Конспект по теме «Взаимодействия между переменными»

Ковариация

Диаграмма рассеивания — это график, который позволяет визуализировать взаимоотношение двух числовых величин.

Ковариация — это мера совместной изменчивости двух величин. Она бывает положительной, отрицательной и равной нулю.

Коэффициент ковариации рассчитывают по формуле:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Ковариация показывает только направление взаимосвязи, но не её силу.

Корреляция

Корреляция — это мера силы и направления линейной взаимосвязи между двумя величинами. Она принимает значения от -1 до 1. Чем ближе абсолютное значение коэффициента корреляции к 1, тем сильнее связь.

Коэффициент корреляции Пирсона рассчитывают по формуле:

$$r_{X,Y} = \frac{\mathrm{C}ov(X,Y)}{s_X \cdot s_Y}$$

Чтобы корректно применять коэффициент корреляции Пирсона, надо убедиться, что в данных нет выбросов и величины имеют линейную зависимость. В этом помогает диаграмма рассеяния.

Сравнение категорий

Чтобы оценить, как категориальная переменная взаимосвязана с числовой:

- можно использовать столбчатую диаграмму со средними значениями по категориям
- или ящик с усами с более полной информацией по распределениям величин.

	Столбчатая диаграмма	Ящик с усами
Наглядность	Проста для восприятия	Нужно научиться читать график
Сколько показателей помогает исследовать один график	Один (например, среднее)	Много (медиану, размах, характер распределения, наличие выбросов)
Удобна ли при близких значениях показателей	Нет	Да
Какой анализ помогает сделать	Краткий	Детальный
В каких случаях полезна	Когда достаточно сравнить категории по одному показателю	Когда нужно провести подробный анализ и принять обоснованное решение

Бинаризация — процесс преобразования числовой переменной в категориальную путём разделения её на интервалы.

Способ бинаризации	Ситуации, в которых этот способ эффективен
Вручную	Данные позволяют легко определить интервалы на основе здравого смысла или экспертных знаний.
Интервалы одинаковой длины	Данные равномерно распределены и нет особого значения, насколько точно определены границы интервалов
Интервалы с одинаковым количеством наблюдений	Важно, чтобы интервалы содержали примерно одинаковое количество точек, например, для сбалансированного сравнения групп.
Математические методы, методы машинного обучения	Сложные данные с неявными зависимостями или важно определить оптимальное количество интервалов.