Классификатор дерева решений

Классификатор дерева решений

- Пусть обучающий набор $\mathbf{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ состоит из n точек в d-мерном пространстве, причем y_i - метка класса для точки $\mathbf{x}_i$. Мы предполагаем, что измерения или атрибуты X_j являются числовыми или категориальными и что существует k различных классов, так что $y_i \in \{c_1, c_2, \ldots, c_k\}$.
- Классификатор дерева решений - это рекурсивная модель дерева на основе секционирования, которая предсказывает класс $\hat{y}_i$ для каждой точки $\mathbf{x}_i$.

- Пусть **R** обозначает пространство данных, охватывающее множество входных точек **D**. Дерево решений использует осепараллельную гиперплоскость для разделения пространства данных **R** на два результирующих полупространства или области, скажем $\mathbf{R}_1$ и $\mathbf{R}_2$, что также вызывает разбиение входных точек на $\mathbf{D}_1$ и $\mathbf{D}_2$ соответственно. Каждая из этих областей рекурсивно разбивается с помощью осепараллельных гиперплоскостей до тех пор, пока точки внутри индуцированного разбиения не станут относительно чистыми с точки зрения их меток классов, то есть большинство точек принадлежат к одному и тому же классу. Результирующая иерархия разделенных решений образует **модель дерева решений**, с конечными узлами, помеченными классом большинства среди точек в этих регионах.
- Чтобы классифицировать новую тестовую точку, мы должны рекурсивно оценить, к какому полупространству она принадлежит, пока не достигнем листового узла в дереве решений, в котором мы предсказываем ее класс как метку листа

Деревья решений

- **Дерево решений** состоит из внутренних узлов, представляющих решения, соответствующие гиперплоскостям или точкам разделения и конечных узлов, представляющих области или разделы пространства данных, которые помечены классом большинства. Область характеризуется подмножеством точек данных, лежащих в этой области.

Осепараллельные гиперплоскости

- Гиперплоскость h(x) определяется как множество всех точек **x**, удовлетворяющих следующему уравнению: $h(\mathbf{x}) : \mathbf{w}^T\mathbf{x} + b = 0$ (19.1)
- Здесь $\mathbf{w} \in R^d$ - ***весовой вектор***, нормальный к гиперплоскости, а b - смещение гиперплоскости от начала координат. Дерево решений рассматривает только ***осепараллельные гиперплоскости***, то есть весовой вектор должен быть параллелен одному из исходных измерений или осям X_j. Иными словами, весовой вектор w ***априори*** ограничен одним из стандартных базисных векторов $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$, где $\mathbf{e}_i \in R^d$ имеет 1 для j-го измерения и 0 для всех других измерений. Если $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ и предполагая $\mathbf{w} = \mathbf{e}_j$, мы можем переписать уравнение (19.1) как $h(\mathbf{x}) : \mathbf{e}_j^T\mathbf{x} + b = 0$, из чего следует $h(\mathbf{x}) : x_j + b = 0$, где выбор смещения b дает различные гиперплоскости вдоль размерности X_j.

Точки Разделения

- Гиперплоскость определяет точку принятия решения или разделения, поскольку она разбивает пространство данных **R** на два полупространства. Все точки x, такие, h(**x**) ≤ 0, находятся на гиперплоскости или на одной стороне гиперплоскости, тогда как все точки, такие, что h(**x**) > 0, находятся на другой стороне. Точки разделения, связанного с осью параллельной гиперплоскости можно записать как h(**x**) ≤ 0, что означает, что $x_i + b \leq 0$ или $x_i \leq -b$. Поскольку x_i - это некоторое значение из измерения X_j и смещение b может быть выбрано как любое значение, общая форма точки разделения для числового атрибута X_j задается как

$$\boldsymbol{X_j \leq v},$$

где v = - b - некоторое значение в области атрибута X_j. Таким образом, решение или точка разделения $X_j \leq v$ разбивает пространство входных данных R на две области $\mathbf{R}_Y$ и $\mathbf{R}_N$, которые обозначают множество *всех возможных точек*, удовлетворяющих решению, и тех, которые не удовлетворяют.

Раздел Данных

- Каждое разбиение **R** на $\mathbf{R}_Y$ и $\mathbf{R}_N$ также вызывает двоичное разбиение соответствующих входных точек данных **D**. То есть точка разбиения вида $X_j \leq v$ вызывает разбиение данных

$$\mathbf{D}_Y = \{x \mid x \in \mathbf{D}, x_j \leq v\}$$

$$\mathbf{D}_N = \{x \mid x \in \mathbf{D}, x_j > v\},$$

где $\mathbf{D}_Y$ - подмножество точек данных, лежащих в области $\mathbf{R}_Y$, а $\mathbf{D}_N$ - подмножество входных точек, лежащих в области $\mathbf{R}_N$.

Чистота

- Чистота области $\mathbf{R}_j$ определяется в терминах смеси классов для точек в соответствующем разделе данных $\mathbf{D}_j$. Формально, чистота — это доля точек с меткой большинства в $\mathbf{D}_j$, т. е.
- $$purity(D_j) = max_i \left\{ \frac{n_{ji}}{n_j} \right\} \quad (19.2),$$
- где $n_j = |\mathbf{D}_j|$ - общее количество точек данных в области $\mathbf{R}_j$, а $n_{j\,i}$ - количество точек в $\mathbf{D}_j$ с меткой класса c_i.

Категориальные Атрибуты

- Для категориального атрибута X_j точки разделения или решения имеют вид $X_j \in V$, где $V \subset \text{dom}(X_j)$, а $\text{dom}(X_j)$ обозначает область для X_j. Интуитивно это расщепление можно считать категориальным аналогом гиперплоскости. Это приводит к двум "полупространствам", одна область $\mathbf{R}_Y$ состоит из точек **x**, удовлетворяющих условию $x_i \in V$, а другая область $\mathbf{R}_N$ состоит из точек, удовлетворяющих условию $x_i \notin V$.

Правила Принятия Решений

- Одним из преимуществ деревьев решений является то, что они создают модели, которые относительно легко интерпретировать. В частности, дерево может быть прочитано как набор решающих правил, причем антецедент каждого правила включает в себя решения по внутренним узлам вдоль пути к листу, а его следствием является метка листового узла. Далее, поскольку все области не пересекаются и охватывают все пространство, набор правил можно интерпретировать как набор альтернатив или дизъюнкций.

Алгоритм дерева решений

- Псевдокод для построения модели дерева решений показан в алгоритме 19.1. Алгоритм принимает в качестве входных данных обучающий набор данных **D** и два параметра η и π, где η - размер листа, а π - порог чистоты листа. Различные точки разделения оцениваются для каждого атрибута в **D**. числовые решения имеют вид Xj ≤ v для некоторого значения v в диапазоне значений атрибута Xj, а категориальные решения имеют вид Xj ∈ V для некоторого подмножества значений в области Xj . Наилучшая точка разбиения выбирается для разбиения данных на два подмножества, $\mathbf{D}_Y$ и $\mathbf{D}_N$, где $\mathbf{D}_Y$ соответствует всем точкам x ∈ **D**, удовлетворяющим решению разбиения, а $\mathbf{D}_N$ соответствует всем точкам, не удовлетворяющим решению разбиения. Затем метод дерева решений рекурсивно вызывается на $\mathbf{D}_Y$ и $\mathbf{D}_N$. Для остановки процесса рекурсивного секционирования можно использовать ряд условий остановки. Самое простое условие основано на размере раздела **D**. Если число точек n в **D** падает ниже заданного пользователем порога размера η, то мы останавливаем процесс разбиения и делаем **D** листом. Это условие предотвращает чрезмерную подгонку модели к обучающему набору, избегая моделировать очень маленькие подмножества данных. Одного размера недостаточно, потому что, если раздел уже чистый, тогда нет смысла разбивать его дальше. Таким образом, рекурсивное разбиение также прекращается, если чистота D выше порога чистоты π. далее приводится подробная информация о том, как оцениваются и выбираются точки разделения.

ALGORITHM 19.1. Decision Tree Algorithm

```
DECISIONTREE (D, η, π):
1  n ← |D| // partition size
2  n_i ← |{x_j | x_j ∈ D, y_j = c_i}| // size of class c_i
3  purity(D) ← max_i {n_i / n}
4  if n ≤ η or purity(D) ≥ π then // stopping condition
5      c* ← arg max_{c_i} {n_i / n} // majority class
6      create leaf node, and label it with class c*
7      return
8  (split point*, score*) ← (∅, 0) // initialize best split point
9  foreach (attribute X_j) do
10     if (X_j is numeric) then
11         (v, score) ← EVALUATE-NUMERIC-ATTRIBUTE(D, X_j)
12         if score > score* then (split point*, score*) ← (X_j ≤ v, score)
13     else if (X_j is categorical) then
14         (V, score) ← EVALUATE-CATEGORICAL-ATTRIBUTE(D, X_j)
15         if score > score* then (split point*, score*) ← (X_j ∈ V, score)
   // partition D into D_Y and D_N using split point*, and call
      recursively
16 D_Y ← {x ∈ D | x satisfies split point*}
17 D_N ← {x ∈ D | x does not satisfy split point*}
18 create internal node split point*, with two child nodes, D_Y and D_N
19 DECISIONTREE(D_Y); DECISIONTREE(D_N)
```

Меры Оценки С Разделением Точек

- Учитывая точку разделения вида $X_j \leq v$ или $X_j \in V$ для числового или категориального атрибута соответственно, нам нужен объективный критерий оценки точки разделения. Интуитивно мы хотим выбрать точку разделения, которая дает лучшее разделение или различение между различными метками классов.

Энтропия

- Энтропия, как правило, измеряет количество беспорядка или неопределенности в системе. В классификационной установке разбиение имеет более низкую энтропию (или низкий беспорядок), если оно относительно чисто, то есть если большинство точек имеют одну и ту же метку. С другой стороны, разделение имеет более высокую энтропию (или больше беспорядка), если метки классов смешаны, и нет класса большинства как такового.

- Энтропия множества помеченных точек **D** определяется следующим образом:

$$H(D) = -\sum_{i=1}^{k} P(C_i|\boldsymbol{D}) \log_2 P(c_i|\boldsymbol{D}) \text{ (19.3)},$$

- где P (c_i |**D**) - вероятность класса c_i в **D**, А k-число классов. Если область чиста, то есть имеет точки из одного класса, то энтропия равна нулю. С другой стороны, если все классы перемешаны и каждый появляется с равной вероятностью P (c_i |**D**) = $\frac{1}{k}$, то энтропия имеет наибольшее значение, *H*(**D**) = $\log_2$ k.
- Предположим, что точка разделения разбивает **D** на $\mathbf{D}_Y$ и $\mathbf{D}_N$. Определите энтропию разделения как взвешенную энтропию каждого из результирующих разбиений, заданную как

$$H(\boldsymbol{D}_Y, \boldsymbol{D}_N) = \frac{n_Y}{n} H(\boldsymbol{D}_Y) + \frac{n_N}{n} H(\boldsymbol{D}_N) \text{ (19.4)},$$

- где n = | **D** | - количество точек в **D**, а n_Y = |$\mathbf{D}_Y$| и n_N = |$\mathbf{D}_N$ | - количество точек в $\mathbf{D}_Y$ и $\mathbf{D}_N$.

- Чтобы увидеть, приводит ли точка разделения к уменьшению общей энтропии, мы определяем информационный прирост для данной точки разделения следующим образом:

$$Gain(\boldsymbol{D}, \boldsymbol{D}_Y, \boldsymbol{D}_N) = H(\boldsymbol{D}) - H(\boldsymbol{D}_Y, \boldsymbol{D}_N) \quad (19.5)$$

- Чем выше прирост информации, тем больше уменьшение энтропии и тем лучше точка разделения. Таким образом, учитывая точки разбиения и соответствующие им разбиения, мы можем оценить каждую точку разбиения и выбрать ту, которая дает наибольший информационный выигрыш.

Индекс Джини

- Другой распространенной мерой для оценки чистоты точки разделения является индекс Джини, определяемый следующим образом:

$$G(\boldsymbol{D}) = 1 - \sum_{i=1}^{k} P(c_i|\boldsymbol{D})^2 \quad (19.6)$$

- Если разбиение чисто, то вероятность мажоритарного класса равна 1, а вероятность всех остальных классов равна 0, и, таким образом, индекс Джини равен 0. С другой стороны, когда каждый класс представлен одинаково, с вероятностью P (c_i |**D**) = $\frac{1}{k}$, то индекс Джини имеет значение $\frac{k-1}{k}$. Таким образом, более высокие значения индекса Джини указывают на больший беспорядок, а более низкие значения указывают на больший порядок с точки зрения меток классов.

- Мы можем вычислить взвешенный индекс Джини точки разделения следующим образом:

$$G(\boldsymbol{D}_Y, \boldsymbol{D}_N) = \frac{n_Y}{n} G(\boldsymbol{D}_Y) + \frac{n_N}{n} G(\boldsymbol{D}_N),$$

- где n, n_Y и n_N обозначают количество точек в областях **D**, $\mathbf{D}_Y$ и $\mathbf{D}_N$ соответственно. Чем ниже значение индекса Джини, тем лучше точка разделения.
- Вместо энтропии и индекса Джини для оценки расщеплений можно использовать и другие показатели. Например, мера деревьев классификации и регрессии (CART) задается следующим образом

$$CART(\boldsymbol{D}_Y, \boldsymbol{D}_N) = 2 \frac{n_Y}{n} \frac{n_N}{n} \sum_{i=1}^{k} |P(c_i|\boldsymbol{D}_Y) - P(c_i|\boldsymbol{D}_N)| \quad (19.7)$$

- Таким образом, эта мера предпочитает точку разбиения, которая максимизирует разницу между массовой функцией вероятности класса для двух разбиений; чем выше мера CART, тем лучше точка разбиения.

Оценка Точек Разделения

- Все меры оценки точки разделения, такие как энтропия [Ур. (19.3)], индекс Джини [Ур. (19.6)] и CART [Ур. (19.7)], рассмотренные в предыдущем разделе, зависят от класса вероятностной массовой функции (ВМФ) для **D**, а именно $P(c_i | \mathbf{D})$, и класса ВМФ для результирующих разбиений $\mathbf{D}_Y$ и $\mathbf{D}_N$, а именно $P(c_i | \mathbf{D}_Y)$ и $P(c_i | \mathbf{D}_N)$. Обратите внимание, что мы должны вычислить класс ВМФ для всех возможных точек разделения; оценка каждой из них независимо приведет к значительным вычислительным затратам. Вместо этого можно постепенно вычислить ВМФ, как описано в следующих параграфах.

Числовые Атрибуты

- Если X является числовым атрибутом, мы должны оценить точки разделения вида $X \leq v$. даже если мы ограничим v лежащим в пределах диапазона значений атрибута X, для V все равно существует бесконечное число вариантов выбора. Один из разумных подходов состоит в том, чтобы рассматривать только средние точки между двумя последовательными различными значениями X в выборке D. Это происходит потому, что точки разбиения вида $X \leq v$ для $v \in [x_a , x_b)$, где x_a и x_b - два последовательных различных значения X в **D**, производят одно и то же разбиение **D** на $\mathbf{D}_Y$ и $\mathbf{D}_N$ и, таким образом, дают одни и те же оценки. Поскольку для X может быть не более n различных значений, необходимо учитывать не более n-1 средних значений.

- Пусть $\{v_1, \ldots, v_m\}$ обозначает множество всех таких средних точек, таких что v1 < v2 < · · · < vm. Для каждой точки x ≤ V, то есть, чтобы оценить класс ВМФ:

$$\hat{P}(c_i|\boldsymbol{D}_Y) = \hat{P}(c_i|X \leq v) \quad (19.8)$$

$$\hat{P}(c_i|\boldsymbol{D}_Y) = \hat{P}(c_i|X > v) \quad (19.9)$$

- Пусть I () - индикаторная переменная, которая принимает значение 1 только тогда, когда ее аргумент истинен, и 0 в противном случае. Используя теорему Байеса, мы имеем

$$\hat{P}(c_i|X \leq v) = \frac{\hat{P}(X \leq v|c_i)\hat{P}(c_i)}{\hat{P}(X \leq v)} = \frac{\hat{P}(X \leq v|c_i)\hat{P}(c_i)}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \hat{P}(X \leq v|c_j)\hat{P}(c_j)} \quad (19.10)$$

- Априорная вероятность для каждого класса в **D** может быть оценена следующим образом:

$$\hat{P}(c_i) = \frac{1}{n}\sum_{j=1}^{n} I(y_i = c_i) = \frac{n_i}{n} \quad (19.11)$$

- где y_j -класс для точки $\mathbf{x}_j$, n = |**D**| - общее количество точек, а n_i - количество точек в **D** с классом c_i .

- Определите N_{vi} как число точек xj ≤ v с классом c_i , где x_j - значение точки данных x_j для атрибута X, заданное как

$$N_{vi} = \sum_{j=1}^{n} I(x_j \leq v \text{ and } y_i = c_i) \text{ (19.12)}$$

- Затем мы можем оценить P ($X \leq v | ci$) следующим образом:

$$\hat{P}(X \leq v|c_i) = \frac{\hat{P}(X \leq v \text{ and } c_i)}{\hat{P}(c_i)} = \left(\frac{1}{n}\sum_{j=1}^{n} I\left(x_j \leq v \text{ and } y_j = c_i\right)\right)/(n_i/n) = \frac{N_{vi}}{n_i} \text{(19.13)}$$

- Подставляя уравнения. (19.11) и (19.13) в уравнение (19.10), и используя уравнение (19.8), мы имеем

$$\hat{P}(c_i|\boldsymbol{D}_Y) = \hat{P}(c_i|X \leq v) = \frac{N_{vi}}{\sum_{j=1}^{k} N_{vi}} \text{ (19.14)}$$

- Мы можем оценить $\hat{P}(X > v|c_i)$ следующим образом:

$$\hat{P}(X > v|c_i) = 1 - \hat{P}(X \leq v|c_i) = 1 - \frac{N_{vi}}{n_i} = \frac{n_i - N_{vi}}{n_i} \text{ (19.15)}$$

- Используя уравнения (19.11), (19.15) класс ВМФ $\hat{P}(c_i|\boldsymbol{D}_N)$ рассчитывается:

$$\hat{P}(c_i|\boldsymbol{D}_N) = \hat{P}(c_i|X > v) = \frac{\hat{P}(X > v|c_i)\hat{P}(c_i)}{\sum_{j=1}^{k} \hat{P}(X > v|c_i)\hat{P}(c_j)} = \frac{n_i - N_{vi}}{\sum_{j=1}^{k}(n_j - N_{vi})} \text{(19.16)}$$

Алгоритм

- Алгоритм 19.2 показывает метод оценки точки разделения для числовых атрибутов. Цикл for в строке 4 перебирает все точки и вычисляет значения средних точек v и количество точек N_{vi} из класса c_i таким образом, что xj ≤ v. Цикл for в строке 12 перечисляет все возможные точки разделения вида X ≤ v, по одной для каждой средней точки v, и оценивает их с помощью критерия усиления [Ур. (19.5)]; лучшая точка разделения и оценка записываются и возвращаются. Можно также использовать любые другие оценочные меры. Однако для индекса Джини и CART более низкий балл лучше, в отличие от прироста, где более высокий балл лучше.

Сложность

- С точки зрения вычислительной сложности начальная сортировка значений X (строка 1) занимает время O (n log n). Стоимость вычисления средних точек и специфичных для класса отсчетов N_{vi} занимает время O (nk) (для цикла в строке 4). Стоимость вычисления балла также ограничена O (nk), поскольку общее число средних точек v может быть не более n (для цикла в строке 12). Таким образом, общая стоимость оценки числового атрибута равна O(n log n + nk). Игнорируя k, поскольку это обычно небольшая константа, общая стоимость выведения числовой оценки точки разделения равна O(n log n).

ALGORITHM 19.2. Evaluate Numeric Attribute (Using Gain)

EVALUATE-NUMERIC-ATTRIBUTE $(\mathbf{D}, X)$:

1 sort $\mathbf{D}$ on attribute X, so that $x_j \leq x_{j+1}, \forall j = 1, \ldots, n-1$
2 $\mathcal{M} \leftarrow \emptyset$ // set of midpoints
3 **for** $i = 1, \ldots, k$ **do** $n_i \leftarrow 0$
4 **for** $j = 1, \ldots, n-1$ **do**
5 **if** $y_j = c_i$ **then** $n_i \leftarrow n_i + 1$ // running count for class c_i
6 **if** $x_{j+1} \neq x_j$ **then**
7 $v \leftarrow \frac{x_{j+1} + x_j}{2}$; $\mathcal{M} \leftarrow \mathcal{M} \cup \{v\}$ // midpoints
8 **for** $i = 1, \ldots, k$ **do**
9 $N_{vi} \leftarrow n_i$ // Number of points such that $x_j \leq v$ and $y_j = c_i$
10 **if** $y_n = c_i$ **then** $n_i \leftarrow n_i + 1$
 // evaluate split points of the form $X \leq v$
11 $v^* \leftarrow \emptyset$; $score^* \leftarrow 0$ // initialize best split point
12 **forall** $v \in \mathcal{M}$ **do**
13 **for** $i = 1, \ldots, k$ **do**
14 $\hat{P}(c_i | \mathbf{D}_Y) \leftarrow \frac{N_{vi}}{\sum_{j=1}^{k} N_{vj}}$
15 $\hat{P}(c_i | \mathbf{D}_N) \leftarrow \frac{n_i - N_{vi}}{\sum_{j=1}^{k} n_j - N_{vj}}$
16 $score(X \leq v) \leftarrow Gain(\mathbf{D}, \mathbf{D}_Y, \mathbf{D}_N)$ // use Eq. (19.5)
17 **if** $score(X \leq v) > score^*$ **then**
18 $v^* \leftarrow v$; $score^* \leftarrow score(X \leq v)$
19 **return** $(v^*, score^*)$

Категориальные Атрибуты

- Если X является категориальным атрибутом, то мы оцениваем точки расщепления вида X ∈ V, где V ⊂ dom (X) и V ∅. Другими словами, рассматриваются все различные разбиения множества значений X. Поскольку точка разделения X ∈ V дает то же разбиение, что и X ∈ V, где V = dom (X) \ V является дополнением к V, общее число различных разбиений задается как

$$\sum_{i=1}^{[m/2]} \binom{m}{i} = O(2^{m-1}) \text{ (19.17)},$$

- где m-число значений в области X, то есть m = / dom (X)|. Таким образом, число возможных точек расщепления для рассмотрения экспоненциально в m, что может создавать проблемы, если m велико. Одно из упрощений состоит в том, чтобы ограничить V размером один, так что существует только m расщепленных точек вида Xj ∈ {v}, где v ∈ dom(Xj).

- Чтобы оценить данный раскол точке x ∈ V, мы должны вычислить следующую вероятность массового функций класса:

$$P(c_i|\boldsymbol{D}_Y) = P(c_i|X \in V)$$
$$P(c_i|\boldsymbol{D}_N) = P(c_i|X \notin V)$$

- Используя теорему Байеса, мы имеем

$$P(c_i|X \in V) = \frac{P(X \in V|c_i)P(c_i)}{P(X \in V)} = \frac{P(X \in V|c_i)P(c_i)}{\sum_{j=1}^{k} P(X \in V|c_i)P(c_i)}$$

- Однако заметим, что данная точка x может принимать только одно значение в области X, и поэтому значения v ∈ dom(X) являются взаимоисключающими. Поэтому мы имеем

$$P(X \in V|c_i) = \sum_{v \in V} P(X = v|c_i)$$

- и мы можем переписать $P(c_i|\boldsymbol{D}_Y)$ как

$$P(c_i|\boldsymbol{D}_Y) = \frac{\sum_{v \in V} P(X=v|c_i)P(c_i)}{\sum_{j=1}^{k} \sum_{v \in V} P(X=v|c_j)P(c_i)} \quad (19.18)$$

- Определим n_{vi} как число точек $\mathbf{x}_j \in \mathbf{D}$ со значением $x_j = v$ для атрибута X и классом $y_j = c_i$:

$$n_{vi} = \sum_{j=1}^{n} I(x_j = v \text{ and } y_j = c_i) \quad (19.19)$$

- Класс условных эмпирических ВМФ для X тогда задается как

$$\hat{P}(X = v|c_i) = \frac{\hat{P}(X=v \text{ and } c_i)}{\hat{P}(c_i)} = \left(\frac{1}{n}\sum_{j=1}^{n} I\left(x_j = v \text{ and } y_i = c_i\right)\right) / \left(c_i / n\right) = n_{vi}/n_i \quad (19.20)$$

- Заметим, что априорные вероятности класса могут быть оценены с помощью уравнения (19.11), как обсуждалось для разбиения $\mathbf{D}_Y$ для точки разбиения $X \in V$ задается как

$$\hat{P}(c_i|\boldsymbol{D}_Y) = \frac{\sum_{v \in V} \hat{P}(X=v|c_i)\hat{P}(c_i)}{\sum_{j=1}^{k} \sum_{v \in V} \hat{P}(X=v|c_j)\hat{P}(c_j)} = \frac{\sum_{v \in V} n_{vi}}{\sum_{j=1}^{k} \sum_{v \in V} n_{uj}} \quad (19.21)$$

- Аналогичным образом класс ВМФ для раздела $\mathbf{D}_N$ задается как

$$\hat{P}(c_i|\boldsymbol{D}_N) = \hat{P}(c_i|X \notin V) = \frac{\sum_{v \notin V} n_{vi}}{\sum_{j=1}^{k} \sum_{v \notin V} n_{vj}} \quad (19.22)$$

Алгоритм

- Алгоритм 19.3 показывает метод оценки точки разделения для категориальных атрибутов. Цикл for в строке 4 перебирает все точки и вычисляет n_{vi}, то есть число точек, имеющих значение $v \in dom(X)$ и класс c_i. Цикл for в строке 7 перечисляет все возможные точки разбиения вида $X \in V$ для $V \subset dom\ (X)$, такие что $|V| \leq l$, где l - заданный пользователем параметр, обозначающий максимальную мощность V. Например, чтобы контролировать количество точек разбиения, мы также можем ограничить V одним элементом, то есть $l = 1$, так что разбиения имеют вид $V \in \{v\}$, с $v \in dom(X)$. Если $l = \lfloor m/2 \rfloor$, то мы должны рассмотреть все возможные различные разбиения V. учитывая точку разбиения $X \in V$, метод оценивает ее с помощью информационного усиления [Ур. (19.5)], хотя также можно использовать любой другой критерий оценки. Лучшая точка разделения и оценка записываются и возвращаются.

Сложность

- С точки зрения вычислительной сложности специфичные для класса подсчеты для каждого значения n_{vi} занимают O (n) времени (для цикла в строке 4). При m = |dom(*X*)| максимальное число разбиений V равно $O(2^{m-1})$, и поскольку каждая точка разбиения может быть оценена за время O(mk), цикл for В строке 7 занимает время $O(mk2^{m-1})$. Таким образом, общая стоимость категориальных атрибутов равна $O(n + mk2^{m-1})$. Если мы предположим, что $2^{m-1} = O(n)$, то есть если мы связываем максимальный размер V с l = O(log n), то стоимость категориальных расщеплений ограничена как O (n log n), игнорируя k.

ALGORITHM 19.3. Evaluate Categorical Attribute (Using Gain)

EVALUATE-CATEGORICAL-ATTRIBUTE $(\mathbf{D}, X, l)$:

```
1  for i = 1, ..., k do
2      n_i ← 0
3      forall v ∈ dom(X) do n_vi ← 0
4  for j = 1, ..., n do
5      if x_j = v and y_j = c_i then n_vi ← n_vi + 1 // frequency statistics
   // evaluate split points of the form X ∈ V
6  V* ← ∅; score* ← 0 // initialize best split point
7  forall V ⊂ dom(X), such that 1 ≤ |V| ≤ l do
8      for i = 1, ..., k do
9          P̂(c_i|D_Y) ← (Σ_{v∈V} n_vi) / (Σ_{j=1}^{k} Σ_{v∈V} n_vj)
10         P̂(c_i|D_N) ← (Σ_{v∉V} n_vi) / (Σ_{j=1}^{k} Σ_{v∉V} n_vj)
11     score(X ∈ V) ← Gain(D, D_Y, D_N) // use Eq. (19.5)
12     if score(X ∈ V) > score* then
13         V* ← V; score* ← score(X ∈ V)
14 return (V*, score*)
```

Вычислительная Сложность

- Для анализа вычислительной сложности метода дерева решений в алгоритме 19.1 мы предполагаем, что стоимость оценки всех точек разбиения для числового или категориального атрибута - O (n log n), где n = |**D**| - размер набора данных. Учитывая **D**, алгоритм дерева решений оценивает все атрибуты d со стоимостью (d n log n). Общая стоимость зависит от глубины дерева решений. В худшем случае дерево может иметь глубину n, и таким образом общая стоимость равна O(d n^2 log n).