Использование методов, основанных на кросс-энтропии, для снижения размерности временных рядов

Махнева Лиза

Энтропия $H(X)$

В случае дискретных случайных величин энтропия показывает минимальное среднее число бит для шифрования информации о значениях, которые принимает случайная величина.

$$H(X) = \sum_{i=1}^{n} p_i \cdot \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i \,,$$

где p_i — вероятность того, что случайная величина X примет i-ое значение.

Кросс-энтропия

Это количество информации, в среднем, необходимое для опознания событий из распределения P, используя оптимальную схему для распределения $\tilde{P}$.

$$H_P(\tilde{P}) = -\sum_{i=1}^{n} p_i \cdot \log_2 \tilde{p}_i$$

Дивергенция Кульбака-Лейблера

Данная величина является разностью кросс-энтропии и энтропии. С ее помощью можно определить, какое количество информации мы потратили сверх необходимого из-за того, что не знаем истинное распределение случайной величины. Поэтому она показывает степень отдаленности одного вероятностного распределения P от другого $\tilde{P}$.

$$D_{KL}(P \,||\, \tilde{P}) = -\sum_{i=1}^{n} p_i \log \tilde{p}_i + \sum_{i=1}^{n} p_i \log p_i$$

Применение кросс-энтропии

В алгоритме UMAP используется дивергенция Кульбака-Лейблера для случайной величины Бернулли $X \sim B(p(x))$:

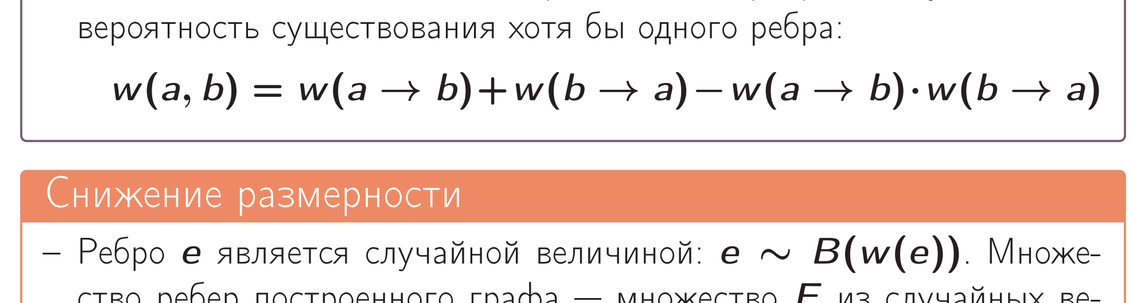
$$p(x) \log \frac{p(x)}{\tilde{p}(x)} + (1 - p(x)) \log \frac{1 - p(x)}{1 - \tilde{p}(x)}$$

Однако алгоритм рассчитывает сумму таких разниц для n случайных величин (для 2 множеств из n случайных величин, S и $\tilde{S}$):

$$\sum_{i=1}^{n} \left(p(x_i) \log \frac{p(x_i)}{\tilde{p}(x_i)} + (1 - p(x_i)) \log \frac{1 - p(x_i)}{1 - \tilde{p}(x_i)} \right)$$

Минимизация $C_S(\tilde{S})$ по $\tilde{p}(x)$ позволяет найти множество $\tilde{S}$, которое наиболее похоже на множество S.

Алгоритм UMAP (Uniform Manifold Approximation and Projection)



Принцип работы алгоритма

Построение графа

- Для каждого объекта из выборки UMAP находит k ближайших соседей, рассчитывает расстояние ρ до ближайшего, а также нормирующую величину σ
- Затем UMAP строит ориентированный взвешенный граф: ребрами соединяются каждый объект с его соседями. Вес ребра из объекта x_i к его соседу t_j определяется по формуле:

$$w(x_i \to t_j) = \exp\left(-\frac{d(x_i, t_j) - \rho_i}{\sigma_i}\right)$$

- Если интерпретировать вес ребра из a в b как вероятность его существования, то мы можем определить вес ребра между a и b как вероятность существования хотя бы одного ребра:

$$w(a, b) = w(a \to b) + w(b \to a) - w(a \to b) \cdot w(b \to a)$$

Снижение размерности

- Ребро e является случайной величиной: $e \sim B(w(e))$. Множество ребер построенного графа — множество E из случайных величин Бернулли
- Чтобы перенести граф в низкоразмерное пространство, UMAP подбирает для множества E_h похожее на него множество E_l с функцией $w_l(e)$, соответствующие низкоразмерному пространству
- UMAP решает задачу минимизации кросс-энтропии:

$$-\sum_{e \in E} w_h(e) \log w_l(e) + (1 - w_h(e)) \log(1 - w_l(e)) \to \min_{w_l}$$

Результатом является граф в низкоразмерном пространстве с подобранной функцией весов w_l.

Реализация UMAP

Применим алгоритм к данным о ценах криптовалют на протяжении 669 дней (см. рис 2). Рассмотрим сформировавшийся кластер №1. Временные ряды, оказавшиеся в кластере, ведут себя наиболее похоже во времени. Они сильно коррелируют между собой:

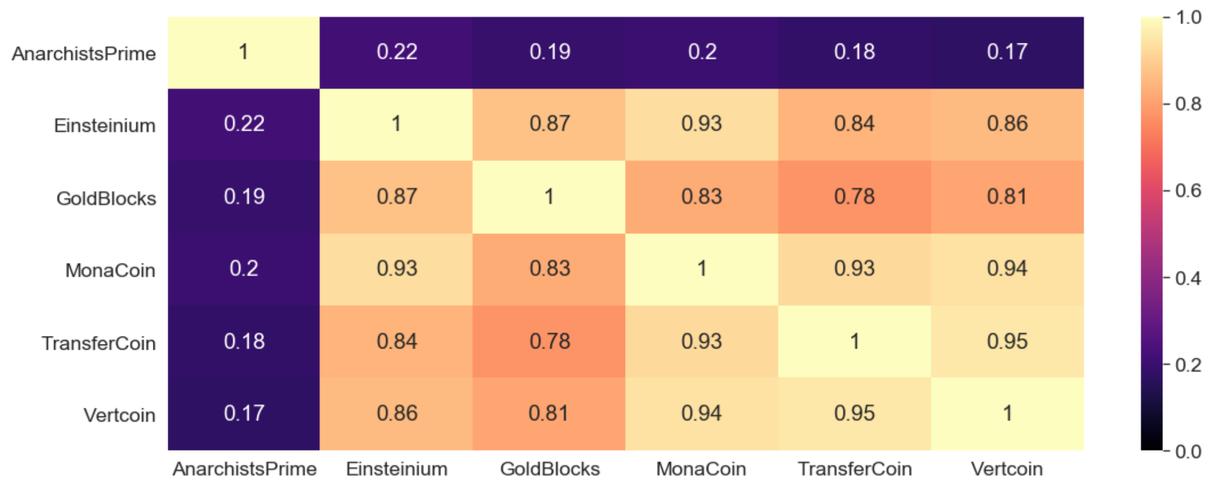


Рис. 1. Коэффициенты корреляции между рядами кластера №1

Однако AnarchistsPrime выбивается из группы криптовалют — у нее самые низкие коэффициенты корреляции. Если посчитать корреляцию AnarchistsPrime со всеми временными рядами из выборки, получается, что наибольшее значение равно $\mathbf{0.22}$. То есть криптовалюта оказалась в данном кластере, так как в нем наиболее близкие к ней объекты среди имеющихся.

Значит, UMAP и коэффициент корреляции по-разному оценивают схожесть рядов. Алгоритм считает похожими те, между которыми оказалось наименьшее «расстояние», даже если это расстояние велико — просто в представленной выборке не оказалось объектов ближе.

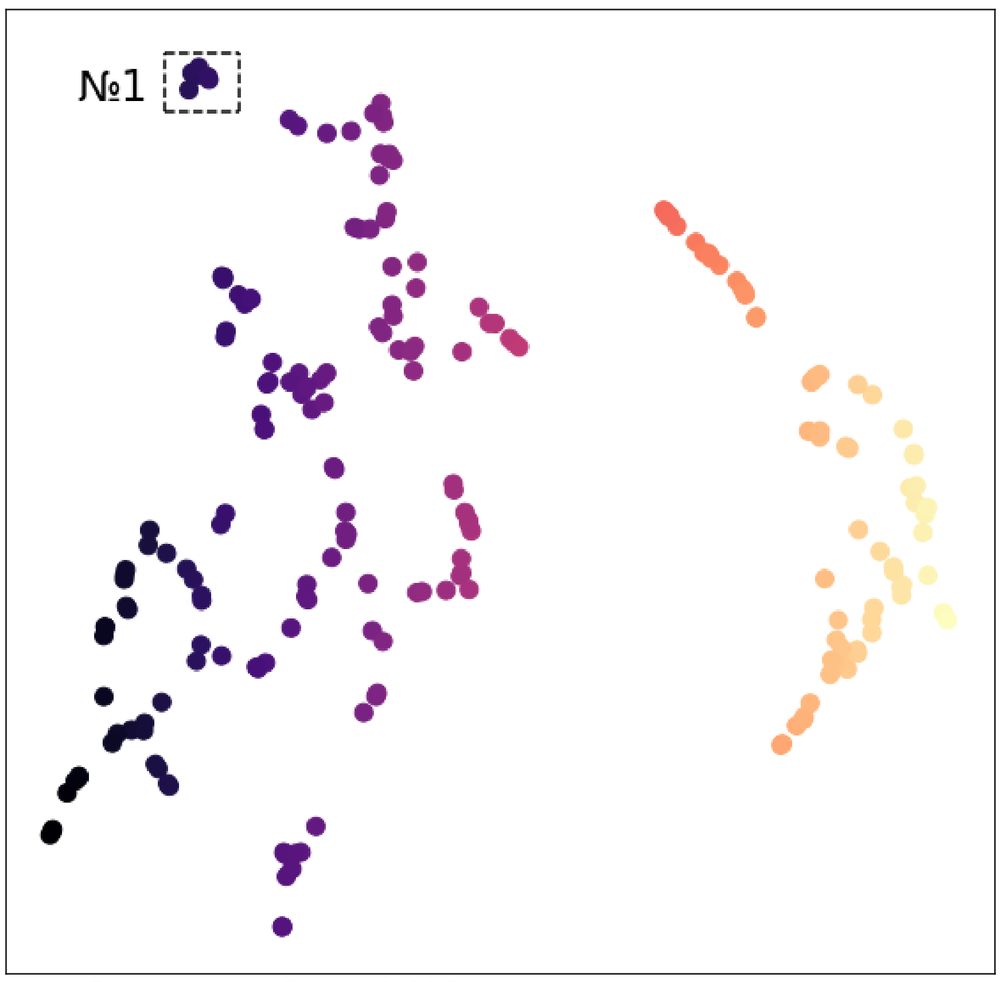


Рис. 2. Результат работы UMAP с временными рядами криптовалют