

MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

ЛЕКЦИЯ 5

ОСНОВНЫЕ ТИПЫ ПРИЗНАКОВ В ДАТАСЕТАХ

1. Числовые (например температура)
2. Числовые ранговые (номер этажа)
2. Бинарные (пол)
3. Категориальные (цвет)
4. Другие (текст)

ПОТЕРЯННЫЕ (ПРОПУЩЕННЫЕ) ЗНАЧЕНИЯ (MISSED VALUES)

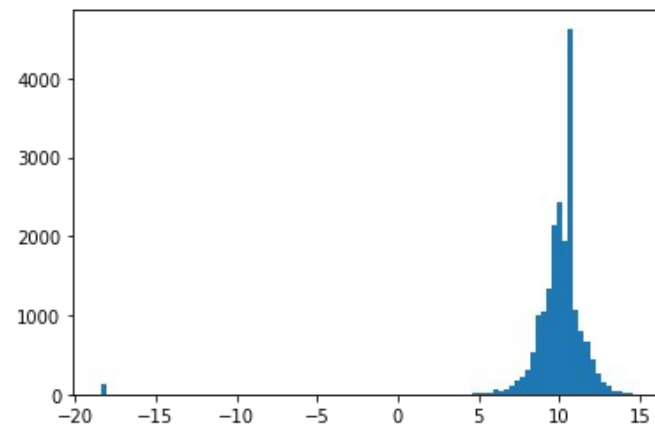
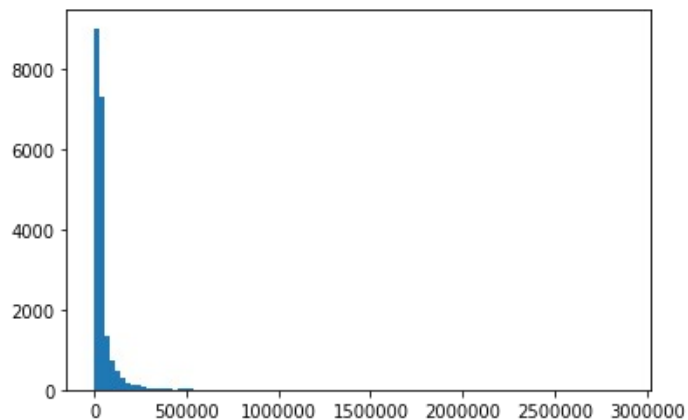
Способы избавления от потерянных значений:

1. Удаление строк с пропусками
2. Замена пропусков средним значением
3. Замена пропусков медианным значением
4. Замена пропусков другим числом (в зависимости от природы признака) или в некоторых случаях случайным числом в заданном диапазоне

Статистические выбросы можно также рассматривать как пропуски

РАСПРЕДЕЛЕНИЕ ЧИСЛОВЫХ ПРИЗНАКОВ

Большинство алгоритмов ML хорошо работает, если значения признаков распределены нормально



ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПРИЗНАКОВ

Основные способы кодирования категориальных признаков

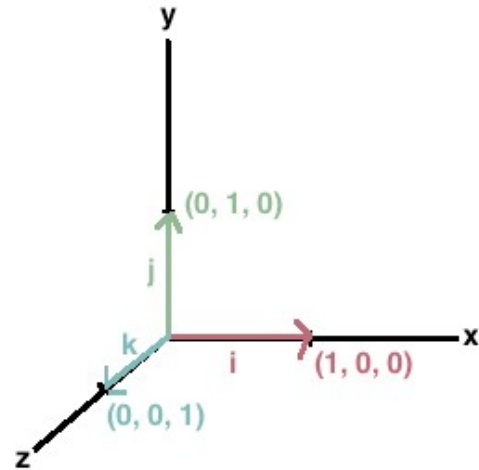
1. One-Hot Encoding – превращение в ортогональные вектора
2. Label Encoding
3. Binary Encoding
4. Замена числовыми значениями (экзотические случаи)

ONE-HOT ENCODING

id	color
1	red
2	blue
3	green
4	blue

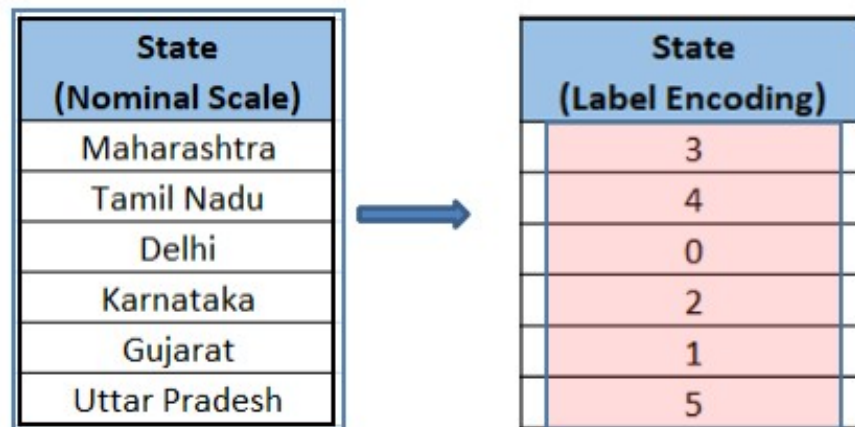


id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0



LABEL ENCODING

Каждому значению признака присваивается
уникальная числовая метка



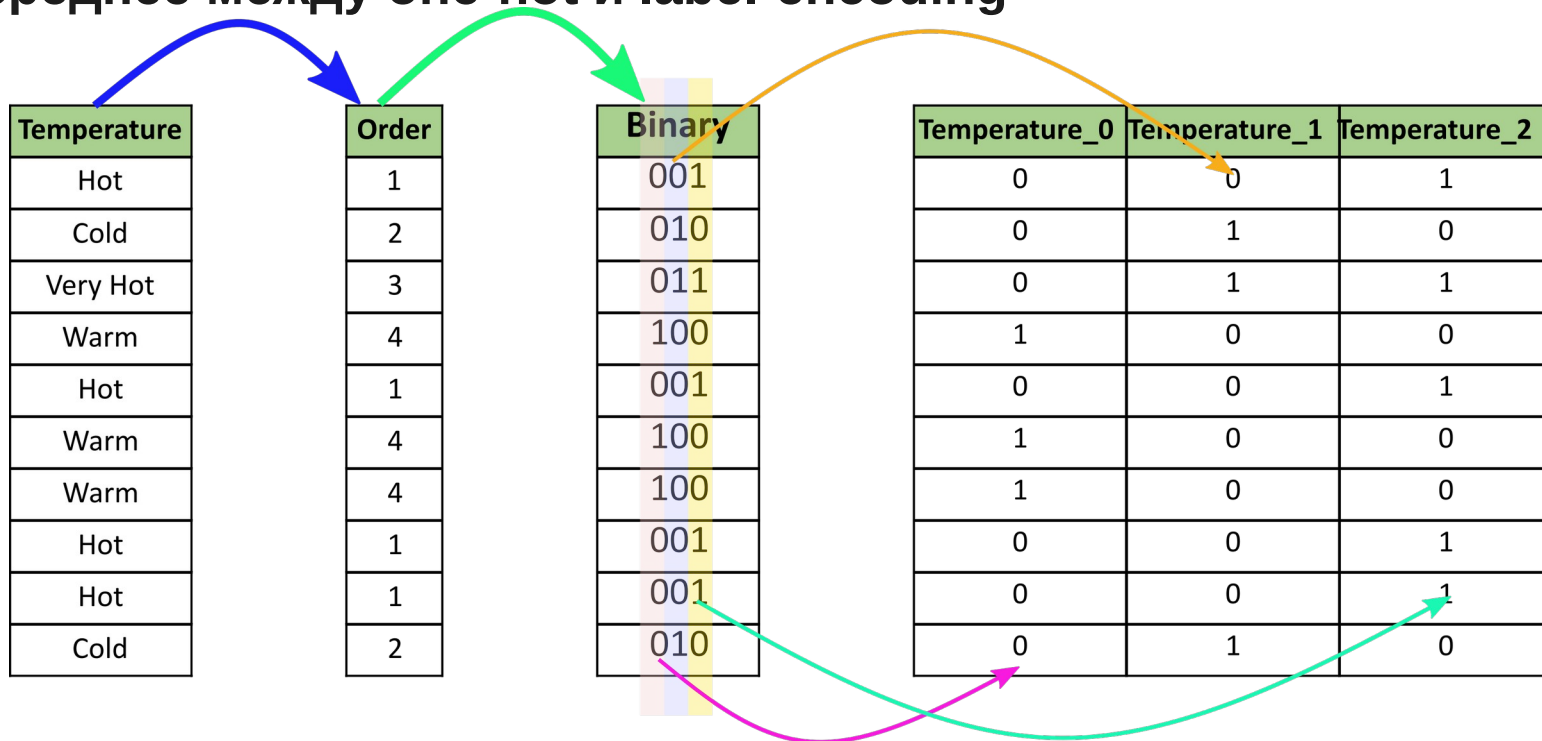
The diagram illustrates the process of label encoding. On the left, a table titled 'State (Nominal Scale)' lists six Indian states. A blue arrow points to the right, where a second table titled 'State (Label Encoding)' shows the same states mapped to unique numerical values. The rows in the second table are highlighted in pink.

State (Nominal Scale)	
Maharashtra	
Tamil Nadu	
Delhi	
Karnataka	
Gujarat	
Uttar Pradesh	

State (Label Encoding)	
	3
	4
	0
	2
	1
	5

BINARY ENCODING

Нечто среднее между one-hot и label-encoding



ЧТО ДЕЛАТЬ ЕСЛИ ЧИСЛО КАТЕГОРИЙ ОГРОМНО?

1. Выбрать TOP-n (например TOP-10) наиболее популярных категорий, а остальные заменить значением “Other”

Далее использовать Label Encoding или One-hot

2. Сделать ONE-HOT или BINARY ENCODING, а затем уменьшить размерность данных методом PCA

3. Если есть возможность, то заменить числовыми значениями

МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PCA)

PRINCIPAL COMPONENT ANALYSIS

$$X = (X_1, X_2, \dots, X_m)^T$$

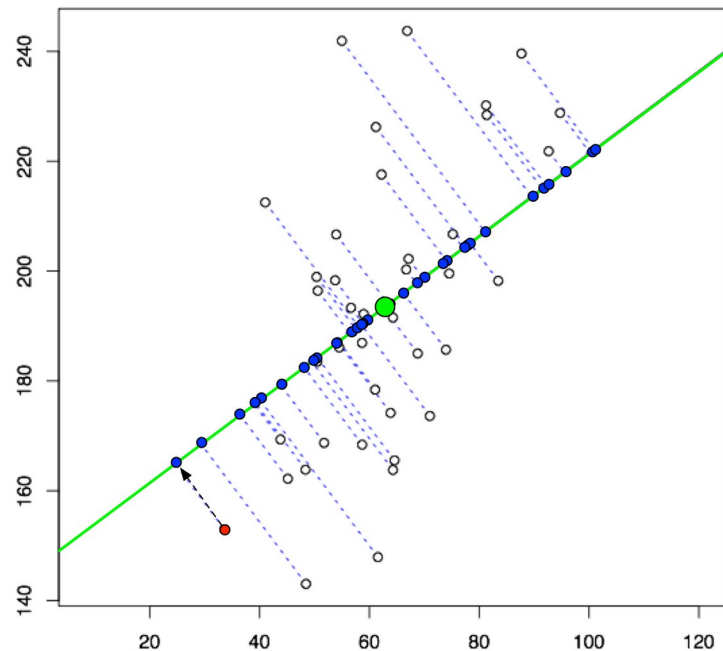
$$\text{cov} = \frac{1}{m} \sum_{i=1}^m (x^{(i)})(x^{(i)})^T$$

$$\text{SVD}(\text{cov}) = U \Sigma W^T$$

$$U \rightarrow [n \times n] \quad U_{\text{reduce}} \rightarrow [n \times k]$$

$$\bar{X} = XU_{\text{reduce}}$$

$$X_{\text{app}} = \bar{X} U_{\text{reduce}}^T$$



МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PCA)

ВЫБОР k ГЛАВНЫХ КОМПОНЕНТ

$$SVD(cov) = U \Sigma W^T$$

$$\frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{app}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq 0.01$$

$$\Sigma = \text{diag}(S_1, S_2, \dots, S_n)$$

$$\frac{\sum_{i=1}^k S_i}{\sum_{i=1}^n S_i} \geq 0.99$$

