# Ordinary least squares method

Marcin Kuta

# Breast cancer dataset

```
1) ID number
2) Diagnosis (M = malignant, B = benign)

3)  radius (mean)
4)  texture (mean)
5)  perimeter (mean)
6)  area (mean)
7)  smoothness (mean)
8)  compactness (mean)
9)  concavity (mean)
10) concave points (mean)
11) symmetry (mean)
12) fractal dimension (mean)
```

# Breast cancer dataset

```
13) radius (stderr)
14) texture (stderr)
15) perimeter (stderr)
16) area (stderr)
17) smoothness (stderr)
18) compactness (stderr)
19) concavity (stderr)
20) concave points (stderr)
21) symmetry (stderr)
22) fractal dimension (stderr)
```

# Breast cancer dataset

```
23) radius (worst)
24) texture (worst)
25) perimeter (worst)
26) area (worst)
27) smoothness (worst)
28) compactness (worst)
29) concavity (worst)
30) concave points (worst)
31) symmetry (worst)
32) fractal dimension (worst)
```

# Least squares method

Linear representation

$$A_{\text{lin}} = \begin{bmatrix} f_{1,1} & \dots & f_{1,m} \\ f_{2,1} & \dots & f_{2,m} \\ \vdots & \dots & \vdots \\ f_{n,1} & \dots & f_{n,m} \end{bmatrix} \tag{1}$$

# Linear system of equations

$$Ax = y, \quad A \in \mathbb{R}^{n\times m}, \quad y \in \mathbb{R}^{n\times 1} \tag{2}$$

System of equations is:

- underdetermined: $n < m$
- overdetermined: $n > m$

Number of solutions:

- Infinitely many solutions:
  - $\text{rank}([A, y]) = \text{rank}(A)$ and $\text{rank}(A) < m$
- Exactly one solution:
  - $\text{rank}([A, y]) = \text{rank}(A)$ and $\text{rank}(A) = m$
- No solution:
  - $\text{rank}([A, y]) = \text{rank}(A) + 1$

# Least squares method

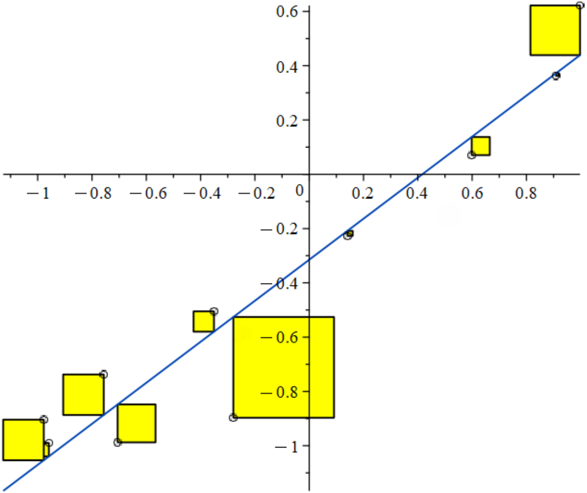
$$Aw \cong y, \quad A \in \mathbb{R}^{n \times m}, \quad y \in \mathbb{R}^{n \times 1} \tag{3}$$

$n$ – number of instances (number of equations)
$m$ – number of features (number of searched weights)

$A \in \mathbb{R}^{n \times m}$ is a known matrix of features
$y \in \mathbb{R}^{n \times 1}$ is a known column vectors of labels
$w \in \mathbb{R}^{m \times 1}$ is a searched vector of feature weights

$$\min_{w} ||Aw - y||_2 \tag{4}$$

# Ordinary least squares method



# Least squares method

$$\min_{w} ||Aw - y||_2 = \min_{w} ||Aw - y||_2^2 = \min_{w} J(w)$$

$$\begin{aligned}
J(w) = ||Aw - y||_2^2 &= (Aw - y)^T(Aw - y) \\
&= (Aw)^T(Aw) - (Aw)^T y - y^T(Aw) + y^T y \\
&= (Aw)^T(Aw) - 2(Aw)^T y + y^T y \\
&= w^T A^T A w - 2w^T A^T y + y^T y
\end{aligned}$$

$$\frac{\partial J}{\partial w} = 2A^T A w - 2A^T y = 0$$

Normal equations

$$A^T A w = A^T y \tag{5}$$

# Least squares method

Linear representation

$$A_{\text{lin}} = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & f_{1,4} \\ f_{2,1} & f_{2,2} & f_{2,3} & f_{2,4} \\ \vdots & \vdots & \vdots & \vdots \\ f_{n,1} & f_{n,2} & f_{n,3} & f_{n,4} \end{bmatrix} \tag{6}$$

# Least squares method

Quadratic representation

$$A_{\text{quad}} = \begin{bmatrix} f_{1,1}, f_{1,2}, f_{1,3}, f_{1,4}, f_{1,1}^2, f_{1,2}^2, f_{1,3}^2, f_{1,4}^2, f_{1,1}f_{1,2}, f_{1,1}f_{1,3}, f_{1,1}f_{1,4}, f_{1,2}f_{1,3}, f_{1,2}f_{1,4}, f_{1,3}f_{1,4} \\ \vdots \\ f_{n,0}, f_{n,1}, f_{n,2}, f_{n,3}, f_{n,0}^2, f_{n,1}^2, f_{n,3}^2, f_{n,4}^2, f_{n,1}f_{n,2}, f_{n,1}f_{n,3}, f_{n,1}f_{n,4}, f_{n,2}f_{n,3}, f_{n,2}f_{n,4}, f_{n,3}f_{n,4} \end{bmatrix}$$

- features $f_{k,i}$
- quadratic features $f_{k,i}^2$
- interaction terms $f_{k,i}f_{k,j}$

# Normal equations

$$w = \underbrace{(A^T A)^{-1} A^T}_{\substack{\text{pseudoinverse} \\ \text{matrix}}} y \tag{7}$$

$$w = A^{\dagger} y \tag{8}$$

$$\min_w J(w) = \min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2 \tag{9}$$

Equations (7) and (8) have only theoretical importance. In practice, explicit matrix inverse should be avoided and normal equations are used to find weights $w$:

$$A^T A w = A^T y \tag{10}$$

# Conditioning of least squares

Condition number of matrix $A$:

$$\kappa(A) = \text{cond}(A) \stackrel{\text{df}}{=} ||A|| \cdot ||A^{-1}|| \tag{11}$$

It can be found with `np.linalg.cond`.

$$\text{cond}(A^T A) = \text{cond}(A)^2 \tag{12}$$

- Conditionining of a square linear system of equations $Aw = y$ depends only on $A$.
- Conditionining of a least squares problem $Aw \cong y$ depends both on $A$ and $y$.
- Ill-conditioning does not harm predictions – residual of normal equations will be small
- The values of weights will be poorly determined

# Normal equations

- $A^T A$ is not guaranteed to be non-singular
- The method is overly sensitive to the condition number of matrix
- QR and SVD are numerically more stable alternatives but computationally slower

| | |
|---|---|
| A is well-conditioned | normal equations |
| A is not well-conditioned but is not rank deficient | QR |
| A is rank deficient | SVD |

# Normal equations

$$w = \underbrace{(A^T A + \lambda I)^{-1} A^T}_{\substack{\text{pseudoinverse} \\ \text{matrix}}} y \tag{13}$$

$$\min_w J(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda w^T w \tag{14}$$

Equation (13) has only theoretical importance. In practice, explicit matrix inverse should be avoided and normal equations are used to find weights $w$:

$$(A^T A + \lambda I) w = A^T y \tag{15}$$

# Beyond least squares method

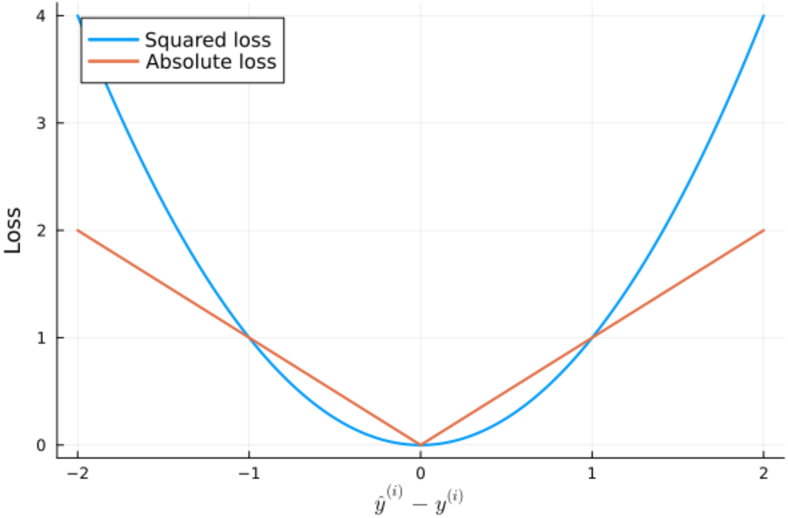
$$\min_w ||Aw - y||_2$$

- Cost function $J(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2$
- Differences $y_i - w^T x_i$ have Gaussian distribution
- There is explicit formula for $w$

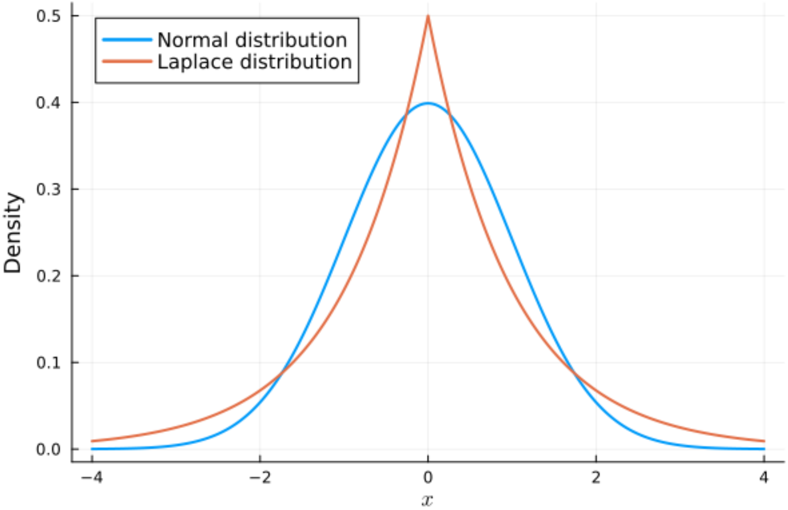
$$\min_w ||Aw - y||_1$$

- Cost function $J(w) = \sum_{i=1}^{n} |y_i - w^T x_i|$
- Differences $y_i - w^T x_i$ have Laplace distribution
- There is no explicit formula for $w$

# Cost functions



# Distributions



# References

[1] Michael T. Heath,
Scientific Computing. An Introductory Survey, 2nd Edition,
Chapter 3: Linear Least Squares
2002

[2] Michael T. Heath,
Chapter 3: Linear Least Squares
http://heath.cs.illinois.edu/scicomp/notes/cs450_chapt03.pdf

[3] Introduction to pandas
https://jakevdp.github.io/PythonDataScienceHandbook/03.00-introduction-to-pandas.html