К вопросу об эффективности бустинга в задаче классификации

Неделько В. М.

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

«Математические методы распознавания образов»
(ММРО-17), г. Светлогорск, 19–25 сентября 2015 г.

Объяснения эффективности бустинга

Цель работы: объяснить механизм высокой эффективности метода бустинга на реальных задачах.

Известный подход:

- максимизация отступа.

Предлагаемое объяснение:

- эксплуатация эффекта независимости.

Предположения о свойствах реальных задач

Для решения задач построения решающих функций используются предположения о свойствах данных:

- гипотеза компактности (простоты модели);
- гипотеза независимости.

Методы, основанные на композициях, ипользуют обе гипотезы.

Проклятие размерности

Для случая независимых переменных «проклятие» размерности превращается в преимущество:

- чем больше зависимых переменных — тем хуже,
- чем больше независимых переменных — тем лучше.

С увеличением числа независимых переменных качество решения только растёт.

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.

Обобщение 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

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

Методы в рамках модели

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

Логистическая функция от линейной комбинации базовых классификаторов — признак использования (ослабленного условия) независимости.

Примеры подобных методов.

- Бустинг на пороговых классификаторах.
- Логистическая регрессия.
- Машина опорных векторов.

Используемый подход

Эффективность метода построения решающих функций зависит от

- аппроксимирующей способности,
- обобщающей способности (статистической устойчивости).

Для исследования аппроксимирующей способности в-отдельности будем на «вход» метода подавать распределения.

Также будут приведены некоторые свойства метода AdaBoost, которые в некотором виде 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

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

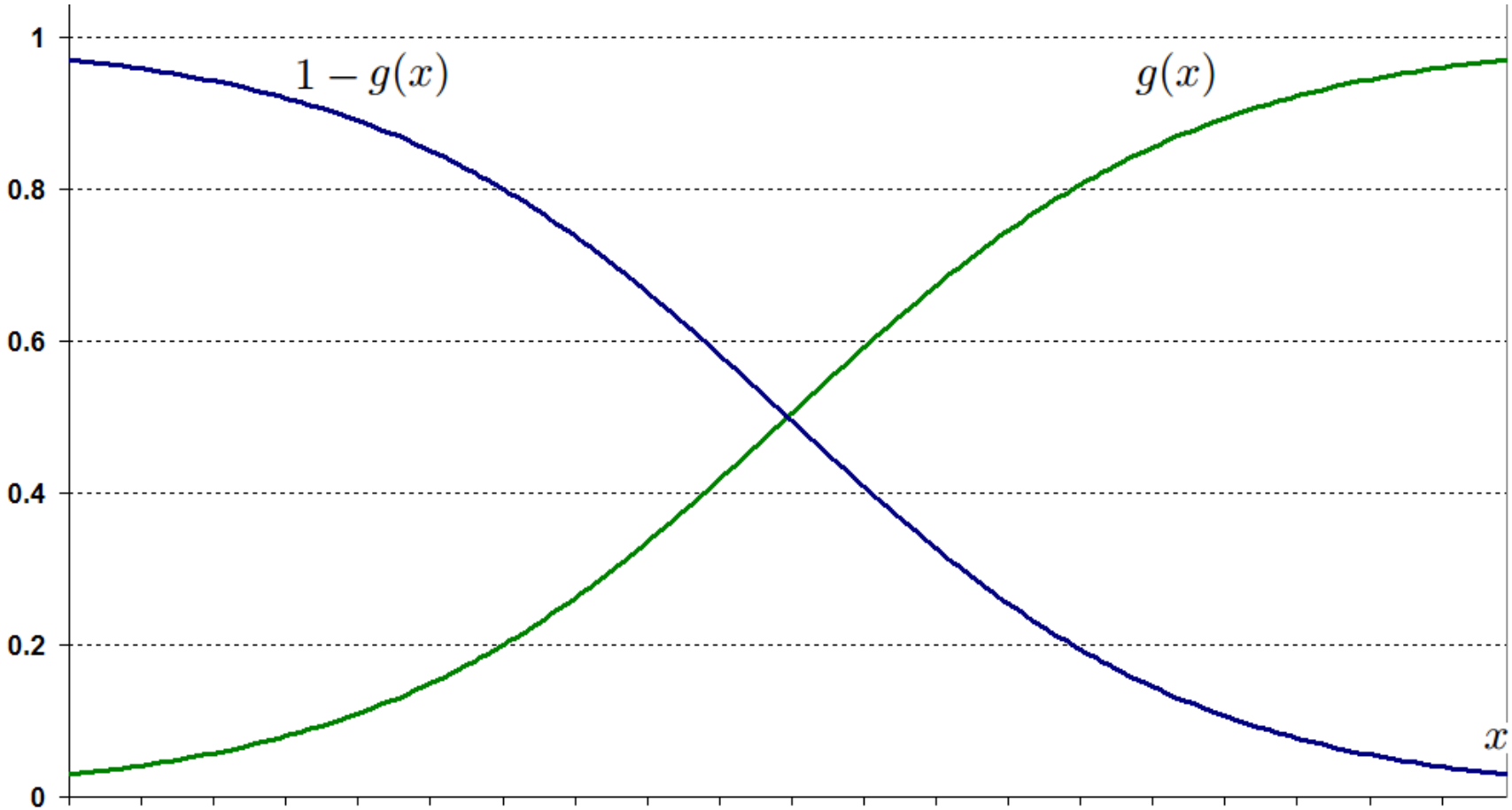
Веса правильно классифицированных объектов умножаются на $e^{-\alpha_t}$, а веса неправильно классифицированных объектов умножаются на e^{α_t}.

Распределение в роли входных данных

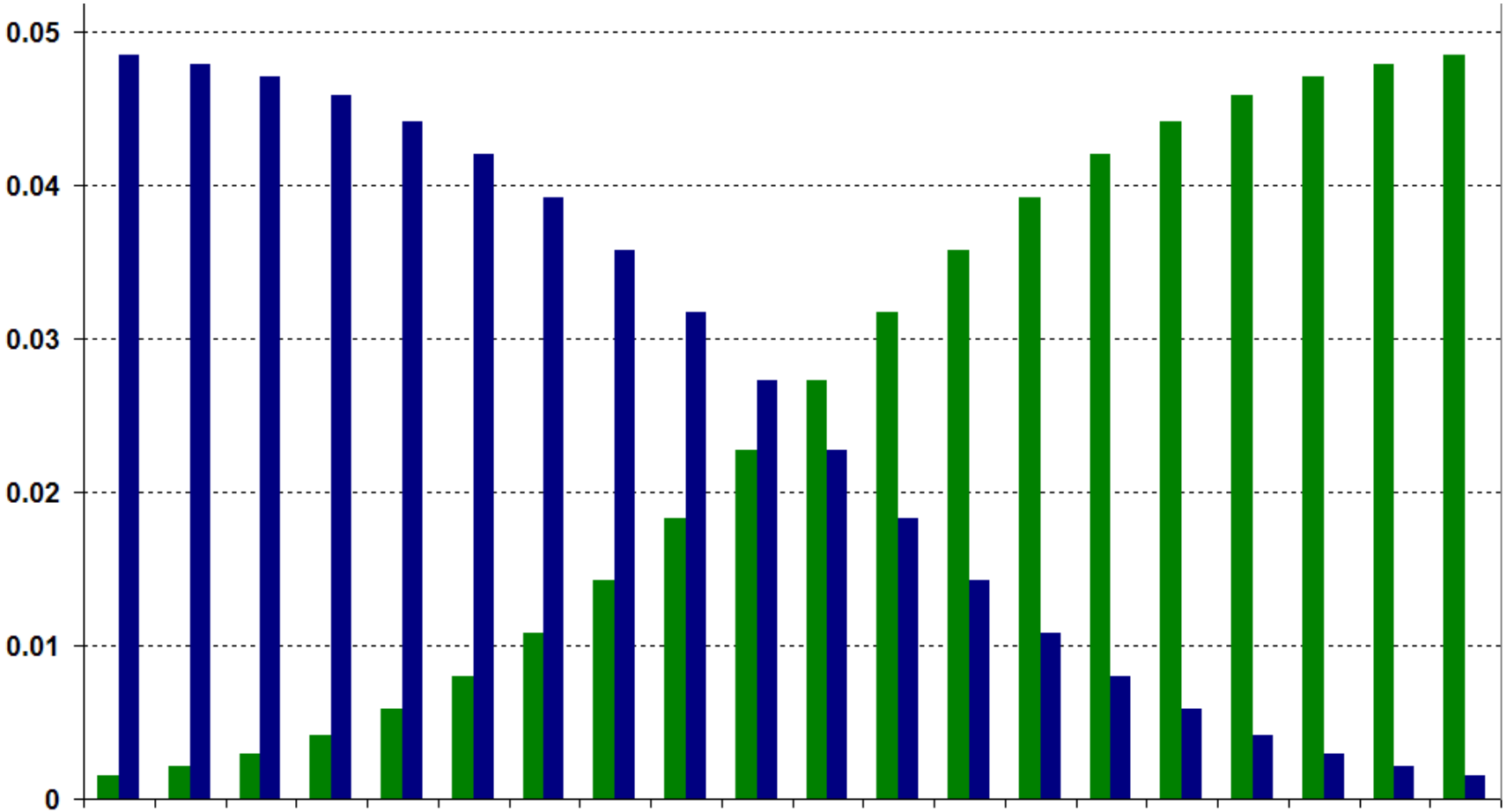
Многие методы классификации (например, бустинг на деревьях) могут вместо выборки использовать произвольные вероятностные меры. Но для простоты ограничимся дискретными распределениями, поскольку дискретное распределение идентично выборке с весами.

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 $w_0(1 - g(x))$.
Множитель w_0 задаётся в соответствии с безусловным распределением $\mathsf{P}(x)$.

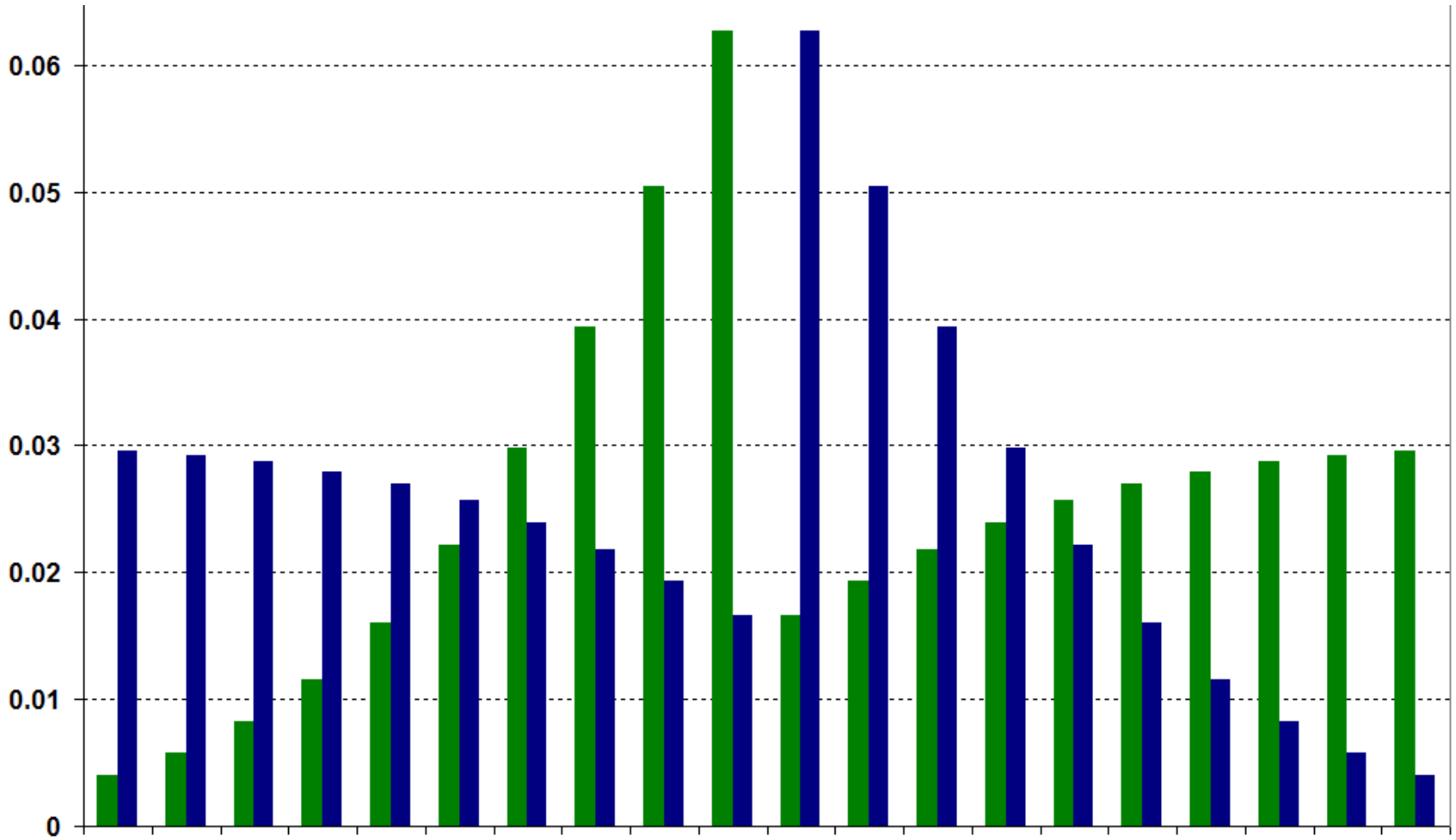
Пример распределения



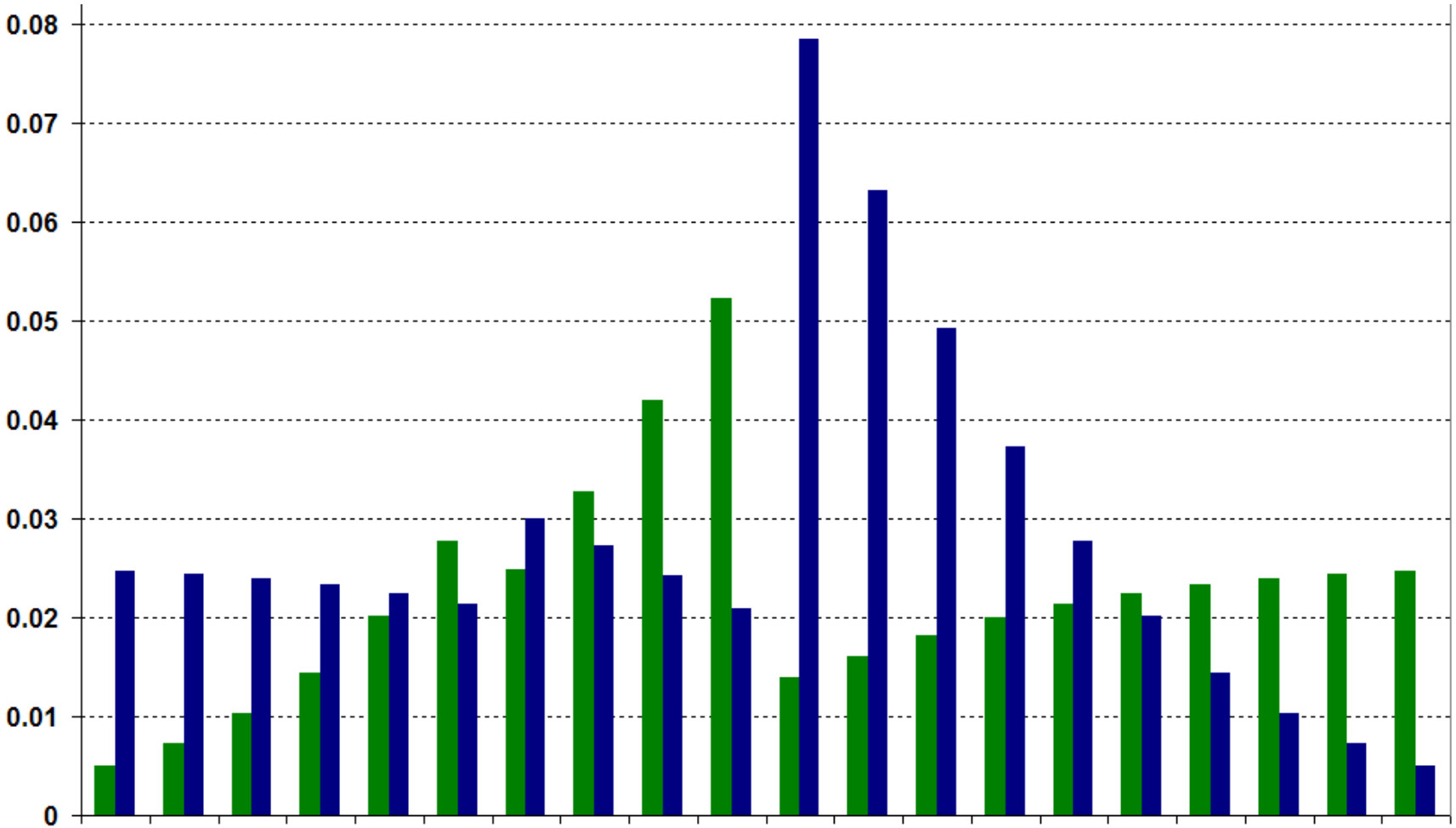
Дискретное распределение в роли выборки



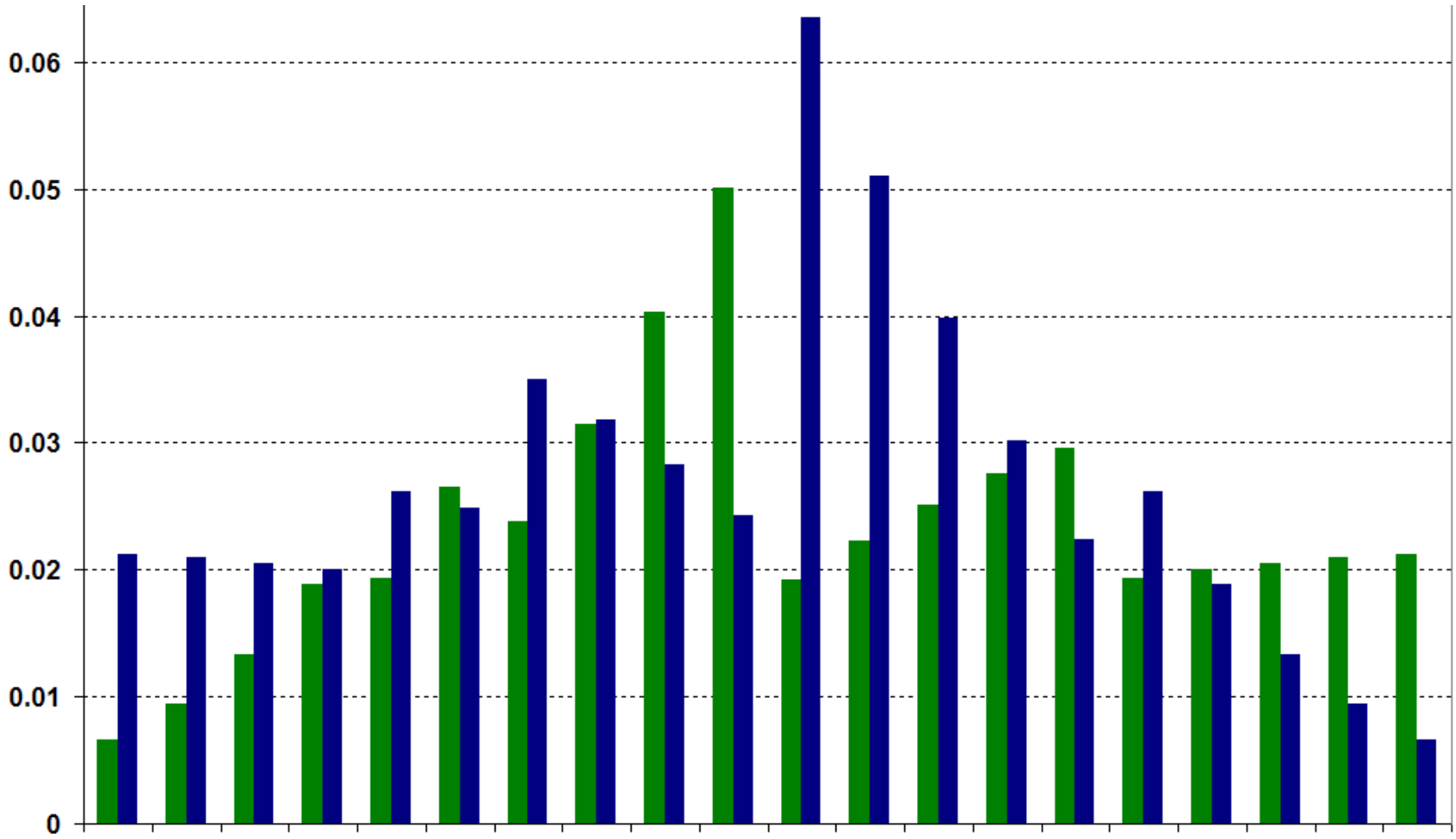
Число пороговых классификаторов: 1



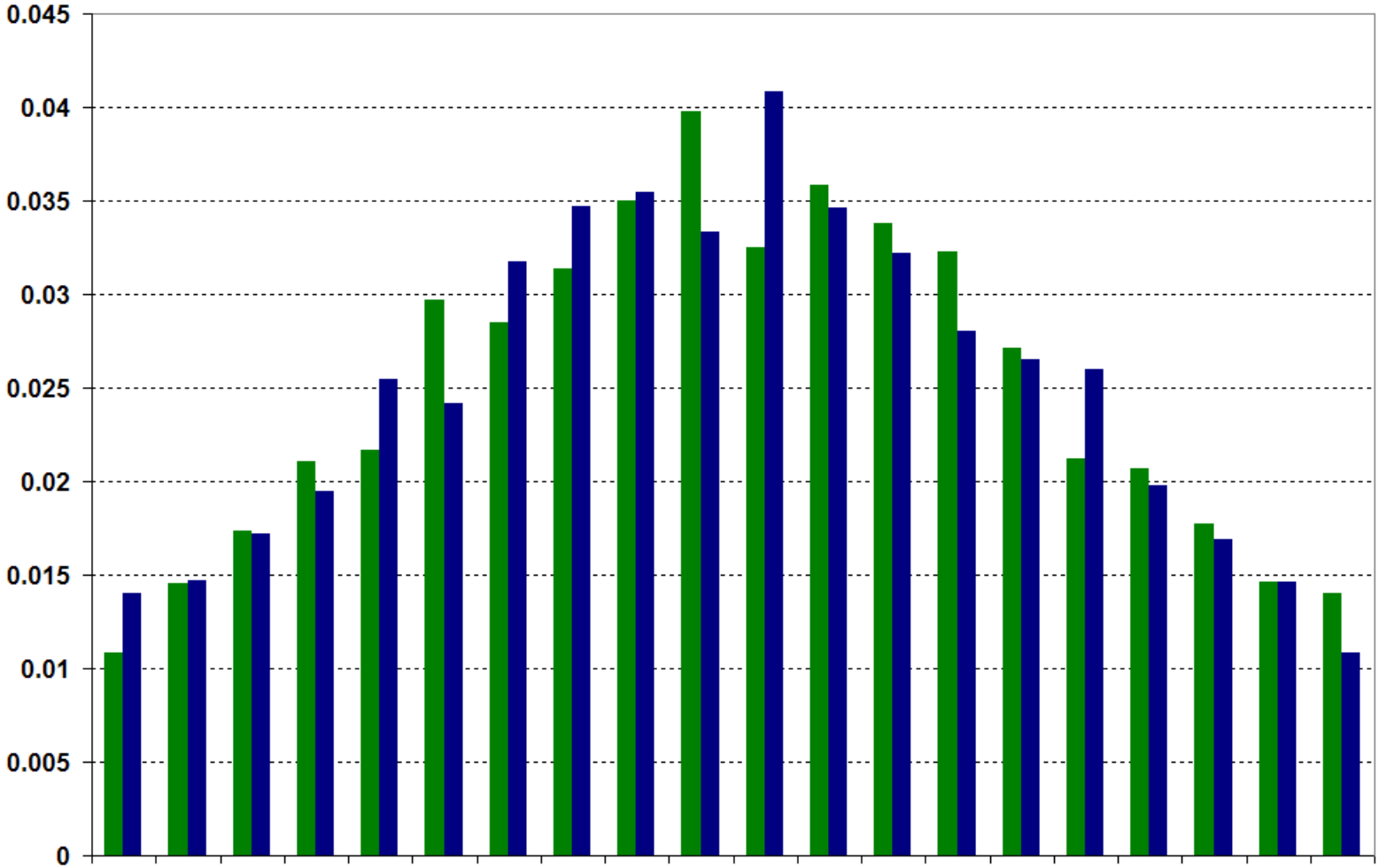
Число пороговых классификаторов: 2



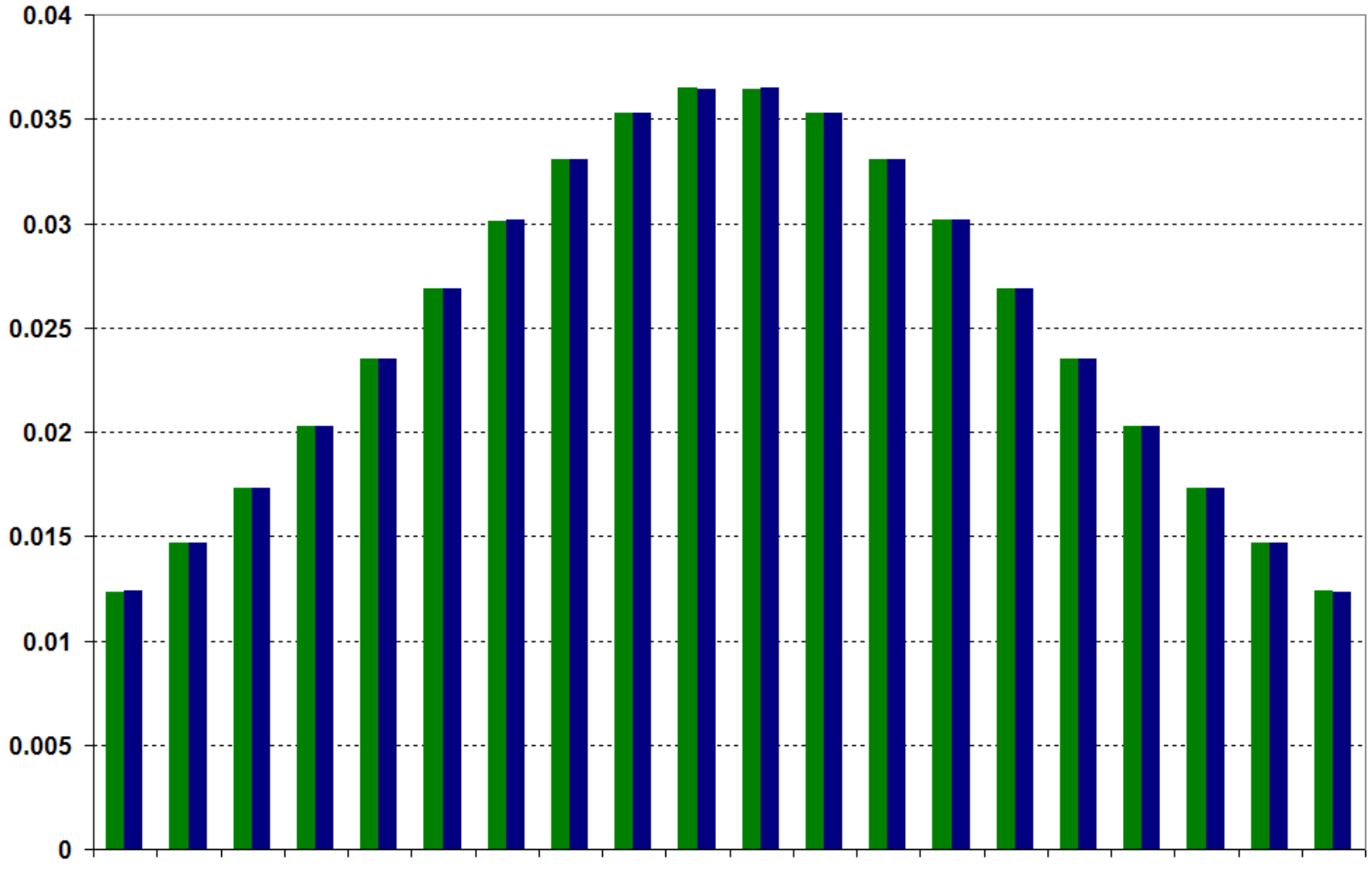
Число пороговых классификаторов: 5



Число пороговых классификаторов: 50



Число пороговых классификаторов: 500



Случай сравнявшихся весов

Выберем некоторую точку x.
В результате выполнения бустинга вес объекта первого класса в этой точке станет равным

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

где $\beta(x) = \sum_{t=1}^{T} \alpha_t \lambda_t(x))$, а константа A есть произведение всех нормировочных множителей.
Конечный вес объекта второго класса есть

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

Оценка условной вероятности

Если приравнять веса объектов, то получим

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

Получили вид, похожий на логистическую регрессию.

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

$$2\beta(x) = \sum_{j=1}^{n} u_j s_j(x_j)).$$

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

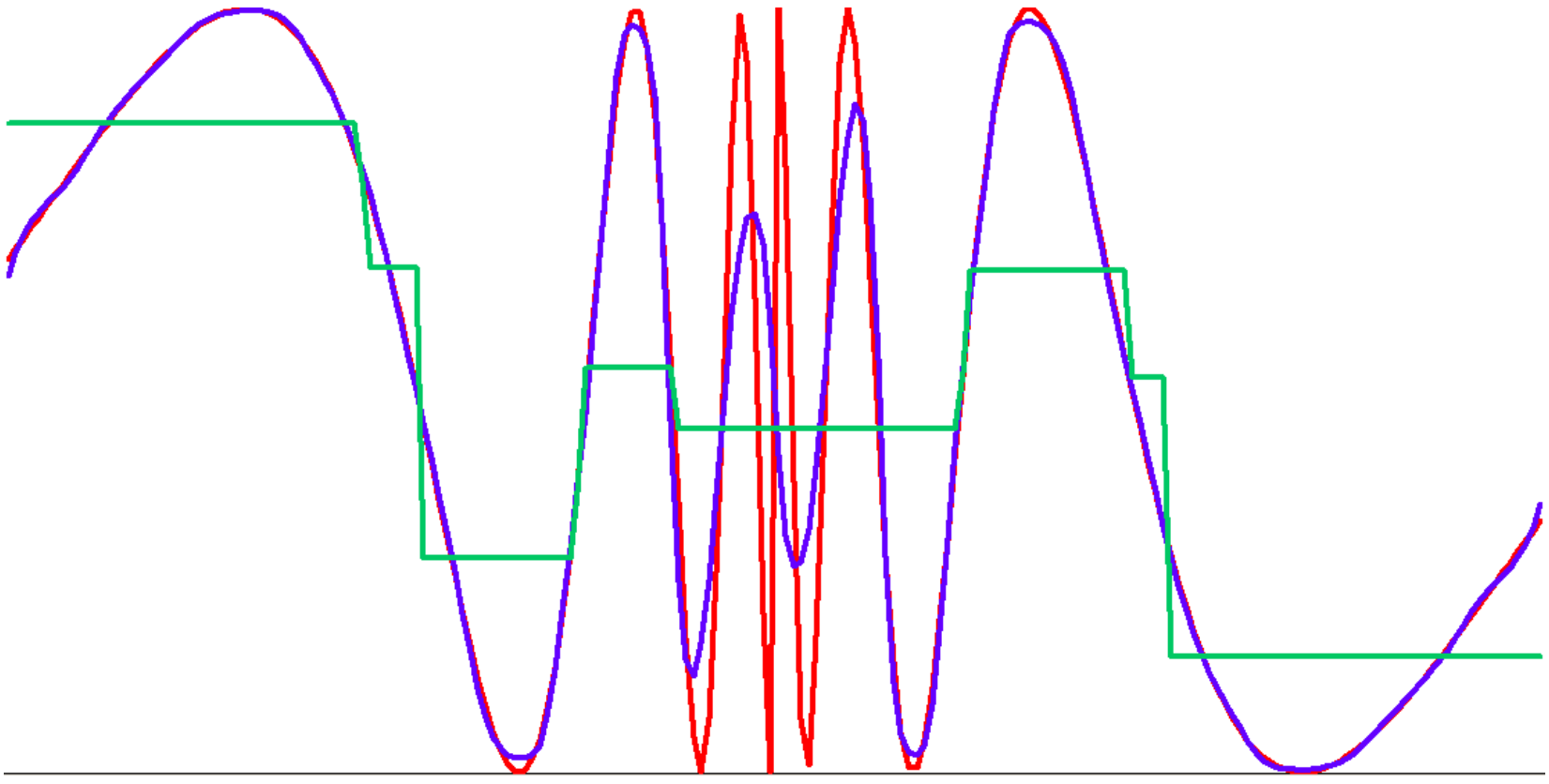
Бустинг на деревьях и ряд Бахадура

Модель можно естественным образом обобщить по аналогии с рядом Бахадура, включив возможность учитывать зависимости между переменными, последовательно добавляя парные зависимости, зависимости в тройках и т.д.

$$g(x) = \sigma\left(u_0 + \sum_{j=1}^{n} u_j s_j(x_j) + \sum_{j,k} u_{jk} s_{jk}(x_j, x_k) + \right.$$

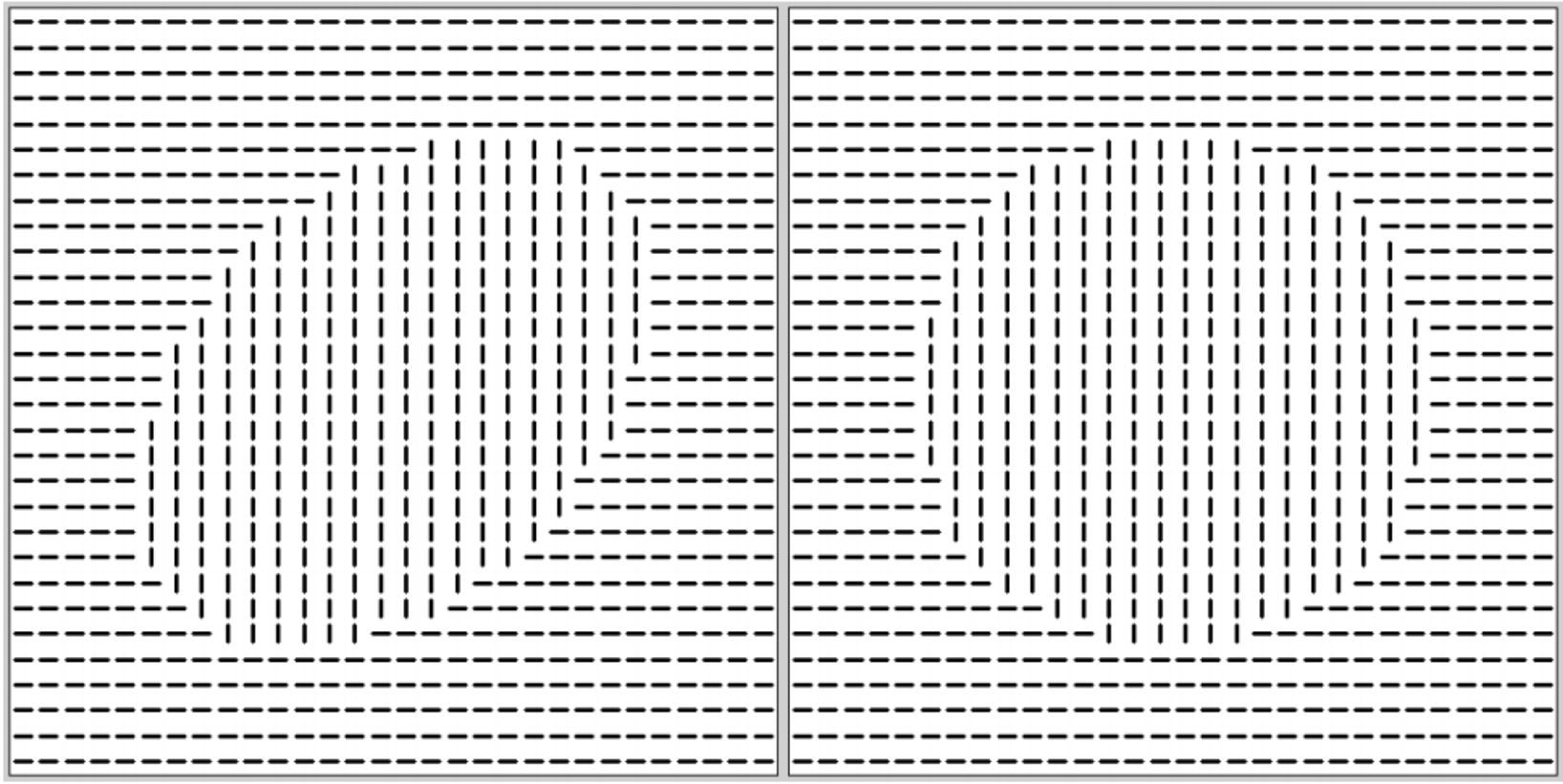
$$\left. + \sum_{j,k,l} u_{jkl} s_{jkl}(x_j, x_k, x_l) + \ldots\right).$$

Аппроксимация функции условной вероятности



Кубический сплайн на 20 интервалов.
AdaBoost 10 итераций.

Пример на двух переменных

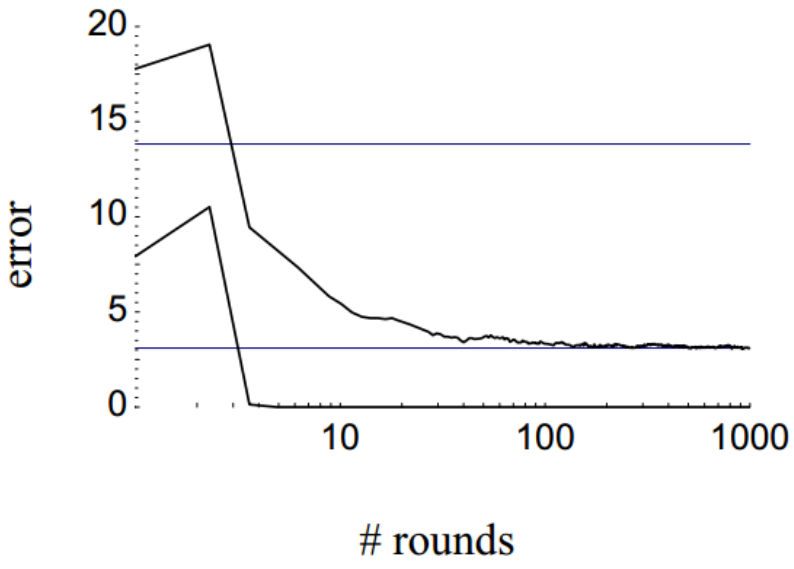


«Регуляризация» выборки

Заметим, что в приведённых примерах бустинг обладает свойством самостоятельного останова и даёт решение в виде сходящегося ряда из классификаторов.

При использовании выборки имеет смысл проводить «регуляризацию», т.е. добавлять к каждому объекту объект противоположного класса с тем же x и маленьким весом.

Классический пример



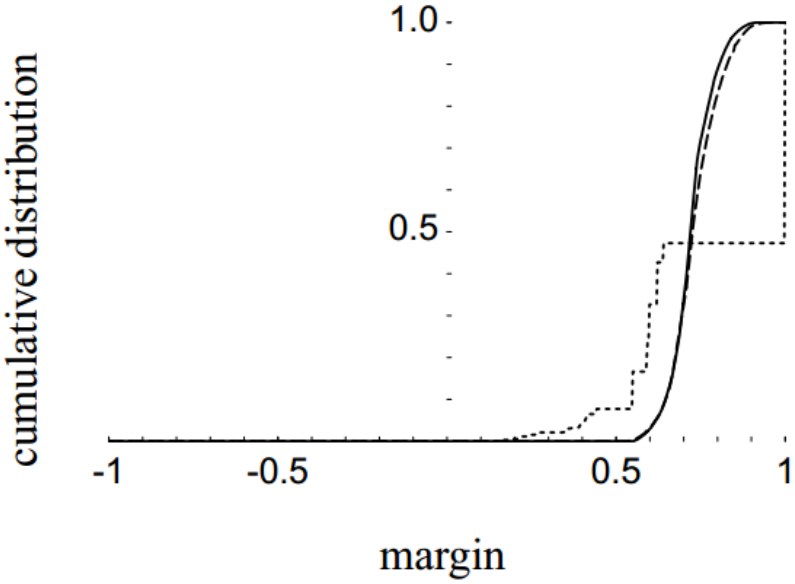


Иллюстрация приводится во многих известных работах.

Оценка риска на основе отступа

Для случая конечного множества Λ базовых классификаторов имеет место оценка риска, в соответствии с которой с вероятностью, не меньшей $1-\delta$, для $\theta > 0$ выполняется неравенство:

$$R(\lambda, c) = \mathsf{P}_c(y\beta(x) \leq 0) \leq \widetilde{\mathsf{P}}(y\beta(x) \leq \theta\varkappa) +$$

$$+O\left(\frac{1}{\sqrt{N}}\left(\frac{\ln N \ln|\Lambda|}{\theta^2} - \ln\delta\right)^{1/2}\right),$$

где $\varkappa = \sum_{t=1}^{T} \alpha_t$, а $\widetilde{\mathsf{P}}(\cdot)$ – частота события на выборке.

Оценка риска на основе отступа

Огрубив, получаем

$$R(\lambda, c) \leq \widetilde{F}(\theta) + O\left(\frac{1}{\theta} \cdot \sqrt{\frac{\ln |\Lambda|}{N}}\right),$$

где $\widetilde{F}(\theta) = \widetilde{\mathsf{P}}(y\beta(x) < \theta\varkappa)$ – эмпирическая (выборочная) функция распредления для величины $\frac{y\beta(x)}{\varkappa}$, которую принято называть отступом.

Механизм максимизации отступа

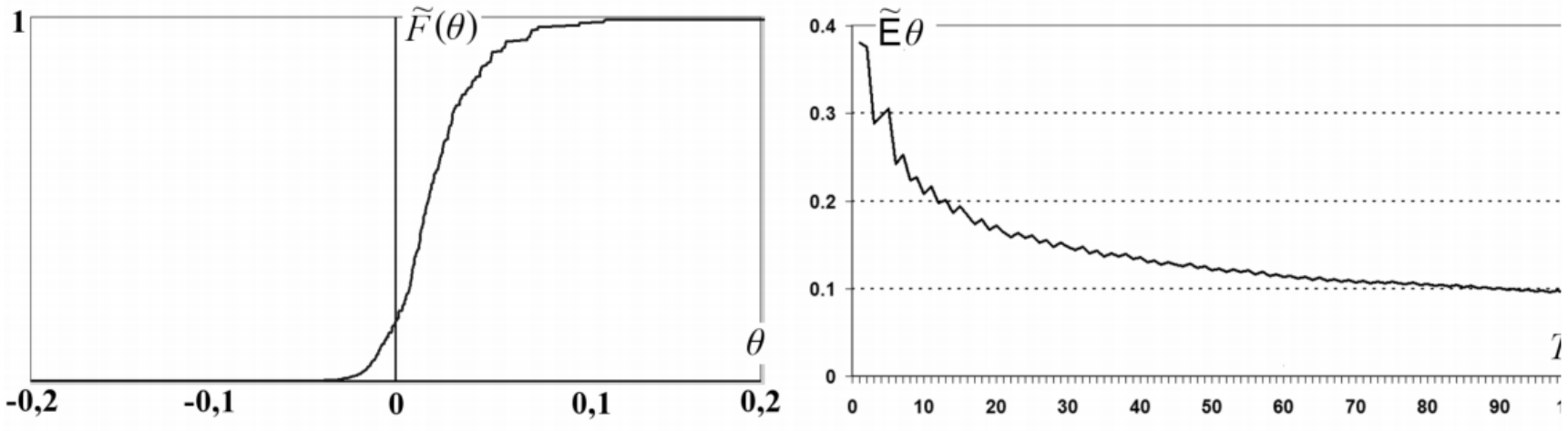
Почему бустинг максимизирует отступ.

Бустинг максимизирует отступ:

- ненормированный — по построению,
- нормированный — потому что эффективно решает задачу.

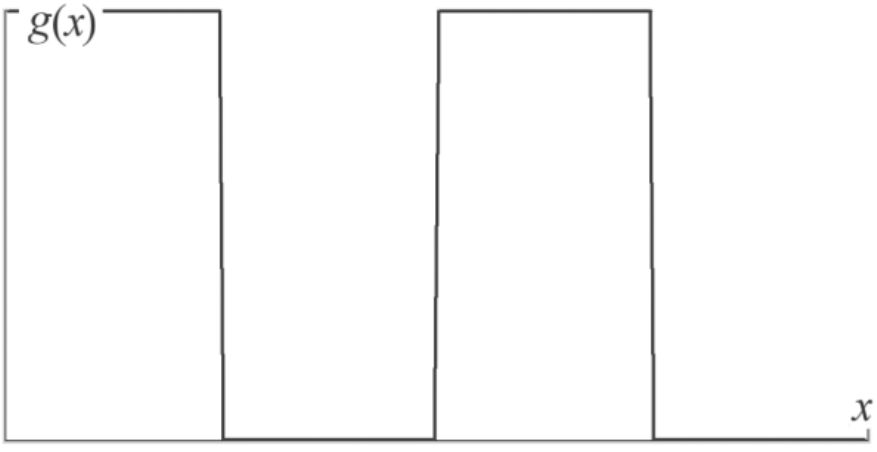
Нормированый отступ с ростом композиции не обязательно растёт.

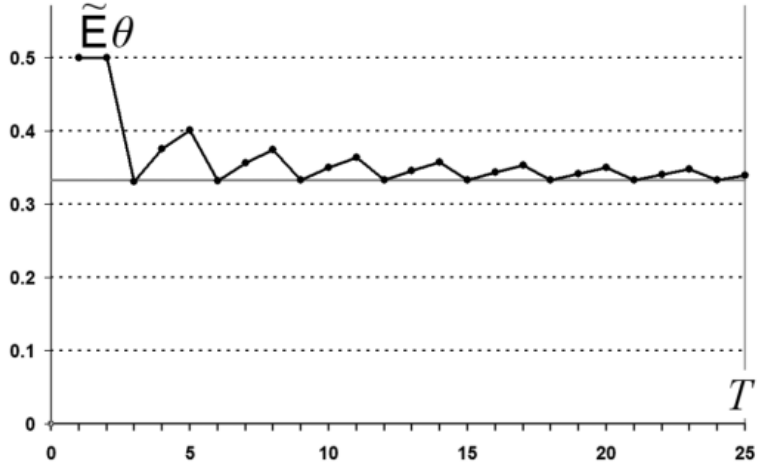
Пример поведения отступа



Эмпирическое распределение отступов после 80000 итераций и зависимость среднего отступа от числа итераций.

Периодическая ступенчатая функция





Функция условной вероятности:

$$g^K(x) = (1 - ([Kx] \mod 2))(1 - \rho) + 0{,}5\rho,$$

где K – параметр, задающий число областей постоянства, а ρ – параметр, определяющий байесовский уровень ошибки.

Интерпретация отступа

Утверждение. Для предельного значения отступа, получаемого методом AdaBoost с пороговыми решающими функциями на приведённой ступенчатой модели для чётных K при увеличении числа итераций, справедливо следующее выражение

$$\lim_{T\to\infty} \widetilde{\mathsf{E}}\theta = \frac{1}{K-1}\cdot(1-\rho)$$

Таким образом, ненормированный отступ является обобщением эмпирического риска, а нормировочный коэффициент отражает сложность решения.

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.
- Нормированный отступ, вообще говоря, уменьшается при росте эффективной сложности композиции (нормировочный коэффициент отражает сложность).
- Объяснять эффективность бустинга максимизацией отступа не более оправдано, чем объяснять максимизацию (нормированного) отступа эффективностью бустинга.

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.