PCA. Breast Cancer

- Цель исследования: классификация опухолей молочной железы на злокачественные и доброкачественные
- Используемые методы: PCA, Logistic Regression

Кац Софья 5030102/10401

Исследовательский анализ данных

- Данные представляют собой таблицу с ID исследования, диагнозом (M - злокачественный, B - доброкачественный) и некоторыми числовыми значениями признаков (радиус, область, периметр и т.д.).
- Характеристики вычисляются по оцифрованному изображению тонкоигольного аспирата (FNA) опухоли молочной железы. Они описывают характеристики клеточных ядер, присутствующих на изображении.

Предварительная обработка данных

- Удалены лишние столбцы (“id”, пустой столбец “Unnamed: 32”)
- Значения диагноза заменены на числовые (M → 1, B → 0)
- Проверено наличие пустых значений (их не оказалось)
- Данные разделены на признаки и целевую функцию и нормализованы

PCA (Principal component analysis)

- Основные шаги алгоритма:
  - Преобразование датасета в матрицу
  - Регулировка значений в каждом столбце матрицы так, чтобы они были сосредоточены вокруг среднего значения
  - Создание корреляционной матрицы между параметрами, затем нахождение собственных векторов и собственных значений этой матрицы
  - Получение на выходе набора новых координат, в итоговой матрице указаны только «существенные» координаты

Logistic Regression (minimization)

- Метод классификации, основанный на логистической функции
- Цель: минимизировать функцию потерь

The logistic function

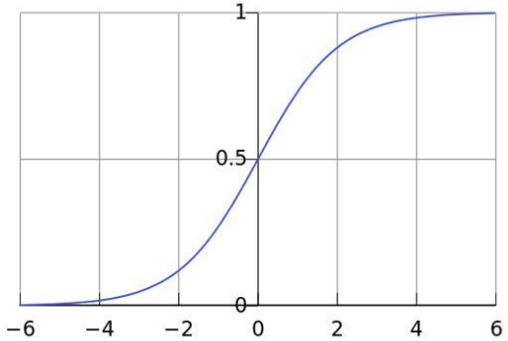
$$f(t) = \frac{1}{1 + e^{-t}}$$

$$P(C_+|x) = \frac{1}{1 + e^{-w \cdot x}}$$

$$P(C_-|x) = \frac{e^{-w \cdot x}}{1 + e^{-w \cdot x}}$$

$$\log \frac{P(C_+|x)}{P(C_-|x)} = w \cdot x$$

Linear regression on the log-odds ratio



Logistic Regression: Find the vector w that maximizes the probability of the observed data