Содержание

Глава 15

Случайные леса

15.1. Введение

В разделе 8.7 мы рассмотрели *баггинг*, или *бутстрэп-агрегацию* — метод уменьшения дисперсии оценочной функции предсказания. Оказалось, что баггинг особенно хорошо подходит для процедур с высокой дисперсией и низким смещением, таких как деревья. При решении задач регрессии мы многократно аппроксимируем одно и то же дерево регрессии по обучающим бутстрэп-версиям и усредняем результат. При решении задач классификации за прогнозируемый класс голосует *комитет* деревьев.

В главе 10 бустинг был первоначально предложен также в качестве метода комитетов, хотя в отличие от баггинга комитет *слабых учеников* со временем эволюционирует, и голоса членов имеют разные веса. Оказалось, что бустинг превосходит баггинг при решении большинства задач и стал предпочтительным выбором.

Случайные леса (random forests) (Breiman, 2001) — существенная модификация баггинга, которая создает большую коллекцию *декоррелированных* деревьев, а затем усредняет их. При решении многих задач точность случайных лесов сопоставима с точностью бустинга, но случайные леса проще обучать и настраивать. Как следствие, случайные леса стали популярными и реализованы в различных пакетах.

15.2. Определение случайных лесов

Основная идея баггинга (см. раздел 8.7) состоит в том, чтобы усреднить большое количество зашумленных, но приблизительно несмещенных моделей и тем самым уменьшить дисперсию. Деревья являются идеальными кандидатами для баггинга, так как могут отражать сложные структуры взаимодействия, скрытые в данных и, если их деревья достаточно глубокие, имеют относительно небольшое смещение. Поскольку общеизвестно, что деревья имеют сильный шум, их усреднение приносит большую пользу. Более того, поскольку все деревья, генерируемые при баггинге, имеют одинаковое распределение, математическое ожидание среднего значения B таких деревьев совпадает с математическим ожиданием каждого из них. Это означает, что смещение деревьев при баггинге такое же, как и у отдельных деревьев (при бутстрэпе), и единственная надежда на улучшение заключается в уменьшении дисперсии. Этим баггинг отличается от бустинга, в котором деревья выращиваются адаптивным способом, чтобы устранить смещение, и, следовательно, не являются одинаково распределенными.

Алгоритм 15.1. Случайный лес для регрессии или классификации

1. Для $b = 1$ до B:
   а). Извлечем бутстрэп-выборку Z^* размера N из обучающих данных.
   б). Вырастим дерево случайного леса T_b по бутстрэп-данным, рекурсивно повторяя следующие шаги для каждого конечного узла дерева, пока не будет достигнут минимальный размер узла n_{min}.
      i. Случайным образом выберем m переменных из p переменных.
      ii. Выберем лучшую переменную или точку разделения среди m переменных.
      iii. Разделяем узел на два дочерних узла.
2. Возвращаем ансамбль деревьев $\{T_b\}_1^B$.

   Регрессия: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.
3. *Классификация*: пусть $\hat{C}_b$ — прогнозируемый класс b-го дерева случайного леса. Тогда $\hat{C}_{rf}^B = \textit{большинство голосов}\left\{\hat{C}_b(x)\right\}_1^B$.

Среднее значение B одинаково распределенных случайных величин, каждая из которых имеет дисперсию σ^2, имеет дисперсию $\frac{1}{B}\sigma^2$. Если переменные одинаково распределены (но не обязательно независимые) с положительной попарной корреляцией ρ, то дисперсия среднего значения (см. упражнение 15.1) равна

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{15.1}$$

При увеличении B второе слагаемое исчезает, но первое остается, и, следовательно, величина корреляции пар деревьев при баггинге ограничивает преимущества усреднения. Идея метода случайных лесов (см. алгоритм 15.1) состоит в том, чтобы снизить уменьшение дисперсии баггинга за счет уменьшения корреляции между деревьями, не слишком сильно увеличивая дисперсию. Это достигается в процессе построения деревьев путем случайного выбора входных переменных.

В частности, при построении дерева на бутстрэп-множестве данных необходимо выполнить следующую операцию.

> *Перед каждым разделением выберите $m \le p$ случайных входных переменных в качестве кандидатов на расщепление.*

Обычно значения m равны $\sqrt{p}$ или единице.

После построения таких деревьев $\left\{T(x;\Theta_b)\right\}_1^B$ предиктор случайного леса (для регрессии) примет вид

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b). \tag{15.2}$$

Как и в разделе 10.9, Θ_b характеризует b-е дерево случайного леса в терминах расщепляющих переменных, точек отсечения в каждом узле и значений терминальных узлов. Интуитивно понятно, что уменьшение m уменьшит корреляцию между любой парой деревьев в ансамбле и, следовательно, (15.1) уменьшит дисперсию среднего значения.

Не все оценки можно улучшить путем возмущения данных, как в данном случае. Оказывается, сильно нелинейные средства оценивания, такие как деревья, приносят наибольшую пользу. Для бутстрэп-деревьев величина ρ обычно невелика (около 0,05 или ниже; см. рис. 15.9), тогда как σ^2 не намного больше, чем дисперсия для исходного дерева. С другой стороны, баггинг не меняет *линейные* оценки, такие как выборочное среднее (а значит, и дисперсию); попарная корреляция между средними значениями после бутстрэпа составляет около 50% (см. упражнение 15.4).

Случайные леса популярны. Лео Брайман (Leo Breiman)[1], сотрудник Адель Катлер (Adele Cutler), поддерживает сайт, посвященный методу случайного леса[2]. Его программное обеспечение находится в свободном доступе, и к 2002 г. было зарегистрировано более 3000 загрузок. Кроме того, существует пакет **`randomForest`** в языке R, поддерживаемый Энди Лиу (Andy Liaw), доступный на сайте CRAN.

Авторы делают громкие заявления об успехе случайных лесов: “наиболее точные”, “наиболее интерпретируемые” и т.п. По нашему опыту, случайные леса работают замечательно и требуют совсем небольшой настройки. На тестовых данных о спаме уровень ошибочной классификации случайного леса снижается до 4,88%, что вполне сопоставимо со всеми другими методами и не намного хуже, чем градиентный бустинг, уровень ошибок которого составил 4,5%. Уровень ошибок баггинга составил 5,4%, что значительно хуже, чем у любого другого метода (с использованием теста Мак-Немара, описанного в упражнении 10.6), поэтому в этом примере дополнительная рандомизация оказалась полезной.

На рис. 15.1 показана прогрессия ошибок тестирования на 2500 деревьях для трех методов. В данном случае есть некоторые свидетельства того, что градиентный бустинг начал переобучаться, хотя 10-блочная перекрестная проверка выбрала все 2500 деревьев.

На рис. 15.2 показаны результаты моделирования[3], сравнивающего случайные леса с градиентным бустингом при решении *задачи о вложенных сферах* (см. уравнение (10.2) в главе 10). Здесь бустинг намного превосходит случайные леса. Обратите

---

[1] Лео Брейман умер в июле 2005 года.

[2] `http://www.math.usu.edu/~adele/forests/`

[3] Детали: случайные леса были обучены с использованием пакета **`randomForest 4.5-11`** языка R и 500 деревьев. Модели градиентного бустинга были обучены с использованием пакета **`gbm 1.5`** языка R с параметром сжатия, равным 0,05, и 2000 деревьев.

внимание на то, что чем меньше m, тем лучше, хотя одна из причин может заключаться в том, что истинная граница решения является аддитивной.

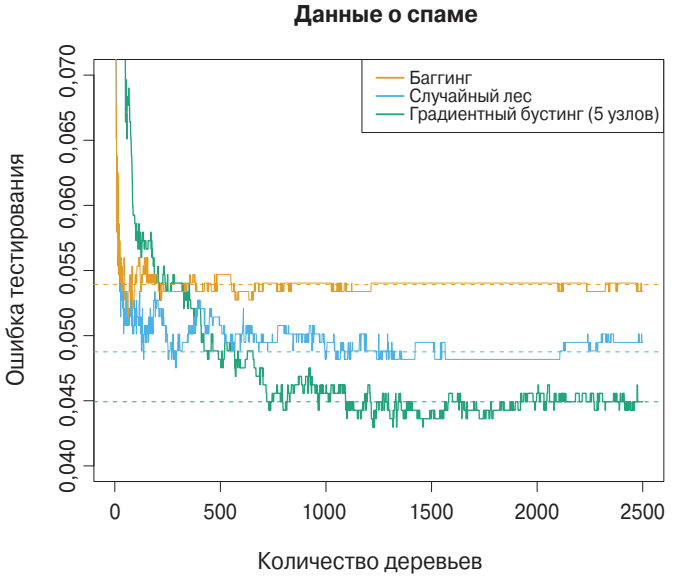

Рис. 15.1. Результаты применения баггинга, случайного леса и градиентного бустинга к данным о спаме. Для бустинга были использованы пятиузловые деревья, а количество деревьев было выбрано путем 10-блочной перекрестной проверки (2500 деревьев). Каждый шаг на рисунке соответствует изменению в одной неправильной классификации (в тестовом множестве из 1536 элементов)

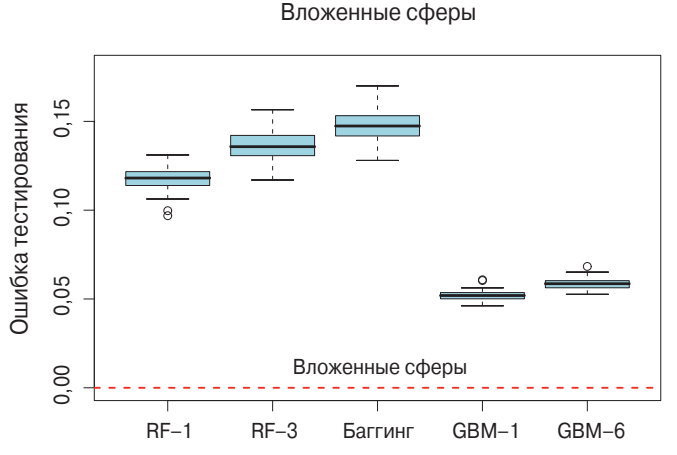

Рис. 15.2. Результаты 50 симуляций по модели вложенных сфер в пространстве $\mathbb{R}^{10}$. Граница байесовского решения — это поверхность сферы (аддитивная). Метка “RF–3” относится к случайному лесу с $m = 3$, а “GBM–6” — к модели с градиентным бустингом с порядком взаимодействия, равным шести; аналогично для “RF–1” и “GBM–1”. Обучающие множества содержали 2000 элементов, а тестовые множества — 10 000

На рис. 15.3 случайные леса сравниваются с бустингом (со сжатием) при решении регрессии с использованием данных о жилье в Калифорнии (см. раздел 10.14.1 главы 10). Здесь проявляются две сильные особенности.

- Случайные леса стабилизируются примерно на 200 деревьях, в то время как на 1000 деревьев бустинг продолжает улучшать точность. Бустинг замедляется из-за сжатия, а также из-за того, что деревья имеют намного меньшую глубину.
- Бустинг в данном случае превосходит случайные леса. При 1000 членах более слабая модель бустинга (GBM с глубиной, равной четырем) имеет меньшую ошибку, чем более сильный случайный лес (RF $m = 6$); p-значение критерия Уилкоксона об абсолютной разности между математическими ожиданиями равно 0,007. При больших m случайные леса работают не лучше.

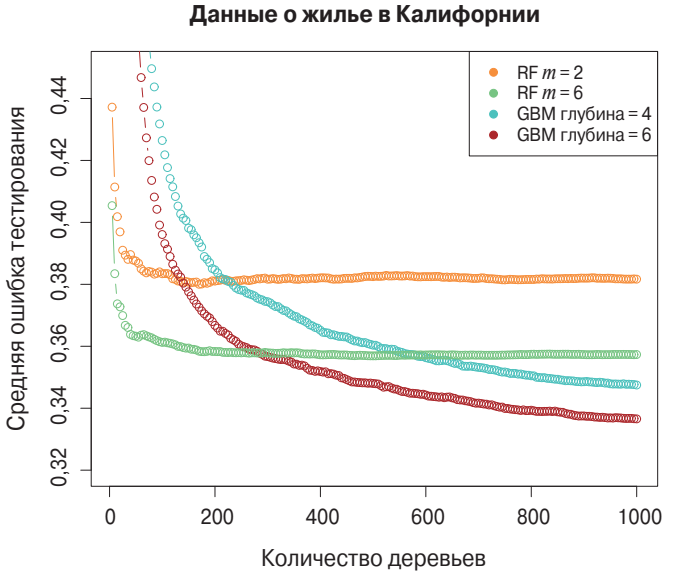

Рис. 15.3. Сравнение случайных лесов с градиентным бустингом на данных о жилье в Калифорнии. Кривые представляют среднюю абсолютную ошибку на тестовых данных как функцию от количества деревьев в моделях. Показаны два случайных леса с $m = 2$ и $m = 6$. Две модели градиентного бустинга используют параметр сжатия $\nu = 0{,}05$ из (10,41). Их глубины взаимодействия равны 4 и 6. Модели бустинга лучше, чем случайные леса

15.3. Детали случайных лесов

Мы не стали подчеркивать различие между случайными лесами при классификации и регрессии. При использовании для классификации случайный лес получает голос класса по каждому дереву, а затем классифицирует его по большинству голосов (аналогичное обсуждение см. в разделе 8.7 главы 8, посвященном баггингу). При использовании для регрессии прогнозы для каждого дерева в целевой точке x просто усредняются, как в (15.2). Кроме того, авторы дают следующие рекомендации.

- Для классификации значение m по умолчанию равно $\lfloor \sqrt{p} \rfloor$, а минимальный размер узла равен единице.
- Для регрессии значение m по умолчанию равно $\lfloor p/3 \rfloor$, а минимальный размер узла равен пяти.

На практике наилучшие значения этих параметров зависят от задачи, и их следует рассматривать как параметры настройки. На рис. 15.3 алгоритм намного лучше работает при $m = 6$, чем при значении по умолчанию, равном $\lfloor 8/3 \rfloor = 2$.

15.3.1. Выборки, не вошедшие в набор

Важной особенностью случайных лесов является использование *выборок, не вошедших в набор* (Out-Of-Bag — OOB).

> *Для каждого наблюдения $z_i = (x_i, y_i)$ создайте его предиктор случайного леса, усредняя только те деревья, которые соответствуют бутстрэп-выборкам, в которых наблюдение z_i не появлялось.*

Оценка ошибки OOB почти идентична оценке, полученной с помощью N-блочной перекрестной проверки (см. упражнение 15.2). Следовательно, в отличие от многих других нелинейных оценок, случайные леса могут быть обучены одновременно с выполнением перекрестной проверки. Как только ошибка OOB стабилизируется, обучение можно прекратить.

На рис. 15.4 показан уровень ошибок классификации OOB для данных о спаме по сравнению с ошибкой тестирования. Хотя здесь усреднено 2500 деревьев, на графике видно, что около 200 уже будет достаточно.

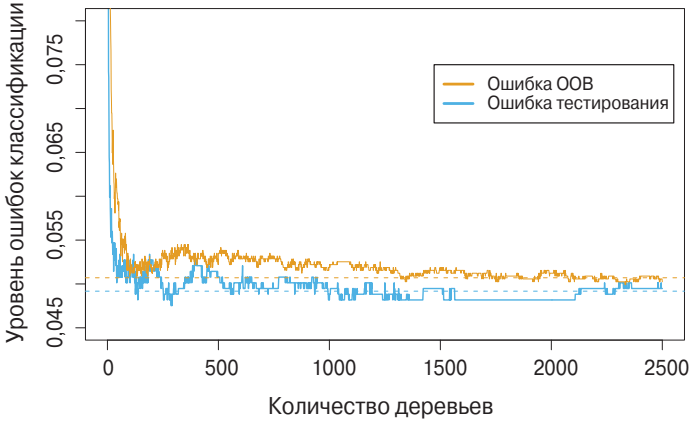

Рис. 15.4. Сравнение ошибки OOB, вычисленной на обучающих данных о спаме, с ошибкой тестирования, вычисленной на тестовом множестве

15.3.2. Значимость переменной

Так же как и для моделей с градиентным бустингом (см. раздел 10.13), для случайных лесов можно строить графики значимости переменных. При каждом разделении в каждом дереве улучшение критерия разделения является показателем значимости, который приписывается переменной разделения и накапливается по всем деревьям, входящим в лес, отдельно для каждой переменной. На левом графике на рис. 15.5 показаны уровни значимости переменных, рассчитанные таким образом по данным о спаме (сравните его с соответствующим рис. 10.6, построенным для градиентного бустинга). Бустинг полностью игнорирует некоторые переменные, а случайный лес — нет. Выбор подходящей переменной разделения увеличивает вероятность того, что любая отдельная переменная будет включена в случайный лес, в то время как при бустинге такой селекции не происходит.

Случайные леса также используют выборки OOB, чтобы построить другую *меру значимости* переменных и тем самым измерить прогностическую силу каждой переменной. После того как b-е дерево построено, через него пропускают выборки OOB и записывают точность прогноза. Затем значения j-й переменной в выборках OOB случайным образом переставляются, и снова вычисляется точность. Уменьшение точности в результате этой перестановки усредняется по всем деревьям и используется как мера значимости переменной j в случайном лесу. Они выражены в процентах от максимума на правом графике на рис. 15.5. Хотя ранжирование в этих двух методах похожи, значимость переменных на правом графике выглядит более равномерной. Рандомизация эффективно аннулирует влияние переменной, аналогично обнулению коэффициента в линейной модели (см. упражнение 15.7). Она не измеряет влияние отсутствия переменной на прогноз, потому что, если модель была обучена без этой переменной, в качестве ее суррогатов могли использоваться другие переменные.

15.3.3. Диаграмма близости

Одним из рекламируемых результатов случайного леса является *диаграмма близости* (proximity plot). На рис. 15.6 показана диаграмма близости для смешанных данных, описанных в разделе 2.3.3 главы 2. При построении случайного леса для обучающих данных накапливается матрица близости $N \times N$. Для каждого дерева близость любой пары наблюдений OOB, совместно использующих конечный узел, увеличивается на единицу. Эта матрица близости затем представляется в двух измерениях с использованием многомерного шкалирования (см. раздел 14.8). Идея состоит в том, что, хотя данные могут быть многомерными, включая смешанные переменные и т.д., диаграмма близости дает представление о том, какие наблюдения фактически близки с точки зрения классификатора на основе случайного леса.

Диаграммы близости для случайных лесов часто выглядят очень похожими, независимо от данных, что ставит под сомнение их полезность. Они, как правило, имеют форму звезды, по одному лучу на класс, которые тем четче выделены, чем лучше классификация.

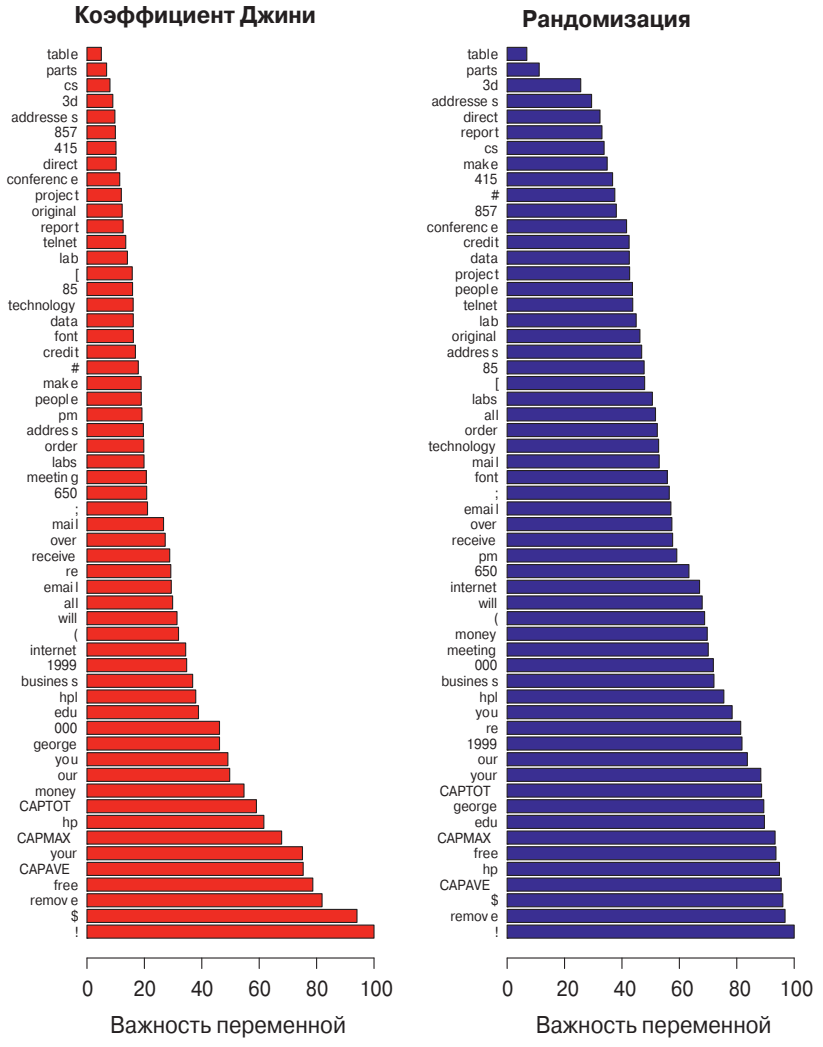

Рис. 15.5. Графики значимости переменных для классификации случайного леса, выращенного по данным о спаме. Левый график основывается на важности индекса расщепления Джини, как при градиентном бустинге. Рейтинги сопоставимы с рейтингами, полученными путем градиентного бустинга (см. рис. 10.6). На правом графике для вычисления значимости переменных использована случайная рандомизация, и значимость распределена более равномерно

Так как смешанные данные являются двумерными, мы можем отобразить точки на диаграмме близости в исходные координаты и лучше понять, что они собой представляют. Похоже, что точки в “чистых” областях отображаются по классам на лучи звезды, а точки, расположенные ближе к границам решения, отображаются ближе к центру. Это не удивительно, если изучить структуру матриц близости. Соседние точки в “чистых” областях часто оказываются в одном и том же сегменте, поскольку,

если конечный узел является “чистым”, он больше не разделяется алгоритмом построения деревьев случайного леса. С другой стороны, пары точек, которые близки, но принадлежат разным классам, иногда совместно используют терминальный узел, хотя и не всегда.

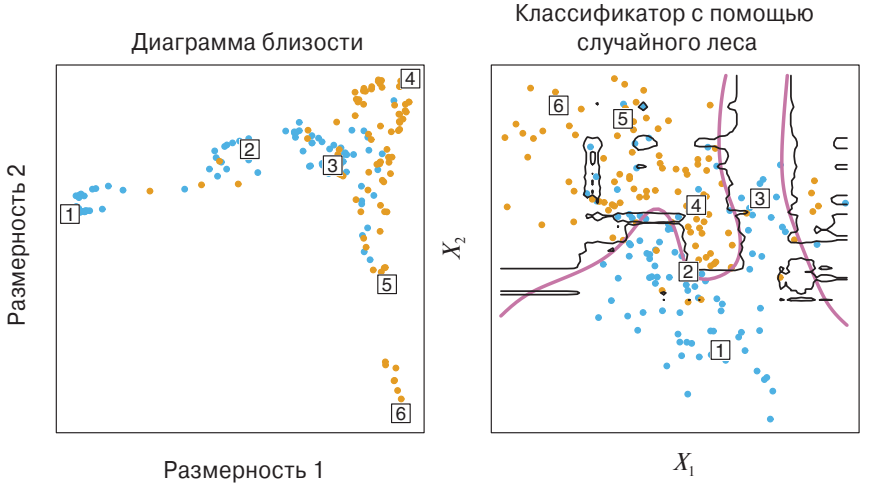

Рис. 15.6. Диаграмма близости для классификатора на основе случайного леса, построенного по смешанным данным (*слева*). Граница решения и обучающие данные для случайного леса, построенного по смешанным данным (*справа*). На каждой диаграмме были определены шесть точек

15.3.4. Случайные леса и переобучение

Если количество переменных велико, но доля релевантных переменных мала, то случайные леса, вероятно, будут плохо работать при малых m. При каждом разделении вероятность того, что будут выбраны соответствующие переменные, может быть небольшой. На рис. 15.7 показаны результаты моделирования, подтверждающего это утверждение. Подробности приведены в подписи к рис. 15.7 и упражнении 15.3. Сверху каждой пары мы видим гипергеометрическую вероятность того, что соответствующая переменная будет выбрана при любом разбиении дерева случайного леса (в этом моделировании все соответствующие переменные равны значимости). Как только эта вероятность становится небольшой, разрыв между бустингом и случайными лесами увеличивается. Когда количество релевантных переменных увеличивается, точность случайных лесов оказывается удивительно устойчивой к увеличению количества шумовых переменных. Например, с шестью релевантными и 100 шумовыми переменными вероятность выбора релевантной переменной при любом разделении равна 0,46, с учетом того, что $m = \sqrt{(6+100)} \approx 10$. Согласно рис. 15.7, это не влияет на точность метода случайного леса по сравнению с бустингом. Эта устойчивость в значительной степени обусловлена относительной нечувствительностью

стоимости ошибочной классификации к смещению и дисперсии оценок вероятности в каждом дереве. Мы рассмотрим случайные леса в контексте регрессии в следующем разделе.

Еще одно утверждение состоит в том, что случайные леса не могут переобучаться. Конечно, верно, что увеличение B не приводит к переобучению последовательности случайных лесов. Как и в случае с баггингом, оценка случайных лесов (15.2) аппроксимирует математическое ожидание

$$\hat{f}_{\text{rf}}(x) = \mathrm{E}_{\Theta} T(x; \Theta) = \lim_{B \to \infty} \hat{f}(x)_{\text{rf}}^{B} \tag{15.3}$$

с помощью среднего значения по B реализаций Θ. Распределение Θ здесь зависит от обучающих данных. Однако *этот предел может переобучаться по данным*; среднее количество полностью построенных деревьев может привести к слишком богатой модели и ненужным отклонениям. Segal (2004) демонстрирует небольшой прирост точности, контролируя глубину отдельных деревьев, построенных в случайных лесах. Наш опыт показывает, что использование полноценных деревьев редко заслуживает внимания и приводит к уменьшению количества параметров настройки всего на единицу.

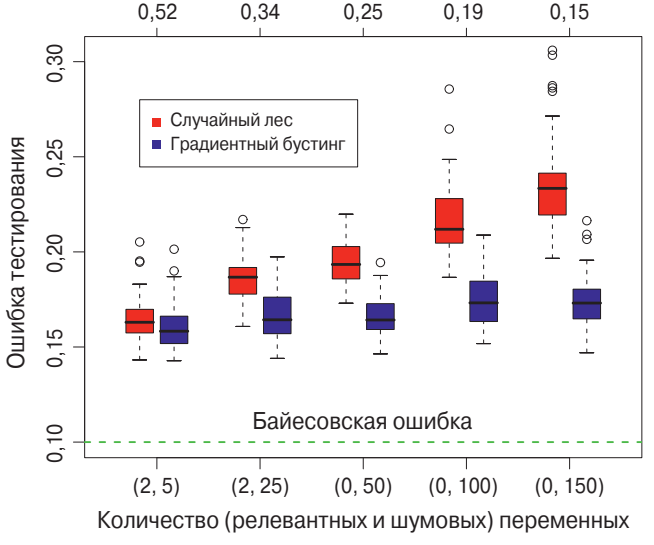

Рис. 15.7. Сравнение случайных лесов и градиентного бустинга на задачах с увеличивающимся количеством шумовых переменных. В каждом случае истинная граница решения зависит от двух переменных, и в задачу включается все больше шумовых переменных. Случайные леса используют значение по умолчанию $m = \sqrt{p}$. Сверху каждой пары указана вероятность того, что одна из релевантных переменных будет выбрана при любом разделении. Результаты основаны на 50 симуляциях для каждой пары, с обучающей выборкой, содержащей 300 элементов, и тестовой выборкой, содержащей 500 элементов. См. упражнение 15.3

На рис. 15.8 показан скромный эффект контроля глубины на простом примере регрессии. Классификаторы менее чувствительны к дисперсии, и этот эффект переобучения редко наблюдается при классификации с помощью случайных лесов.

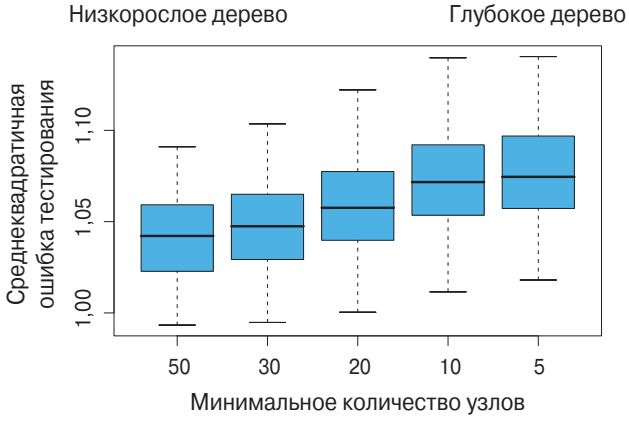

Рис. 15.8. Влияние размера дерева на ошибку регрессии с помощью случайного лесна. В этом примере истинная поверхность была аддитивной по двум из 12 переменных, плюс аддитивная единичная дисперсия нормально распределенного шума. Глубина дерева здесь контролируется минимальным размером узла; чем меньше минимальный размер узла, тем глубже деревья

15.4. Анализ случайных лесов

В этом разделе анализируются механизмы дополнительной рандомизации, используемой в случайных лесах. В этом обсуждении мы сконцентрируемся на регрессии с квадратичной функцией потерь, поскольку анализ смещения и дисперсии для бинарной функции потерь намного сложнее (см. раздел 7.3.1). Кроме того, даже в случае классификации мы можем рассматривать среднее значение по случайному лесу как оценку апостериорных вероятностей класса, для которых смещение и дисперсия являются подходящими дескрипторами.

15.4.1. Дисперсия и эффект декорреляции

Предельная форма ($B \to \infty$) оценки регрессии по методу случайного леса имеет вид

$$\hat{f}_{rf}(x) = \mathrm{E}_{\Theta|\mathbf{Z}} T\left(x;\ \Theta(\mathbf{Z})\right), \tag{15.4}$$

где мы подчеркнули зависимость от обучающих данных $\mathbf{Z}$. Здесь мы рассмотрим оценку в одной целевой точке x. Из (15.1) следует, что

$$\mathrm{Var}\,\hat{f}_{\mathrm{rf}}(x) = \rho(x)\sigma^2(x). \tag{15.5}$$

Здесь

- $\rho(x)$ — *выборочная* корреляция между любой парой деревьев, используемой при усреднении:

$$\rho(x) = \text{corr}[T(x; \Theta_1(\mathbf{Z})), T(x; \Theta_2(\mathbf{Z}))], \tag{15.6}$$

- где $\Theta_1(\mathbf{Z})$ и $\Theta_2(\mathbf{Z})$ — случайно извлеченная пара деревьев случайного леса, построенных по случайно выбранной переменной $\mathbf{Z}$;
- $\sigma^2(x)$ — выборочная дисперсия любого произвольно построенного дерева

$$\sigma^2(x) = \text{Var}\, T(x; \Theta(\mathbf{Z})). \tag{15.7}$$

Величину $\rho(x)$ легко спутать со средней корреляцией между обученными деревьями в *заданном* ансамбле случайных лесов. Иначе говоря, обученные деревья можно интерпретировать как вектор из N элементов и вычислить среднюю попарную корреляцию между этими векторами, обусловленную данными. Это *неправильно*; эта условная корреляция не имеет прямого отношения к процессу усреднения, и зависимость $\rho(x)$ от x свидетельствует об этом отличии. Скорее $\rho(x)$ — это теоретическая корреляция между парой деревьев случайного леса, оцененных в точке x, индуцированная многократным извлечением обучающей выборки $\mathbf{Z}$ из генеральной совокупности с последующим извлечением пары деревьев случайного леса. На статистическом жаргоне эта величина называется *выборочной корреляцией* (sampling distribution) переменных $\mathbf{Z}$ и Θ.

Точнее говоря, переменная, усредненная по вычислениям (15.6) и (15.7), является

- зависимой от $\mathbf{Z}$: из-за бутстрэп-выборки и выборки признаков при каждом разделении
- и результатом изменчивости выборки самой переменной $\mathbf{Z}$.

Фактически условная ковариация пары деревьев, обученных в точке x, равна нулю, поскольку бутстрэп-выборка и выборка признаков являются независимыми и одинаково распределенными (см. упражнение 15.5).

Следующие примеры основаны на имитационной модели:

$$Y = \frac{1}{\sqrt{50}} \sum_{j=1}^{50} X_j + \varepsilon, \tag{15.8}$$

где все X_j и ε являются независимыми и одинаково нормально распределенными. Мы используем 500 обучающих выборок размером 100 и один набор тестовых точек размером 600. Поскольку деревья регрессии нелинейны относительно $\mathbf{Z}$, шаблоны, которые приведены ниже, будут несколько отличаться в зависимости от структуры модели.

На рис. 15.9 показано, как корреляция (15.6) между парами деревьев уменьшается с уменьшением m: пары предсказаний деревьев в точке x для разных обучающих наборов $\mathbf{Z}$, вероятно, будут менее похожими, если они не используют одинаковые переменные расщепления.

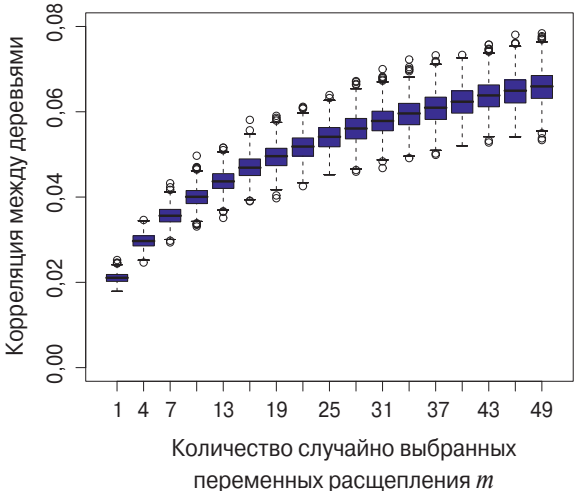

Рис. 15.9. Значения корреляции между парами деревьев, построенными алгоритмом регрессии с помощью случайного леса, в зависимости от m. Квадратные диаграммы представляют корреляции в 600 случайно выбранных точках прогнозирования x

На левой панели рис. 15.10 мы рассматриваем дисперсии предикторов одного дерева, $\mathrm{Var}\,T(x;\Theta(\mathbf{Z}))$ (усредненные по 600 точкам прогнозирования x, взятым случайным образом из нашей имитационной модели). Это общая дисперсия, которая может быть разложена на две части с использованием стандартных аргументов условной дисперсии (см. упражнение 15.5):

$$\mathrm{Var}_{\Theta|\mathbf{Z}}T(x;\Theta(\mathbf{Z})) = \mathrm{Var}_{\mathbf{Z}}\mathrm{E}_{\Theta|\mathbf{Z}}T(x;\Theta(\mathbf{Z})) + \mathrm{E}_{\mathbf{Z}}\mathrm{Var}_{\Theta|\mathbf{Z}}T(x;\Theta(\mathbf{Z})),$$

$$\textit{Общая дисперсия} = \mathrm{Var}_{\mathbf{Z}}\hat{f}_{\mathrm{rf}}(x) \qquad + \textit{внутренняя дисперсия } \mathbf{Z} \quad (15.9)$$

Второе слагаемое — это внутренняя дисперсия переменной $\mathbf{Z}$ (within-$\mathbf{Z}$ variance), представляющая собой результат рандомизации, который увеличивается с уменьшением m. Первое слагаемое фактически является выборочной дисперсией ансамбля случайных лесов (показан на правой панели), которая уменьшается с уменьшением m. Дисперсия отдельных деревьев не изменяется заметно в большей части диапазона m, поэтому в свете (15.5) дисперсия ансамбля значительно ниже, чем эта дисперсия дерева.

15.4.2. Смещение

Как и в случае с баггингом, смещение случайного леса такое же, как у любого отдельного дерева $T(x;\Theta(\mathbf{Z}))$:

$$\begin{aligned}\mathrm{Bias}(x) &= \mu(x) - \mathrm{E}_{\mathbf{Z}}\hat{f}_{\mathrm{rf}}(x) = \\ &= \mu(x) - \mathrm{E}_{\mathbf{Z}}\mathrm{E}_{\Theta|\mathbf{Z}}T\big(x;\Theta(\mathbf{Z})\big).\end{aligned} \quad (15.10)$$

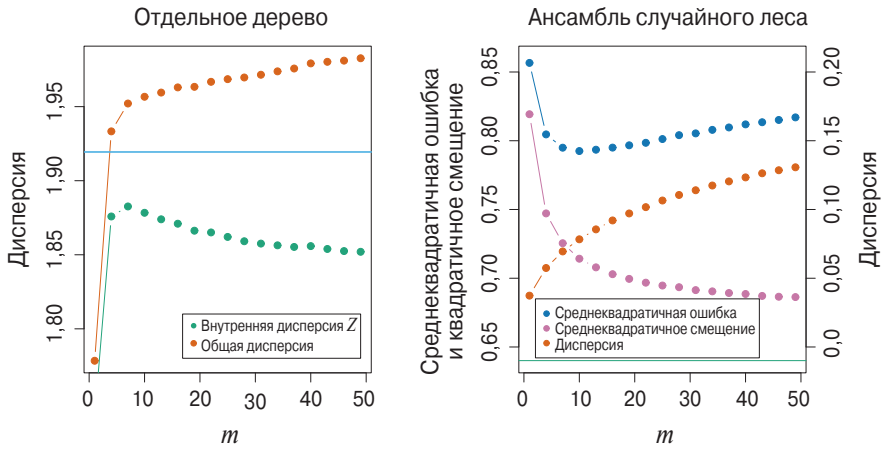

Рис. 15.10. Результаты моделирования. На левой панели показана средняя дисперсия одного дерева случайного леса как функция от m. Внутренняя дисперсия $\mathbf{Z}$ — это средний вклад внутри выборки в дисперсию, полученную в результате бутстрэпа и выборки разделяющей переменной (15.9). Общая дисперсия включает в себя изменчивость выборки $\mathbf{Z}$. Горизонтальная линия представляет собой среднюю дисперсию одного полностью построенного дерева (без бутстрэпа). На правой панели показаны среднеквадратичная ошибка, квадратичное смещение и дисперсия ансамбля как функция от m. Обратите внимание на то, что ось отклонения находится справа (тот же масштаб, другой уровень). Горизонтальная линия представляет собой среднеквадратическое смещение полностью построенного дерева

Она также, как правило, больше (в абсолютном выражении), чем смещение необрезанного дерева, выращенного до $\mathbf{Z}$, поскольку рандомизация и уменьшенное выборочное пространство накладывают свои ограничения. Следовательно, улучшения в прогнозировании, полученные с помощью баггинга или случайных лесов, являются *исключительно результатом уменьшения дисперсии*.

Любое обсуждение смещения зависит от неизвестной истинной функции. На рис. 15.1, *справа*, показан квадрат смещения для модели аддитивной модели (оценен по 500 реализациям). Хотя для разных моделей форма и скорость кривых смещения могут отличаться, общая тенденция заключается в том, что с уменьшением m смещение увеличивается. На рисунке показана среднеквадратичная ошибка, и мы видим классический компромисс между смещением и дисперсией при выборе m. Для всех m квадрат смещения случайного леса больше, чем для одного дерева (горизонтальная линия).

Эти закономерности предполагают сходство с гребневой регрессией (см. раздел 3.4.1). Гребневая регрессия полезна (в линейных моделях), когда имеется большое количество переменных с коэффициентами одинакового размера; она сжимает их коэффициенты к нулю, а коэффициенты сильно коррелированных переменных — друг к другу. Хотя размер обучающей выборки может не позволить включить в модель все переменные, эта регуляризация стабилизирует модель и позволяет всем переменным иметь свое влияние (хотя и уменьшенное). Случайные леса с небольшим m выполня-

ют аналогичное усреднение. Каждая из соответствующих переменных получает свою очередь для первичного разделения, а усреднение по ансамблю уменьшает вклад любой отдельной переменной. Так как этот пример моделирования (15.8) основан на линейной модели по всем переменным, гребневая регрессия достигает более низкой среднеквадратичной ошибки (около 0,45 с $\mathrm{df}(\lambda_{\mathrm{opt}}) \approx 29$).

15.4.3. Адаптивный метод ближайших соседей

Классификатор на основе случайного леса имеет много общего с классификатором по методу k ближайших соседей (см. раздел 13.3); фактически, его взвешенная версия. Поскольку каждое дерево строится до максимального размера, для конкретного Θ^* величина $T(x; \Theta^*(\mathbf{Z}))$ является значением отклика для одной из обучающих выборок[4]. Алгоритм построения деревьев находит оптимальный путь к этому наблюдению, выбирая наиболее информативные предикторы из имеющихся в его распоряжении. Процесс усреднения присваивает веса этим обучающим ответам, которые в конечном итоге голосуют за прогноз. Следовательно, посредством механизма голосования в случайном лесу этим наблюдениям, близким к целевой точке, присваиваются веса (эквивалентное ядро), которые объединяются для формирования решения о классификации.

Рис. 15.11 демонстрирует сходство между границей решения по методу трех ближайших соседей и методом случайного леса на смешанных данных.

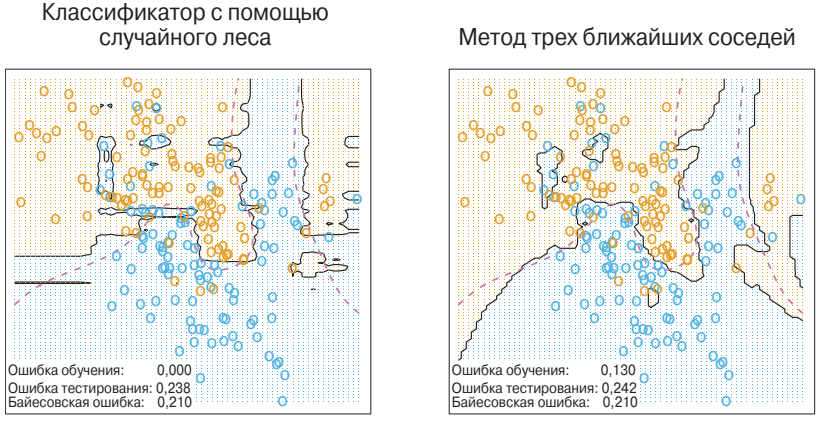

Рис. 15.11. Результаты применения метода случайного леса и метода 3NN к данным о смеси. Ориентация отдельных деревьев в случайном лесу вдоль осей приводит к областям решений с границами, ориентированными примерно по осям

[4] Мы игнорируем тот факт, что чистые узлы не разделяются дальше, и, следовательно, в терминальном узле может быть более одного наблюдения.

Библиографические заметки

Случайные леса, описанные здесь, были введены в работе Breiman (2001), хотя многие идеи ранее встречались в литературе в разных формах. В частности, Но (1995) ввел термин “случайный лес” и использовал консенсус деревьев, выращенных в случайных подпространствах признаков. Идея использования стохастического возмущения и усреднения для избежания переобучения была предложена в работах Kleinberg (1990), а затем Kleinberg (1996). Amit and Geman (1997) использовали рандомизированные деревья, построенные на признаках изображений для задач классификации изображений. Breiman (1996a) изобрел баггинг — предшественника его версии случайных лесов. Dietterich (2000b) также предложил улучшение баггинга с помощью дополнительной рандомизации. Его подход состоял в том, чтобы ранжировать 20 лучших вариантов разделения в каждом узле, а затем выбирать разделение из списка случайным образом. Он показал с помощью моделирования и реальных примеров, что эта дополнительная рандомизация улучшает точность по сравнению с точностью баггинга. Friedman and Hall (2007) показали, что отбор проб (без замены) является эффективной альтернативой баггингу. Они показали, что построение и усреднение деревьев на выборках размера $N/2$ приблизительно эквивалентно (с точки зрения смещения/дисперсии) баггингу, в то время как использование меньших долей N еще больше уменьшает дисперсию (благодаря декорреляции).

Существует несколько бесплатных программных реализаций случайных лесов. В этой главе мы использовали пакет **`randomForest`** в языке R, поддерживаемый Энди Лиу (Andy Liaw), который доступен на веб-сайте CRAN. Он позволяет как выбирать переменную разделения, так и субдискретизацию. Адель Катлер поддерживает сайт случайных лесов **`http://www.math.usu.edu/~adele/forests/`**, где (по состоянию на август 2008 года) свободно распространяются программы, написанные Лео Брейманом и Адель Катлер. Их код и название “случайные леса” исключительно лицензированы компанией Salford Systems для коммерческого использования. Архив по машинному обучению **`Weka http://www.cs.waikato.ac.nz/ml/weka/`** в Университете Вайкато, Новая Зеландия, предлагает бесплатную Java-реализацию случайных лесов.

Упражнения

15.1. Выведите формулу для дисперсии (15.1). Она не работает, если ρ меньше нуля. Опишите проблему, которая возникает в этом случае.

15.2. Покажите, что по мере того, как количество бутстрэп-выборок B становится большим, оценка ошибки OOB для случайного леса приближается к оценке ошибки N-блочной перекрестной проверки и что в пределе тождество является точным.

15.3. Рассмотрим имитационную модель, использованную на рис. 15.7 (Mease and Wyner, 2008). Бинарные наблюдения генерируются с вероятностями

$$\Pr\left(Y=1\middle|X\right)=q+\left(1-2q\right)I\left[\sum_{j=1}^{J}X_j>J/2\right], \tag{15.11}$$

где $X\sim U\left[0,\,1\right]^p$, $0\le q\le 1/2$, а $J\le p$ — некоторое заранее заданное (четное) число. Опишите эту поверхность вероятности и вычислите байесовскую ошибку.

15.4. Пусть x_i, $i = 1, \ldots, N$ являются независимыми и одинаково распределенными с параметрами (μ, σ^2). Пусть $\bar{x}_1^*$ и $\bar{x}_2^*$ — две бутстрэп-реализации выборочного среднего. Покажите, что корреляция выборки $\operatorname{corr}\left(\bar{x}_1^*,\,\bar{x}_2^*\right)=\dfrac{n}{2n-1}\approx 50\%$. Попутно выведите $\operatorname{var}\left(\bar{x}_1^*\right)$ и дисперсию среднего значения при баггинге $\bar{x}_{bag}$. Здесь $\bar{x}$ — *линейная* статистика; баггинг не уменьшает дисперсию для линейной статистики.

15.5. Покажите, что выборочная корреляция между парой случайных деревьев в точке x определяется как

$$\rho\left(x\right)=\frac{\operatorname{Var}_{\mathbf{Z}}\left[\operatorname{E}_{\Theta|\mathbf{Z}}T\left(x;\Theta\left(\mathbf{Z}\right)\right)\right]}{\operatorname{Var}_{\mathbf{Z}}\left[\operatorname{E}_{\Theta|\mathbf{Z}}T\left(x;\Theta\left(\mathbf{Z}\right)\right)\right]+\operatorname{E}_{\mathbf{Z}}\operatorname{Var}_{\Theta|\mathbf{Z}}\left[T\left(x;\Theta\left(\mathbf{Z}\right)\right)\right]}. \tag{15.12}$$

Числитель равен $\operatorname{Var}_{\mathbf{Z}}\left[\hat{f}_{\mathrm{rf}}\left(x\right)\right]$, а второе слагаемое в знаменателе является математическим ожиданием условной дисперсии из-за рандомизации в случайных лесах.

15.6. Обучите ряд классификаторов на основе случайного леса для данных о спаме, чтобы изучить их чувствительность к параметру m. Отобразите как ошибку OOB, так и ошибку тестирования с соответствующим выбранным диапазоном значений для m.

15.7. Предположим, мы обучаем модель линейной регрессии по N наблюдениям с откликом y_i и предикторами $x_{i1}, \ldots, x_{ip}$. Предположим, что все переменные стандартизированы, т.е имеют нулевое математическое ожидание и единичное стандартное отклонение. Пусть RSS — среднеквадратичная невязка обучающих данных и $\hat{\beta}$ — оцениваемый коэффициент. Обозначим через RSS_j^* среднеквадратичную невязку на обучающих данных, использующих ту же самую оценку $\hat{\beta}$, но с N значениями для j-й переменной, случайным образом переставленными до вычисления прогнозов. Покажите, что

$$\operatorname{E}_P\left[RSS_j^*-RSS\right]=2\hat{\beta}_j^2, \tag{15.13}$$

где E_P — математическое ожидание относительно распределения перестановок. Докажите, что это приблизительно верно, если оценки выполняются с использованием независимого тестового множества.