная *PO* устойчивость – необходимое и достаточное условие

В теории В. Н. Вапника в определении равномерной сходимости фигурирует только семейство H. Универсальное эмпирическое обобщение не оговаривает явно ни семейство G, ни семейство H. Тем не менее, при лю-

бом подходе к машинному обучению его результатом является некоторая выбранная алгоритмом A гипотеза $h = h(A, X_l)$. Для разных обучающих выборок $X_l \in \mathrm{X}^l$ эта выбранная гипотеза, вообще говоря, может оказаться различной. Поэтому $h \in S(A, \mathrm{X}^l) \subseteq H$, где $S(A, \mathrm{X}^l) = \mathrm{Im}\, A$ – множество всевозможных порождаемых алгоритмом A гипотез, а H – любой содержащий это множество класс, имеющий некоторое точное математическое определение. На практике семейство H непосредственно определено выбором для решения задачи машинного обучения некоторой модели: нейронных сетей, решающих деревьев, SVM или др. Но именно алгоритм обучения A определяет сужение $S(A, \mathrm{X}^l)$, оценка ёмкости которого $VCD(S(A, \mathrm{X}^l))$ не превышает $VCD(H)$, и чем она меньше $VCD(H)$, тем точнее окажется оценка обучаемости, использующая VC размерность.

Считается, что фундаментальным результатом статистической теории обучения является следующий строго доказанный факт [5,23]. Если H – класс концептов (решающих правил) над проблемной областью X с произвольной вероятностной мерой и выполняются все необходимее условия измеримости, то *следующие три утверждения эквивалентны*:

i. Для класса H имеет место PAC обучаемость для любой вероятностной меры на X.
ii. H является равномерным классом Гливенко-Кантелли.
iii. $VCD(H)$ является конечной.

Рассмотренные выше подходы к определению обучаемости и устойчивости и полученные на их основе результаты позволяют расширить представление о статистической теории обучения.

Теория равномерной сходимости, PAC обучаемость и универсальная способность к обобщению представляют собой достаточно широко определённые модели. В них не оговариваются ни свойства распределения вероятностей, ни особенности алгоритма обучения, которые могут быть произвольными.

Фиксация свойств алгоритма обучения (в частности, его заведомая устойчивость) позволяют сузить модель обучения и вследствие этого получить обучаемость даже в случае бесконечной VC размерности семейства гипотез, в которое вложен образ $\mathrm{Im}\, A$ алгоритма обучения A.

Конечность VC размерности также перестаёт быть необходимым условием в некоторых случаях при конкретизации вероятностной меры (например, в случае диффузных или атомарных мер).

Дополнительно выявленные фундаментальные положения дают объяснение практически наблюдаемой обучаемости при использовании некоторых алгоритмов и моделей обучения, несмотря на кажущееся противоречие с VC теорией: в действительности этого противоречия нет.

Литература

1. Вапник В. Н. Восстановление зависимостей по эмпирическим данным / В. Н. Вапник. – М. Наука, 1979. – 447 с.
2. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов / В. Н. Вапник, А. Я. Червоненкис. – М.: Наука, 1974. – 416 с.
3. Воронцов К. В. Обзор современных исследований по проблеме качества обучения алгоритмов / К. В. Воронцов // Таврический вестник информатики и математики, 2004. – № 1. – С. 5–24.
4. Донской В. И. Эмпирическое обобщение и распознавание: классы задач, классы математических моделей и применимость теорий. Часть I; Часть II / В. И. Донской // Таврический вестник информатики и математики, 2011. – №1. – С. 15 – 26; №2. – С. 31 – 42.
5. Blumer A. Learnability and the Vapnik-Chervonenkis Dimension / A. Blumer, A.Ehrenfeucht, D. Haussler, M. Warmuth // J. Assoc. Comp. Mach., 1989. – 35. – P. 929 – 965.
6. Blumer A. Occam's Razor / A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth // Information Processing Letters, 1987. – Vol. 24(6). – P.377 – 380.
7. Blumer A., Littlestont N. Learning faster than promises by the Vapnik-Chervonenkis dimension / Anselm Blumer, Nick Littlestone // Discrete Applied Mathematics, 1989. – Vol. 24. – Iss. 1-3, – P. 47 – 63.
8. Bousquet O., Elisseeff A. Algorithmic Stability and Generalization Performance / Olivier Bousquet , André Elisseeff // Advances in Neural Information Processing Systems. – 2001. – 13. – P. 196 – 202.
9. Bousquet O., Elisseeff A. Stability and Generalization / Olivier Bousquet , André Elisseeff // Journal of Machine Learning Research. – 2002. – 2. – P. 499-526.
10. Elisseeff F. A Study About Algorithmic Stability and Their Relation to Generalization Performances // Andre Elisseeff. – Technical report. – Laboratoire ERIC, Univ. Lyon 2,2000. – 19 P.
11. Devroye L., Wagner T. Distribution-free performance bounds for potential function rules / Luc Devroye, T. Wagner // IEEE Transactions on Information Theory. – 1979. – 25. – P. 601 – 604.
    https://www.researchgate.net/publication/3083261_Distribution-free_performance_bounds_for_potential_function_rules
12. Ehrenfeucht A. A general lower bound on the number of examples needed for learning / A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant // Inform. Computations, 1989. – 82. – P. 247 – 261.
13. Floyd S., Warmuth M. Sample Compression, learnability, and the Vapnik-Chervonenkis dimension / Sally Floyd, Manfred Warmuth // J. Machine Learning, 1995. – Vol. 21. – Iss. 3. – P. 269 – 304.
14. Freund Y. Self bounded learning algorithms / Y. Freund // In Proc. Of the 11th Ann. Conf. on Computational Learning Theory (COLT-98). – N.Y.: ACM Press. – 1998. – P. 247 – 258.
15. Haussler D. Overview of the Probably Approximately Correct (PAC) Learning Framework / David Haussler // AAAI'90 Proceedings of the eighth National conference on Artificial intelligence, 1990. – Volume 2. – P. 1101–1108.
    http://www.cbse.ucsc.edu/sites/default/files/smo_0.pdf
16. Hutter M. Algoritmic complexity // Scholarpedia. – 2008. – 3(1):2573 http://www.scholarpedia.org/article/Algorithmic_complexity#Prefix_Turing_machine

17. Kearns M. J., Vazirani U. V. An Introduction to Computational Learning Theory / M. Kearns, U. Vazirani. – MIT Press 1994. – 221 p.
18. Littlestone L., Warmuth M. Relaring Data Compression and Learnability / Nick Littlestone, Manfred K. Warmuth. – Technical Report. – Santa-Cruz: University of California, 1986. – 13 p.
http://users.soe.ucsc.edu/~manfred/pubs/T1.pdf
19. McDiarmid C. On the method of bounded differences / Colin McDiarmid // In Surveys in Combinatorics. – Cambridge University Press, Cambridge, 1989. – London Math. Soc. Lectures Notes. – 141. **–** P. 148–188.
20. Mukherjee S. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization / Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin // Advances in Computational Mathematics. – 2006. – 25. – P. 161–193.
21. Noga A., Shai B. D. Scale-sentitive Dimensions, Uniform Convergence, and Learnability / Alon Noga, Ben David Shai // Journal of the ACM. – 1997. – 44(4). – p. 615 – 631.
22. Ogielski A. T. Information, Probability, and Learning from Examples. Survey / Andrew Ogielski. – Bell Communication Research, 1990. – 87 p.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.9797&rep=rep1&type=pdf
23. Pestov V. PAC learnability under non-atomic measures: a problem by Vidyasagar / Vladimir Pestov // 21st Int. Conf. "Algorithmic Learning Theory"(ALT 2010). – Canberra, Australia, 2010. – P. 134 – 147.
24. Rifkin M. R. Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning / Ryan Michael Rifkin. Ph.D. in Operation Research. Thesis, MIT, 2002. – 221 P.
25. Sridharan K. Learning from an Optimization Viepoint / Karthik Sridharan. – Thesis for degree of Philosophy in Computer Science. – Chicago:TTIC, 2012. – 217 p.
http://ttic.uchicago.edu/~karthik/thesis.pdf
26. Valiant L. G. A Theory of the Learnable / Leslie G. Valiant // Communications of the ACM, 1984. – Vol. 27. – N11. – P. 1134 – 1142.
27. Vapnik V. N. The Nature of Statistical Learning Theory / Vladimir N. Vapnik. – 2nd ed. – New York: Springer-Verlag, 2000. – 314 p.