Некоторые вопросы оценивания качества методов построения решающих функций

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

«Интеллектуализация обработки информации»
(ИОИ-10), Греция, г. Ираклион, 6–10 октября 2014 г.

Вопросы к рассмотрению

- Как определить, что один метод анализа данных лучше другого.
- Что означает: один метод анализа данных лучше другого.
- За счёт чего один метод может превосходить другой.
- За счёт чего бустинг зачастую превосходит другие методы.

Основные понятия

Пусть X – пространство значений переменных, используемых для прогноза,
$Y = \{0, 1\}$ – пространство значений прогнозируемых переменных,
C – множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$.

При каждом $c \in C$ имеем вероятностное пространство: $\langle D, B, \mathsf{P}_c \rangle$, где B – σ-алгебра, P_c – вероятностная мера.

Метод построения решающих функций

Решающей функцией (алгоритмом классификации) называется соответствие $\lambda\colon X \to Y$.
Качество принятого решения оценивается заданной функцией потерь $\mathcal{L}\colon Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери:

$$R(c, \lambda) = \mathsf{E}\mathcal{L}(y, \lambda(x)) = \int_D \mathcal{L}(y, \lambda(x))\ \mathsf{P}_c(dx, dy),$$

$x \in X$, $y \in Y$.

Отображение $Q\colon D^N \to \Lambda$ называется методом (алгоритмом) построения решающих функций.

Типичный пример сравнения методов

Таблица из одной из научных статей.

DATASET	Adaboost[27]	SVM [26]	LMKL [25]	PBGD3 [12]	SFM-GMM
Breast Cancer	93.24 ± 1.26	$\mathbf{96.80 \pm 1.79}$	96.41 ± 0.97	93.98 ± 1.52	95.26 ± 0.93
Breast Tissue	88.55 ± 5.91	80.45 ± 4.27	87.69 ± 5.24	88.14 ± 4.50	$\mathbf{89.61 \pm 3.84}$
Wine	92.98 ± 3.42	$\mathbf{97.73 \pm 1.86}$	95.48 ± 4.10	92.22 ± 12.63	96.11 ± 1.38
Sonar	70.87 ± 4.76	73.11 ± 3.25	80.21 ± 1.52	75.52 ± 5.70	$\mathbf{81.54 \pm 2.93}$
Credit Approval	84.74 ± 1.30	84.74 ± 1.79	81.92 ± 1.41	83.53 ± 1.82	$\mathbf{85.06 \pm 1.31}$
SPECTF Heart	78.65 ± 2.08	74.66 ± 3.56	80.38 ± 3.40	79.70 ± 0.65	$\mathbf{81.34 \pm 0.46}$
Libras Movement	92.97 ± 1.76	87.54 ± 7.01	$\mathbf{96.58 \pm 1.78}$	94.52 ± 2.80	95.97 ± 3.12
Steel Plates Faults	89.47 ± 9.08	86.43 ± 9.16	$\mathbf{92.63 \pm 8.14}$	87.30 ± 8.26	90.24 ± 8.23

- Нет уверенности, что все методы запускались при наилучших значениях параметров.
- Неизвестно, задачи выбраны наугад или отобраны.
- Непонятна область применимости метода.

Сравнение методов построения решающих функций

- Решение практических задач (kaggle.com).
- Выбор эталонного набора тестовых задач.
- Формальное введение понятия оптимальности метода.

Задача построения решающих функций близка задаче проверки статистических гипотез, при этом понятия оптимального статистического критерия (в случае отсутствия альтернатив) не существует.

Система «Полигон» — 1980-е

Лбов Г.С., Старцева Н.Г. Сравнение алгоритмов распознавания с помощью программной системы «Полигон» // Анализ данных и знаний в экспертных системах. Новосибирск, 1990. Вып. 134: Вычислительные системы. С. 56–66.

Принципы:

- для каждого метода включается «эталонная» задача,
- на «своей» задаче метод должен работать лучше других,
- возможно оценить степень универсальности метода,
- тестовая единица - таблица данных.

Система «Полигон» — 2000-е

Воронцов К.В., Ивахненко А.А., Инякин А.С., Лисица А.В., Минаев П.Ю. «Полигон» — распределённая система для эмпирического анализа задач и алгоритмов классификации // Всеросс. конф. Математические методы распознавания образов-14 - М.: МАКС Пресс, 2009. С. 503–506.

Принципы:

- использование реальных задач,
- большое число характеристик качества,
- основной критерий - скользящий экзамен.

Тестовые единицы

Возможные тестовые единицы:

- таблица данных,
- распределение,
- класс распределений.

Проблема определения оптимальности метода

Напомним, что метод — это отображение выборок в решения.

- Для таблицы данных понятие оптимального метода не имеет смысла.
- Для заданного распределения оптимальный метод вырожден — он любой выборке сопоставляет байесовское решающее правило.
- Об оптимальности метода можно говорить только для класса распределений.
- Даже для нормальных распределений оптимальный метод неизвестен.

Минимаксный подход к оцениванию качества

Усреднение по задачам не вполне содержательно.
Привлекателен минимаксный подход.

- Максимальное по классу распределений значение риска для всех методов одинаково.
- Вводить ограничения сверху на Байесовский уровень ошибки не имеет смысла.
- Использование максимума не абсолютного риска, а отнесённого к достижимому уровню, позволяет ввести осмысленное понятие метода, оптимального на классе распределений.

Достижимый уровень качества

- Интересует не Байесовский уровень ошибки, а тот, который реально достижим.
- Необходимо задавать для каждого распределения из класса.
- Определяется на основе эталонного метода.

Выбор классов распределений

- Класс распределений не должен быть ни узким ни широким — иначе получаем соответственно вырожденный метод или аналог NFL.
- Представительность для исследований: все значения достижимого риска, максимально смещённые оценки.
- Параметр сложности. Универсальность.
- Замкнутость относительно допустимых преобразований пространства переменных.

Варианты классов распределений

- Класс кусочно–постоянных распределений.
- Класс нормальных распределений.
- Класс, сформированный случайными решающими деревьями.
- Ядерные функции для условных вероятностей.

Исследование бустинга

Для иллюстрации подхода к исследованию качества методов построения решающих функций с использованием эталонных распределений рассмотрим метод бустинга.

Также будут приведены некоторые свойства метода AdaBoost, которые предположительно известны, но в предлагаемом простом изложении автору не встречались.

Алгоритм AdaBoost

В методе AdaBoost решение строится в виде композиции

$$\lambda(x) = sign(\beta(x)), \quad \beta(x) = \sum_{t=1}^{T} \alpha_t \lambda_t(x)),$$

где базовые классификаторы $\lambda_t(x)$ и их веса α_t находятся следующим образом.
Первый базовый классификатор строится базовым методом на основе исходной выборки, объектам которой приписаны начальные веса $w^1 = (w_1^1, ..., w_N^1)$.
Заметим, что мы будем задавать начальные веса объектом в соответствии с выбранным распределением, но в стандартном варианте метода начальные веса выбираются одинаковыми, т.е. $w_i^1 = \frac{1}{N}$.

Пересчёт весов

Вес построенного базового классификатора в композиции определяется по формуле

$$\alpha_t = \frac{1}{2} \ln \frac{\widetilde{M}^+(V, w^t, \lambda_t)}{\widetilde{M}^-(V, w^t, \lambda_t)},$$

где

$$\widetilde{M}^+(V, w, \lambda) = \sum_{i=1}^{N} w_i \cdot I(y^i = \lambda(x^i)),$$

$$\widetilde{M}^-(V, w, \lambda) = \sum_{i=1}^{N} w_i \cdot I(y^i = -\lambda(x^i)).$$

Итерационный процесс

Следующие базовые классификаторы строятся тем же базовым методом по выборке, веса объектов в которой вычисляются по формулам

$$w_i^{t+1} = \frac{\bar{w}_i^{t+1}}{\sum_{i=1}^{N} \bar{w}_i^{t+1}}, \qquad \bar{w}_i^{t+1} = w_i^t \cdot e^{-\alpha_t y^i \lambda_t(x^i)}.$$

Веса правильно классифицированных объектов умножаются на $e^{-\alpha_t}$, а веса неправильно классифицированных объектов умножаются на e^{α_t}.

Случай независимых переменных

Из формулы Байеса можем записать

$$g(x) = \mathsf{P}(y = 1 \mid x) = \frac{\mathsf{P}(dx, y = 1)}{\mathsf{P}(dx, y = 1) + \mathsf{P}(dx, y = -1)}$$

$$g(x) = \frac{1}{1 + \frac{1-p}{p} \cdot \frac{\mathsf{P}(dx|y=-1)}{\mathsf{P}(dx|y=1)}}.$$

Пусть условные распределения всех переменных X_j при условии обоих классов независимы, т.е.
$\mathsf{P}(dx \mid y) = \prod_{j=1}^{n} \mathsf{P}(dx_j \mid y).$

Сведение к логистической функции

Подставив это произведение в предыдущее выражение, после преобразований имеем

$$\frac{p}{1-p}\cdot\left(\frac{1}{g(x)}-1\right)=\prod_{j=1}^{n}\frac{p}{1-p}\cdot\left(\frac{1}{g_j(x_j)}-1\right),$$

где $g_j(x_j)=\mathsf{P}(y=1\mid x_j)=\frac{\mathsf{P}(dx_j,y=1)}{\mathsf{P}(dx_j)}$.
Логарифмируем последнее выражение и получаем

$$\sigma^{-1}(g(x))=(n-1)(\ln p-\ln(1-p))+\sum_{j=1}^{n}\sigma^{-1}(g_j(x_j)),$$

где $\sigma^{-1}(\cdot)$ — функция, обратная сигмоиду $\sigma(z)=\frac{1}{1+e^{-z}}$.

Обобщённый наивный байесовский классификатор

Заметим, что полученное выражение имеет вид логистической регрессии, а именно

$$g(x) = \sigma\left(u_0 + \sum_{j=1}^{n} u_j \sigma^{-1}(g_j(x_j))\right),$$

при $u_0 = (n-1)(\ln p - \ln(1-p))$, $u_j = 1$.
Обычно логистическую кривую получают, исходя из предположений о виде распределения, однако сейчас мы предположили независимость переменных, но не ограничивали вид распределений.

Использование модели

Данное выражение справедливо не только при независимых переменных, а в несколько более общем случае, поскольку из предыдущего соотношения независимость переменных не следует.
Ещё более расширить область применимости можно, если считать веса свободными параметрами.
Дальнейшее обобщение возможно, если допустить произвольные оценочные функции

$$g(x) = \sigma\left(u_0 + \sum_{j=1}^{n} u_j s(x_j)\right).$$

Оптимизация модели

Мы получили метод, который можно считать разновидностью метода логистической регрессии, а также разновидностью наивного байесовского классификатора.

Функции $s(x_j)$ можно оптимизировать напрямую (вместе с весами u_j), например, по критерию максимального правдоподбия.

В примере будут использованы кубические сплайны.

Оценивание условной вероятности

Условную вероятность $g(x) = \mathsf{P}(y = 1 \,|\, x)$ представим как находящиеся в точке x два объекта: класса 1 с весом $w_0 g(x)$ и класса -1 с весом $w_0(1 - g(x))$.
В результате выполнения бустинга вес первого объекта станет равным

$$w^{+1}(x) = w_0 g(x) \cdot A e^{-\beta(x)},$$

где константа A есть произведение всех нормировочных множителей.
Конечный вес второго объекта есть

$$w^{-1}(x) = w_0(1 - g(x)) \cdot A e^{\beta(x)}.$$

Если приравнять веса объектов, то получим

$$g(x) = \frac{1}{1 + e^{-2\beta(x)}}.$$

Бустинг на пороговых классификаторах

Бустинг на пороговых классификаторах («пнях») является разновидностью обобщённого наивного байесовского классификатора.
Действительно, каждая $\lambda_t(x)$ в композиции

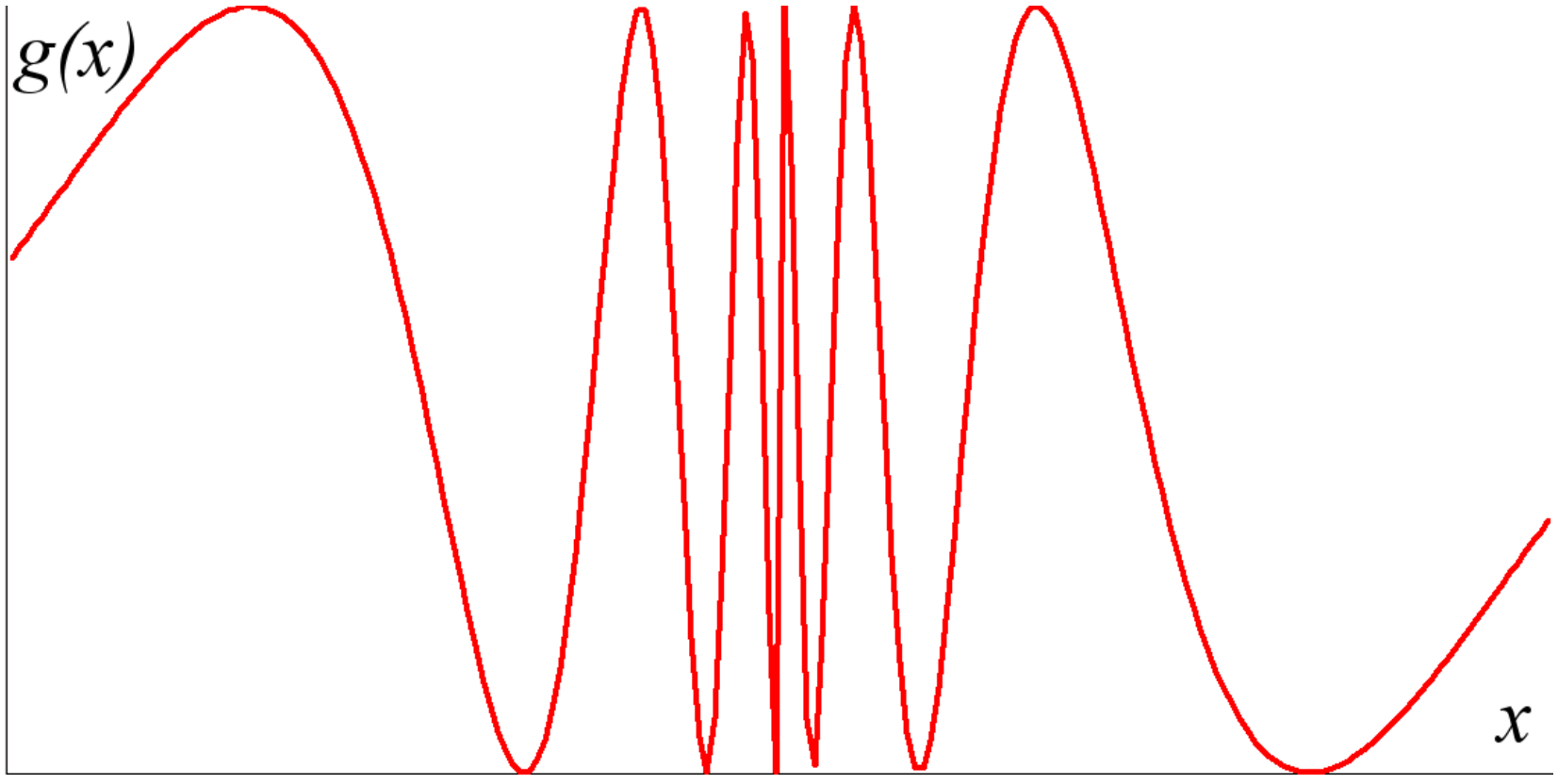
$$\beta(x) = \sum_{t=1}^{T} \alpha_t \lambda_t(x))$$

зависит только от одной переменной X_{i_t}, поэтому после группировки слагаемых выражение можно привести к виду

$$2\beta(x) = \sum_{i=1}^{n} u_i s(x)).$$

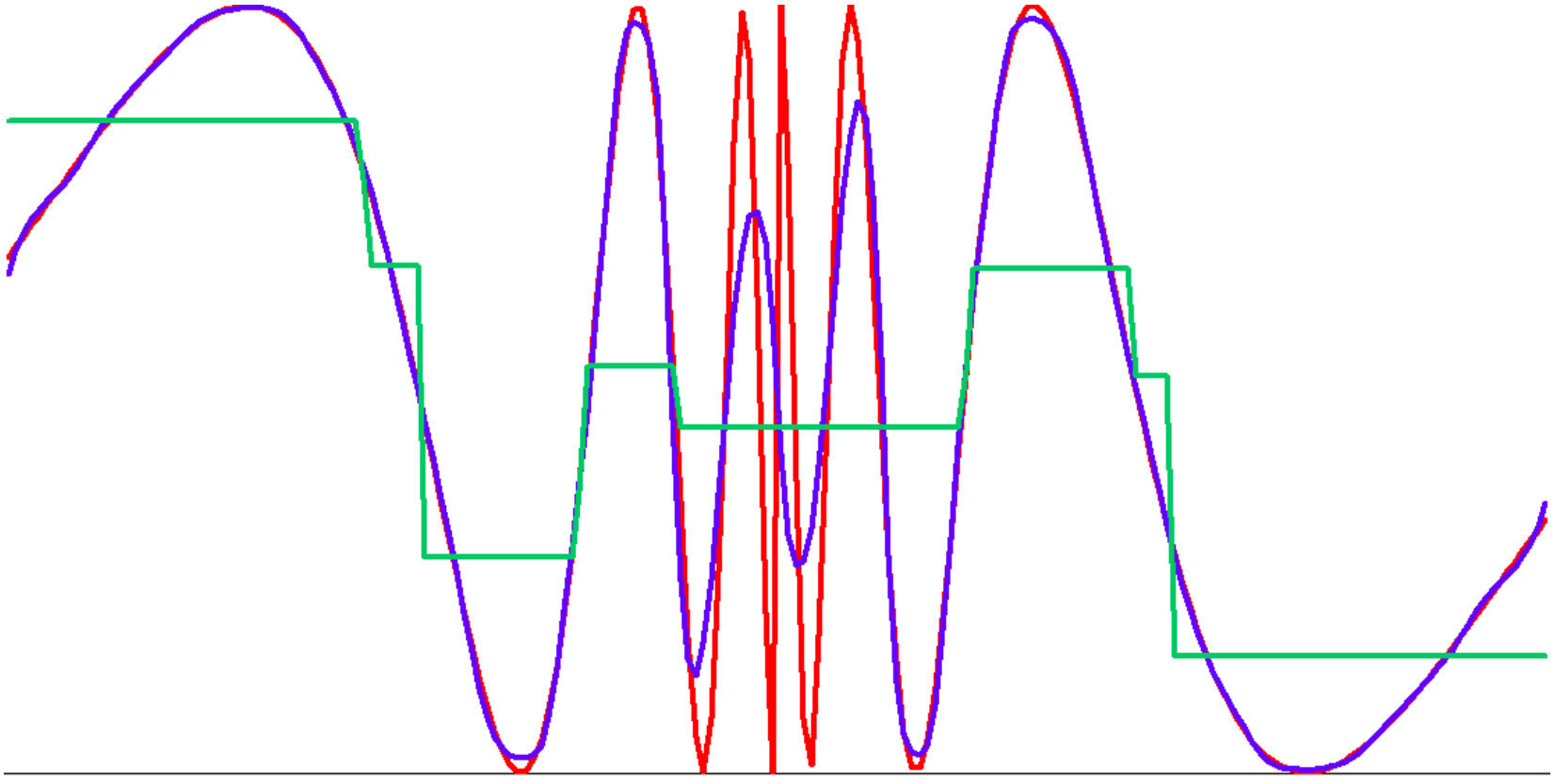
Подставив в выражение для $g(x)$, получим искомый вид.

Модельный пример



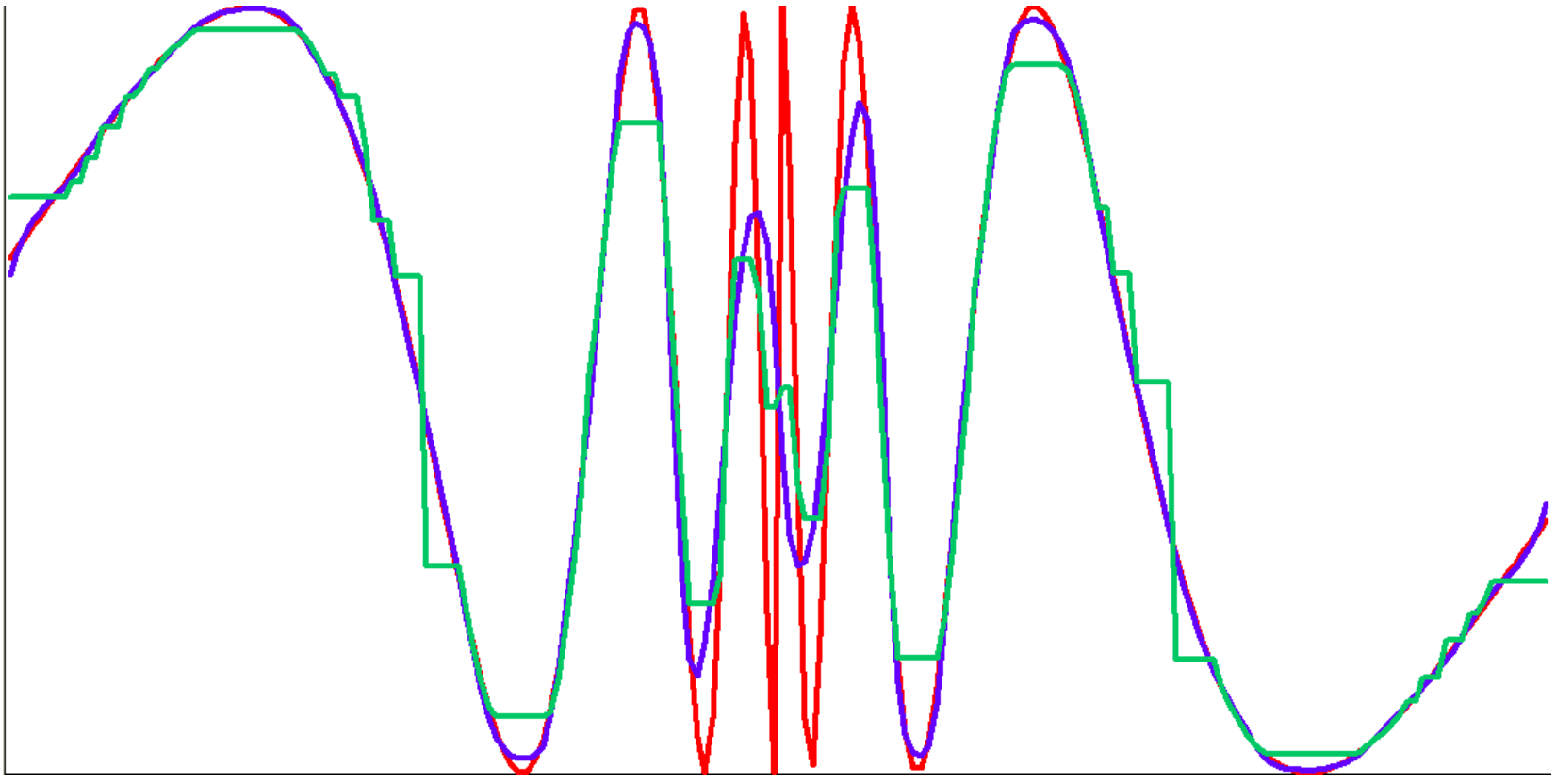
Функция условной вероятности.

Аппроксимация сплайном



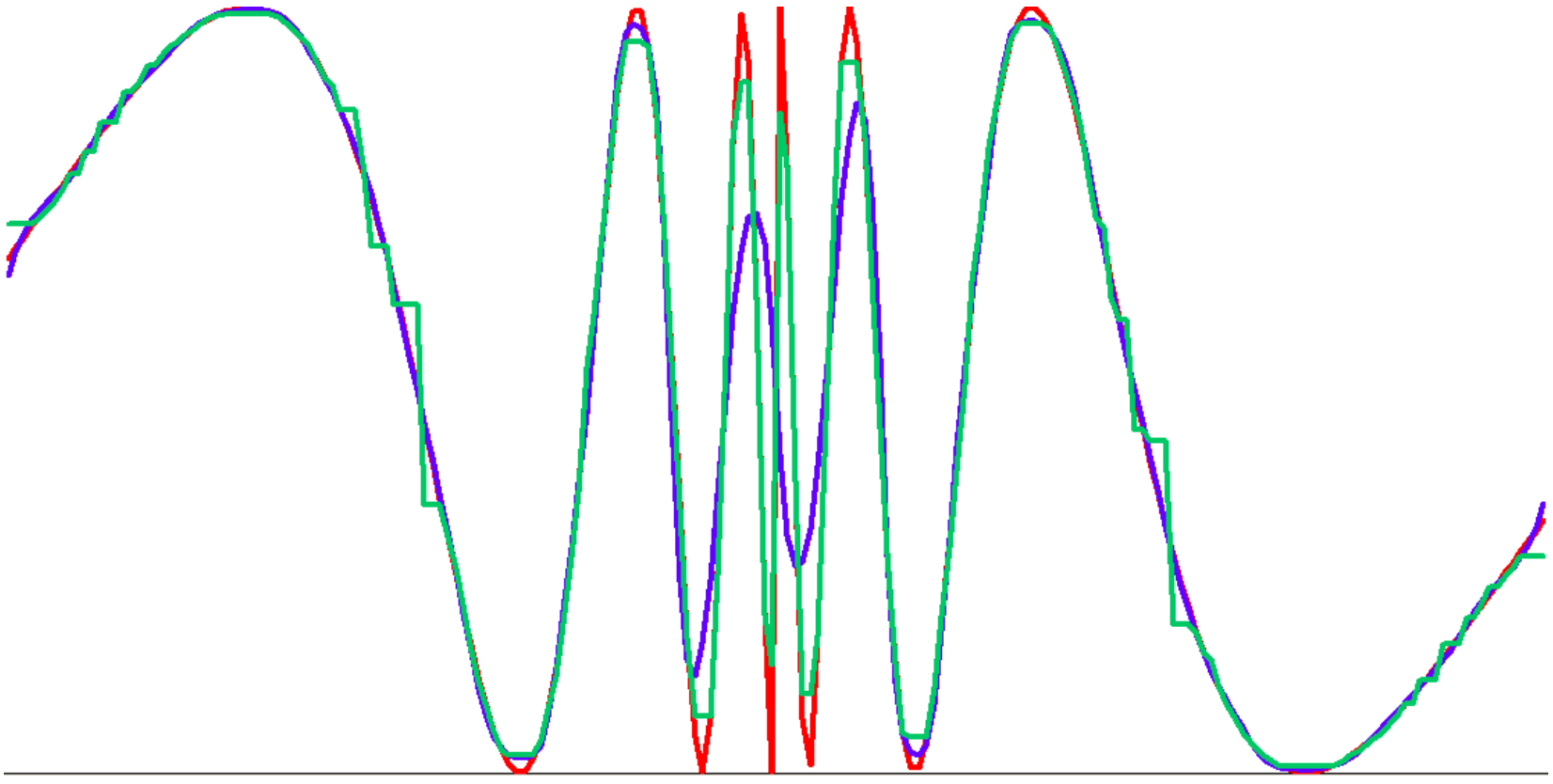
Кубический сплайн на 20 интервалов.
AdaBoost 10 итераций.

Boosting



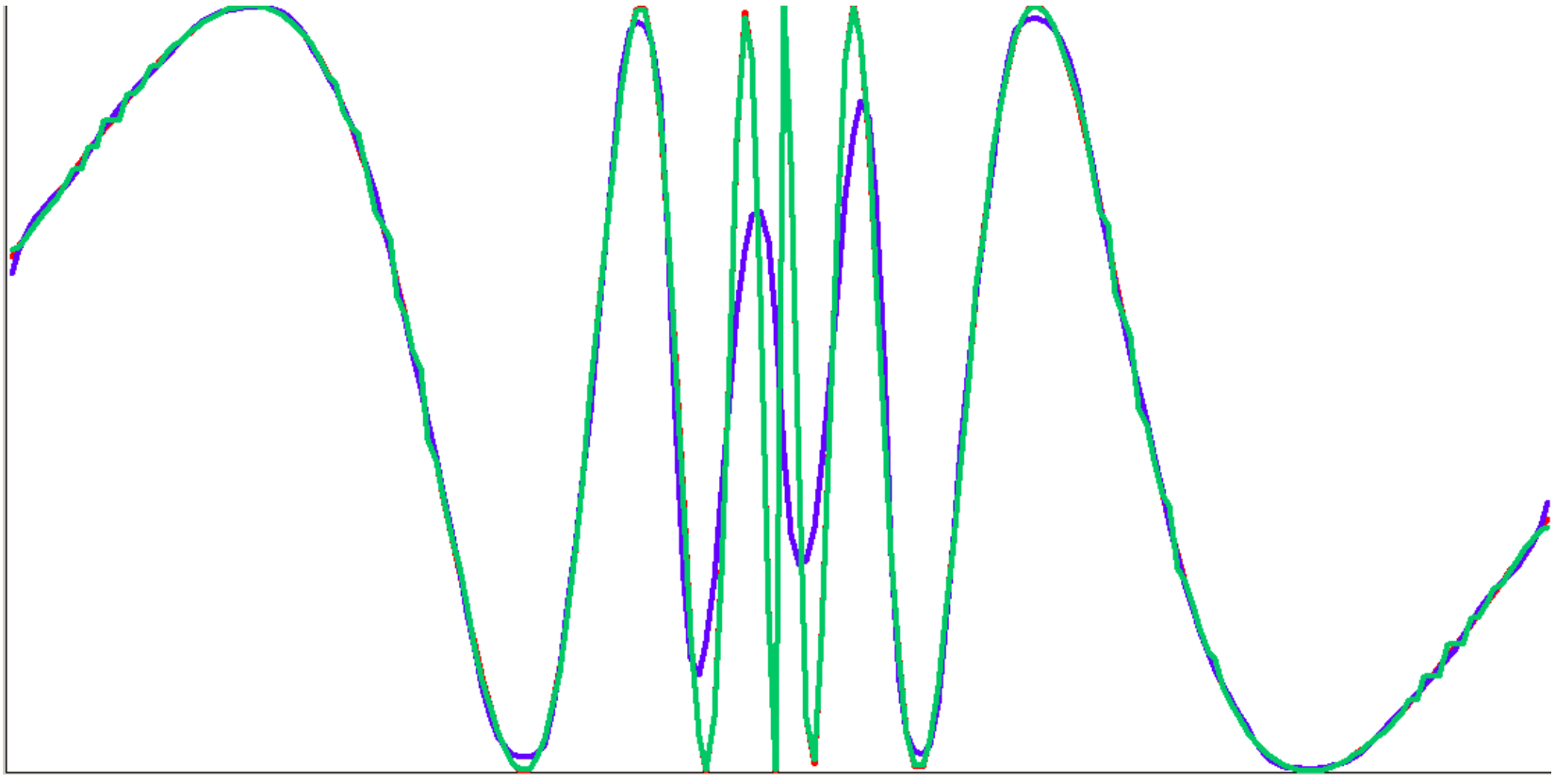
AdaBoost 100 итераций.

Boosting



AdaBoost 1000 итераций.

Boosting



AdaBoost 10000 итераций.

Выводы

Предложено в качестве тестовых единиц для исследования качества методов построения решающих функций использовать специальным образом подобранные классы распределений. Классы распределений подбираются таким образом, чтобы статистическое моделирование на них по возможности полно отражало особенности тестируемого метода обучения.

Это, в частности, означает, что класс распределений должен являться параметрическим семейством, один из параметров которого есть наименьшее достигаемое значение риска в заданном классе решающих правил.

Показана принципиальная возможность введения понятия оптимального метода построения решающих функций путём задания величины эталонного уровня риска.

Выводы

- Важнейшей причиной эффективности бустинга является использование эффекта независимости (переменных, подпространств, моделей).
- Бустинг на пороговых классификаторах является разновидностью непараметрической логистической регрессии, также его можно считать разновидностью (существенно обобщённого) наивного байесовского классификатора.
- Бустинг реализует «удачный» вариант непараметрической аппроксимации условной вероятности.

Интересное обсуждение бустинга:
David Mease and Abraham Wyner. Evidence Contrary to the Statistical View of Boosting // J. Mach. Learn. Res. 9 (June 2008), 131–156.

Список литературы I


Лбов Г. С., Старцева Н. Г. Логические решающие функции и вопросы статистической устойчивости решений. — Новосибирск: Издательство Института математики, 1999. — 212 с.


Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классиикации // Проблемы кибернетики. Вып. 33. М.: Наука, 1978. — С. 5–68.


Лбов Г. С. Выбор эффективной системы зависимых признаков // Выч. системы, вып. 19. Новосибирск. 1965. — с. 21–34.

Список литературы II

Неделько В. М. Некоторые вопросы оценивания качества методов построения решающих функций // Вестник Томского государственного университета. Управление, вычислительная техника и информатика, Томск: ТГУ, 2013. № 3 (24). — С. 123–132

Freund Y., Schapire R. E. Experiments with a new boosting algorithm // In Machine Learning: Proceedings of the Thirteenth International Conference, 1996. Pp. 148–156.

David Mease and Abraham Wyner Evidence Contrary to the Statistical View of Boosting // J. Mach. Learn. Res. 9 (June 2008), 131-156.