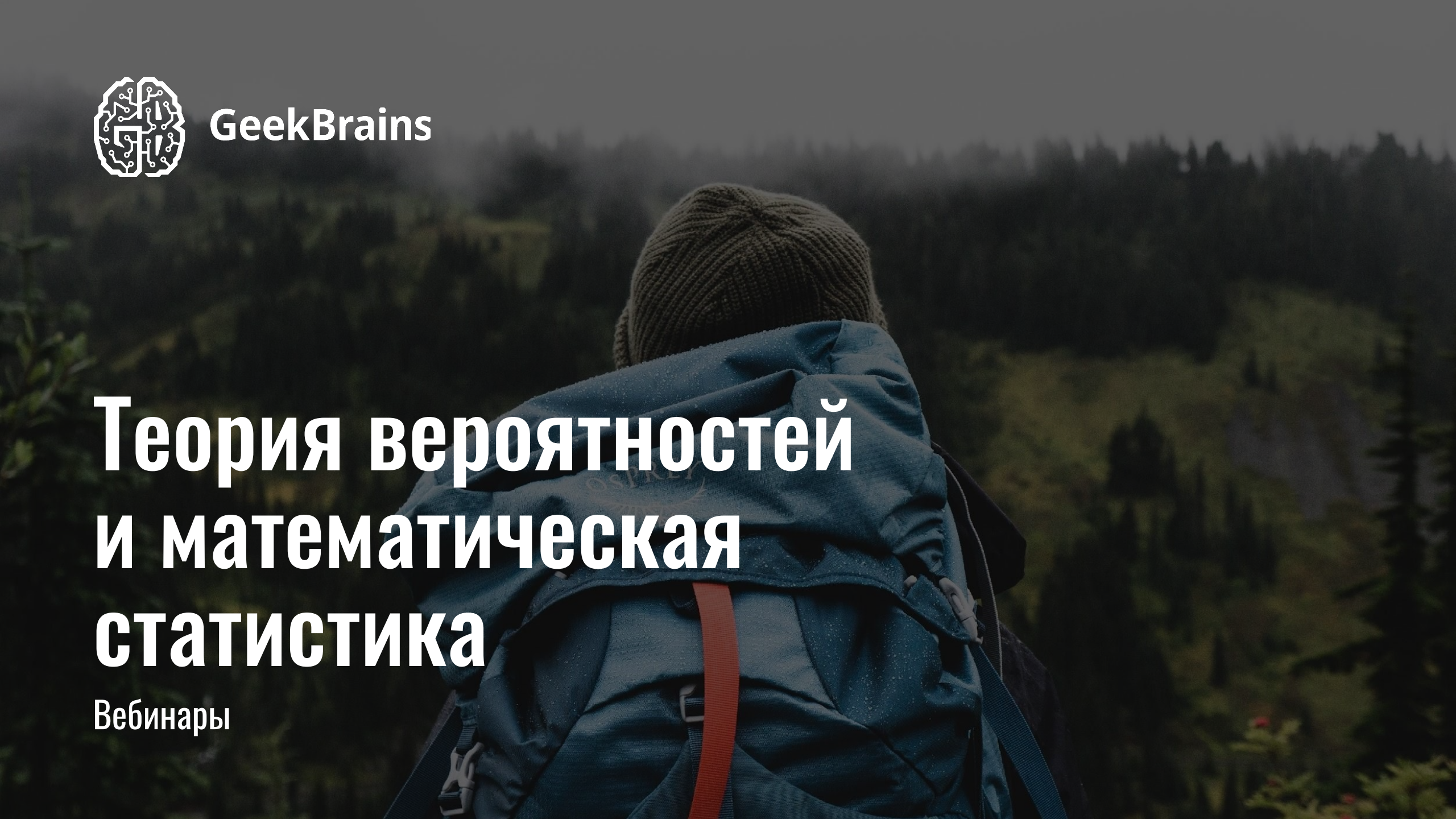
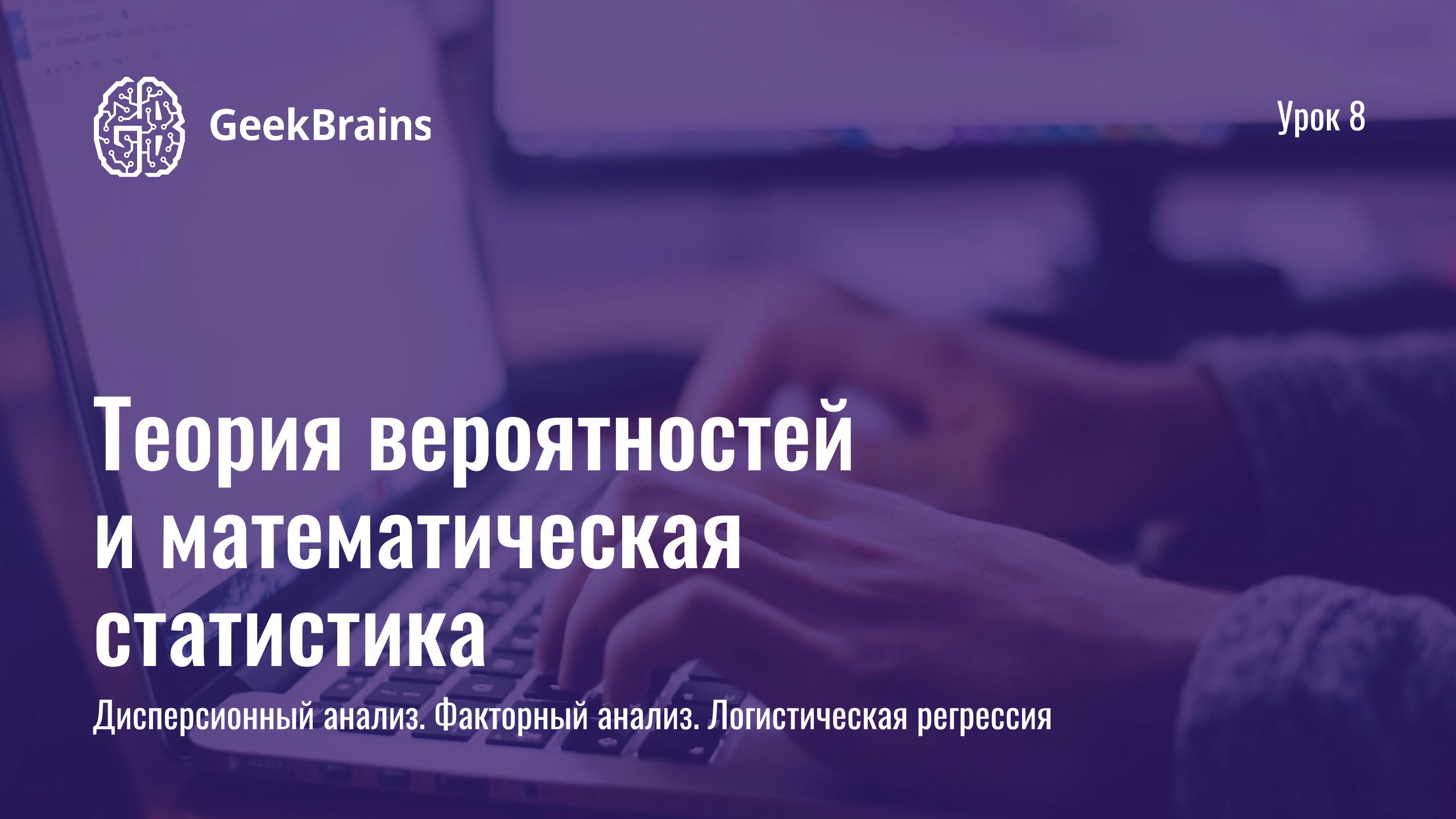


GeekBrains

Теория вероятностей и математическая статистика

Вебинары



GeekBrains

Урок 8

Теория вероятностей и математическая статистика

Дисперсионный анализ. Факторный анализ. Логистическая регрессия

На этом уроке мы изучим

1. Однофакторный и двухфакторный дисперсионный анализ.
2. Факторный анализ.
3. Логистическая регрессия.

Дисперсионный анализ

Дисперсионный анализ

Дисперсионный анализ — метод в математической статистике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях.

В дисперсионном анализе изучается влияние одного или нескольких факторов на зависимую переменную, причём факторы являются *номинативными* (*категориальными*), а целевая переменная является *абсолютной* (*количественной*).

Однофакторный дисперсионный анализ

В *однофакторном дисперсионном анализе* на одну переменную Y влияет один фактор, наблюдаемый на k *уровнях*, т.е. имеем k выборок для переменной Y. Проверяется гипотеза H_0 о равенстве средних значений по каждой выборке:

$$H_0 : \overline{y_1} = \cdots = \overline{y_k}$$

Здесь y_i — i-я выборка.

Двухфакторный дисперсионный анализ

В *двухфакторном дисперсионном анализе* на одну переменную Y влияют два фактора A, B, каждый из которых является категориальным. Проверяются гипотезы о влиянии каждого фактора на значение Y. Отличие теперь в том, что влияния факторов на значение Y могут "пересекаться", и это нужно учитывать.

Факторный анализ

Факторный анализ

Факторный анализ — это способ приведения множества непосредственно наблюдаемых факторов x_j, $j = 1, \ldots, m$, к меньшему числу новых линейно независимых факторов y_j, $j = 1, \ldots, q$, $q < m$.

Метод главных компонент

Метод главных компонент заключается в вычислении собственных значений и собственных векторов ковариационной матрицы.

Далее отбираются собственные векторы, соответствующие наибольшим собственным значениям, и эти векторы используются для синтеза новых признаков.

Логистическая регрессия

Логистическая регрессия

Ранее мы познакомились с моделью линейной регрессии:

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon$$

Такая модель используется в задачах *регрессии*, т.е. когда нам нужно в результате получить какое-то число.

Логистическая регрессия

Логистическая регрессия применяется в задачах *бинарной классификации*, когда нам нужно получить на выходе метку класса: *1* или *-1* (иногда вместо *-1* используют *0*).

Логистическая регрессия представляет собой модель линейной регрессии, *поверх* которой используется *логистическая функция* (или *сигмоида*):

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Логистическая регрессия

Сигмоида принимает в качестве аргумента вещественное число, а отдаёт число из промежутка $[0, 1]$.

Такая модель на вход получает значения факторов, а на выходе отдаёт число из промежутка $[0, 1]$, которое можно интерпретировать как вероятность объекта принадлежать классу 1.

Для нахождения оптимальных параметров модели используют *градиентный спуск*.

Итого

1. Однофакторный и двухфакторный дисперсионный анализ.
2. Факторный анализ.
3. Логистическая регрессия.