Предобработка характеристик.

Непрерывные характеристики

Нормируем.

Дискретизация характеристик

Биним на группы.
Например, хорошо подходит для характеристики **возраст**.

Категориальные характеристики

Либо создаем **one-hot encoding** вектор, состоящий из нулей и единицы:
(Dog, Cat) => [(0, 1), (1, 0), (0, 1)]
Либо каждой категории ставим в соответствие вектор $\in R^n$,который является обучаемым -- **embedding**
В PyTorch это **nn.Embedding([кол-во категорий], [размер вектора])**

Категориальные характеристики

Когда категорий слишком (!) много, то можно использовать **хэширование**:
Для каждой категории вычисляется свое хэш значение и выбирается соответствующий бакет, то есть некое число.
И дальше уже каждому числу (бакету) ставить в соответствие эмбеддинг.
Количество бакетов задаем самостоятельно.

Циклические характеристики

Такие характеристики, как день недели, месяц, час и т.п. неверно ставить в соответствие вектор и делать их категориальными.
Лучше даже оставить непрерывными.

Циклические характеристики

Но у таких характеристик есть особенность - они циклические. И чтобы отобразить эту закономерность и показать, что воскресенье ближе к понедельнику, а среда к четвергу, то кодировать их нужно cos и sin:

$$x_{sin} = sin(\frac{2\pi x}{max(x)})$$

$$x_{cos} = cos(\frac{2\pi x}{max(x)})$$

Циклические характеристики

Например, время суток - 4 часа:

$$cos(\frac{2\pi}{24} * 4); sin(\frac{2\pi}{24} * 4)$$