Энтропия

Терверессы: Лиза, Олеся, Даша, Марина, Яна

БЭК 171

Откуда появилась?

- 1948 год - Клод Шеннон создал первую, истинно математическую, теорию энтропии
- Его идеи послужили основой разработки двух основных направлений: теории информации и теории кодирования

Для дискретных случайных величин

-

Пример. Текст задачи

Свойства

-

Кросс-энтропия и дивергенция Кульбака-Лейблера

-

Пример. Текст задачи

Для непрерывных случайных величин

- Curabitur pellentesque dignissim
- Eu facilisis est tempus quis
- Duis porta consequat lorem
- Eu facilisis est tempus quis

1. Curabitur pellentesque dignissim
2. Eu facilisis est tempus quis
3. Duis porta consequat lorem
4. Curabitur pellentesque dignissim

Пример. Текст задачи

Применение

Энтропийное кодирование
Энтропия показывает наименьшее среднее число бит, необходимое для кодирования некоторой информации. С целью минимизации энтропии и оптимизации кода элементы с большой вероятностью появления кодируются меньшим числом символов. Это позволяет передавать большее количество информации, затрачивая меньший объем памяти.
Построение решающих деревьев
Каждое ветвление дерева представляет собой разделение выборки на две части по порогу некоторого признака. Расчет энтропии помогает определить оптимальный порог для каждого узла — при котором взвешенная сумма энтропий получившихся выборок минимальна среди возможных разбиений.
Например, у нас есть выборка объектов с одним признаком, длина: обыкновенный удав (22 попугая), анаконда (46 попугаев), анаконда (40 попугаев), обыкновенный удав (31 попугай).
Попробуем разделить выборку по 38 попугаям:

Длина $<$ 38 попугаев

Да — Обыкновенный удав, обыкновенный удав

Нет — Анаконда, анаконда

При расчете энтропии $0 \cdot \log_2 0$ считается равным 0, несмотря на $\log_2 0$. За вероятность принимается вероятность встретить данный класс в новой выборке.
Энтропия левой части: $-(1 \cdot \log_2 1 + 0 \cdot \log_2 0) = 0$.
Энтропия правой части: $-(1 \cdot \log_2 1 + 0 \cdot \log_2 0) = 0$.
Суммарная энтропия получилась: $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0 = 0$, $\frac{1}{2}$ — доля каждой выборки в исходной.
Так как 0 — минимально возможное значение энтропии, критерий «длина $<$ 38 попугаев» дает оптимальный результат.

Применение в алгоритме UMAP
В анализе данных алгоритмы снижения размерности используют кросс-энтропию как показатель эффективности перенесения свойств объектов. Чем меньше кросс-энтропия, тем ближе к истинному оказалось подобранное отображение.
Приведем пример работы алгоритма UMAP. Мы возьмем набор данных об одежде, который включает в себя 70000 черно-белых изображений различной одежды по 10 классам: футболки, брюки, свитеры, платья, кроссовки и т.д. Каждая картинка имеет размер 28x28 пикселей или 784 пикселя.
Результатом преобразования будет следующее отображение:

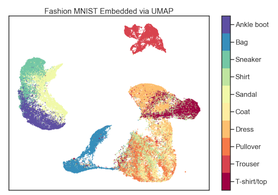

Рис. 1:Алгоритм UMAP

Для оптимального подбора признаков в UMAP используется дивергенция Кульбака-Лейблера для случайной величины Бернулли $X \sim B(p(x))$:

$$D_{KL}(P \,||\, \tilde{P}) = -p(x) \log \tilde{p}(x) - (1-p(x)) \log(1-\tilde{p}(x)) + p$$
$$= p(x) \log \frac{p(x)}{\tilde{p}(x)} + (1-p(x)) \log \frac{1-p(x)}{1-\tilde{p}(x)}$$

Однако алгоритм рассчитывает не просто разницу между двумя распределениями для одной случайной величины, а сумму таких разниц для n случайных величин:

$$S(P||\tilde{P}) = \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{\tilde{p}(x_i)} + (1-p(x_i)) \log \frac{1-p(x_i)}{1-\tilde{p}(x_i)}$$

Данная величина показывает степень отдаленности друг от друга множеств из случайных величин: P и $\tilde{P}$. Минимизация $S(P||\tilde{P})$ по $\tilde{p}(x)$ позволяет найти множество $\tilde{P}$, которое наиболее похоже на множество P.
Задача минимизации заключается в поиске оптимального $\tilde{p}(x)$. Если вернуться к исходной записи $D_{KL} = CE(P||\tilde{P}) - H(p)$, то видно, что энтропия не зависит от $\tilde{p}(x)$, соответственно, является константой при минимизации. Тогда