Machine Learning

Linear Regression and its variants. ML and MAP principles. Regression Quality Metrics.

Aleksandr Petiushko

ML Research

February 19th, 2024

SOFIA UNIVERSITY

Content

Content

Content

1. Linear Regression formulation
2. ML and MAP principles
3. Least Squares method

Content

Content

Content

1. Linear Regression formulation
2. ML and MAP principles
3. Least Squares method
4. Polynomial Regression
5. Ridge Regression, LASSO and Elastic Net
6. Quality metrics for Regression

Main math concepts: a reminder

- Bayes' rule: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$

Main math concepts: a reminder

- Bayes' rule: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
- PDF of Normal distribution $x \sim N(\mu, \sigma^2)$: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Main math concepts: a reminder

- Bayes' rule: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
- PDF of Normal distribution $x \sim N(\mu, \sigma^2)$: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Square norm (common, euclidean, L_2 — usually w/o saying explicitly) of a vector $w = (w_1, \ldots, w_n)$: $||w||^2 = \sum_{i=1}^{n} (w_i)^2$

Main math concepts: a reminder

- Bayes' rule: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$
- PDF of Normal distribution $x \sim N(\mu, \sigma^2)$: $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Square norm (common, euclidean, L_2 — usually w/o saying explicitly) of a vector $w = (w_1, \ldots, w_n)$: $||w||^2 = \sum_{i=1}^{n} (w_i)^2$
- L_1-norm of a vector $w = (w_1, \ldots, w_n)$: $|w| = \sum_{i=1}^{n} |w_i|$

Problem Statement: Linear Regression

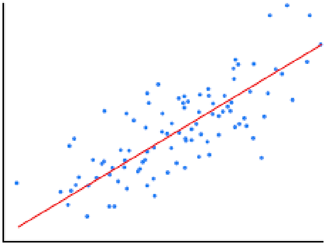
Given

$$y_i = w^T x_i + \varepsilon_i \Rightarrow y_i \sim N(w^T x_i, \sigma^2),$$

for $i = 1, \ldots, m$, where $w \in \mathbf{R}^{n+1}$, $\varepsilon_i \sim N(0, \sigma^2)$, $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ — the training dataset.

Task

Find w



Two kinds of parameter estimation

Maximum Likelihood (**ML**) Principle

$$w_{ML} = \arg\max_{w} \; p(y|w, x)$$

Two kinds of parameter estimation

Maximum Likelihood (**ML**) Principle

$$w_{ML} = \arg\max_{w}\ p(y|w, x)$$

Principle of Maximum A Posterior (**MAP**) Probability

$$w_{MAP} = \arg\max_{w}\ p(w|x, y)$$

Maximum likelihood estimator

$$w_{ML} = \arg\max_{w} p(y|w, x)$$

Maximum likelihood estimator

$$w_{ML} = \arg\max_{w} p(y|w, x)$$

$$w_{ML} = \arg\max_{w} \prod_{i} p(y_i|w, x_i)$$

Maximum likelihood estimator

$$w_{ML} = \arg\max_w p(y|w,x)$$

$$w_{ML} = \arg\max_w \prod_i p(y_i|w,x_i)$$

$$p(y_i|w,x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

Maximum likelihood estimator

$$w_{ML} = \arg\max_w p(y|w,x)$$

$$w_{ML} = \arg\max_w \prod_i p(y_i|w,x_i)$$

$$p(y_i|w,x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

$$w_{ML} = \arg\max_w \prod_i \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) = \arg\max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2}$$

Maximum likelihood estimator

$$w_{ML} = \arg\max_w p(y|w, x)$$

$$w_{ML} = \arg\max_w \prod_i p(y_i|w, x_i)$$

$$p(y_i|w, x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

$$w_{ML} = \arg\max_w \prod_i \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) = \arg\max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2}$$

$$w_{ML} = \arg\min_w \sum_i (y_i - w^T x_i)^2$$

Least squares method

Problem statement and assumptions

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$

Least squares method

Problem statement and assumptions

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$
- $a(x) = f_w(x) = w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_n x^n$, where $w = (w_0, w_1, \ldots, w_n)^T \in \mathbb{R}^{n+1}$ — model parameters.

Least squares method

Problem statement and assumptions

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$
- $a(x) = f_w(x) = w_0 + w_1 x^1 + w_2 x^2 + \cdots + w_n x^n$, where $w = (w_0, w_1, \ldots, w_n)^T \in \mathbb{R}^{n+1}$ — model parameters.
- It is convenient to write in vector form

$$a(x) = w^T \cdot x,$$

where $x = (1, x^1, \ldots, x^n)^T \in \mathbb{R}^{n+1}$.

Least squares method

Problem statement and assumptions

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$
- $a(x) = f_w(x) = w_0 + w_1 x^1 + w_2 x^2 + \dots + w_n x^n$, where $w = (w_0, w_1, \dots, w_n)^T \in \mathbb{R}^{n+1}$ — model parameters.
- It is convenient to write in vector form

$$a(x) = w^T \cdot x,$$

where $x = (1, x^1, \dots, x^n)^T \in \mathbb{R}^{n+1}$.

Least squares method

- $L(w, X_{train}) = MSE(w, X_{train}) = \frac{1}{m} \sum_i (w^T \cdot x_i - y_i)^2$ — loss function

Least squares method

Problem statement and assumptions

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$
- $a(x) = f_w(x) = w_0 + w_1 x^1 + w_2 x^2 + \dots + w_n x^n$, where $w = (w_0, w_1, \dots, w_n)^T \in \mathbb{R}^{n+1}$ — model parameters.
- It is convenient to write in vector form

$$a(x) = w^T \cdot x,$$

where $x = (1, x^1, \dots, x^n)^T \in \mathbb{R}^{n+1}$.

Least squares method

- $L(w, X_{train}) = MSE(w, X_{train}) = \frac{1}{m} \sum_i (w^T \cdot x_i - y_i)^2$ — loss function
- The task is to find $\hat{w} = \arg\min_w (L(w, X_{train}))$

Analytical solution

Theorem

The solution to the problem $\arg\min_w (\sum_{i=1}^{m} (w^T \cdot x_i - y_i)^2)$ is $\hat{w} = (X^T X)^{-1} \cdot X^T \cdot y$, where $X_{i,j} = x_i^j$, $y = (y_1, \ldots, y_m)^T$.

Analytical solution

Theorem

The solution to the problem $\arg\min_w(\sum_{i=1}^{m}(w^T \cdot x_i - y_i)^2)$ is $\hat{w} = (X^T X)^{-1} \cdot X^T \cdot y$, where $X_{i,j} = x_i^j$, $y = (y_1, \dots, y_m)^T$.

Proof idea

Let's write the problem in vector form $||Xw - y||^2 \to \min_w$. The necessary condition for a minimum in matrix form is:

$$\frac{\partial}{\partial w}||Xw - y||^2 = 0$$

Polynomial Regression

Idea

It is possible to generate new features based on existing ones by applying non-linear functions

Polynomial Regression

Idea

It is possible to generate new features based on existing ones by applying non-linear functions

Transformation examples

- Exponentiation
- Pairwise products
- Square root

Pros and cons of linear regression

Pros and cons of linear regression

Advantages

- Simple algorithm, not computationally complex
- Linear regression is a well interpretable model
- Despite its simplicity, it can describe quite complex dependencies (for example, polynomials)

Pros and cons of linear regression

Advantages

- Simple algorithm, not computationally complex
- Linear regression is a well interpretable model
- Despite its simplicity, it can describe quite complex dependencies (for example, polynomials)

Disadvantages

- The algorithm assumes that all features are numeric
- The algorithm assumes that the data is normally distributed, which is not always the case
- The algorithm is highly sensitive to outliers

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\max_{w} p(w|x_1, \ldots, x_m, y_1, \ldots, y_m)$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\max_{w} p(w|x_1, \dots, x_m, y_1, \dots, y_m)$$

$$w_{MAP} = \arg\max_{w} \prod_{i} [p(y_i|x_i, w)p(w)]$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\max_w p(w|x_1, \dots, x_m, y_1, \dots, y_m)$$

$$w_{MAP} = \arg\max_w \prod_i [p(y_i|x_i, w)p(w)]$$

$$w_{MAP} = \arg\max_w \sum_i \ln p(y_i|x_i, w) + m \ln p(w)$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\max_w p(w|x_1, \ldots, x_m, y_1, \ldots, y_m)$$

$$w_{MAP} = \arg\max_w \prod_i \left[p(y_i|x_i, w)p(w)\right]$$

$$w_{MAP} = \arg\max_w \sum_i \ln p(y_i|x_i, w) + m \ln p(w)$$

$$w_{MAP} = \arg\max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2} + m \ln p(w)$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\max_w p(w|x_1, \ldots, x_m, y_1, \ldots, y_m)$$

$$w_{MAP} = \arg\max_w \prod_i [p(y_i|x_i, w)p(w)]$$

$$w_{MAP} = \arg\max_w \sum_i \ln p(y_i|x_i, w) + m \ln p(w)$$

$$w_{MAP} = \arg\max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2} + m \ln p(w)$$

$$w_{MAP} = \arg\min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - m \ln p(w)$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\max_w p(w|x_1, \ldots, x_m, y_1, \ldots, y_m)$$

$$w_{MAP} = \arg\max_w \prod_i [p(y_i|x_i, w)p(w)]$$

$$w_{MAP} = \arg\max_w \sum_i \ln p(y_i|x_i, w) + m \ln p(w)$$

$$w_{MAP} = \arg\max_w \sum_i -\frac{(y_i - w^T x_i)^2}{2\sigma^2} + m \ln p(w)$$

$$w_{MAP} = \arg\min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - m \ln p(w)$$

An additional term appeared in the minimization problem, which depends only on the prior distribution on the weights w

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\min_{w} \sum_{i} \frac{(y_i - w^T x_i)^2}{2\sigma^2} - m \ln p(w)$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\min_{w} \sum_{i} \frac{(y_i - w^T x_i)^2}{2\sigma^2} - m \ln p(w)$$

Let's assume that $p(w) \sim N(0, \tau^2)$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - m \ln p(w)$$

Let's assume that $p(w) \sim N(0, \tau^2)$

$$w_{MAP} = \arg\min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} + \frac{m||w||^2}{2\tau^2}$$

Maximum A Posterior Probability Method

$$w_{MAP} = \arg\min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} - m \ln p(w)$$

Let's assume that $p(w) \sim N(0, \tau^2)$

$$w_{MAP} = \arg\min_w \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} + \frac{m||w||^2}{2\tau^2}$$

$$w_{MAP} = \arg\min_w \frac{1}{m} \sum_i \frac{(y_i - w^T x_i)^2}{2\sigma^2} + \frac{1}{2\tau^2}||w||^2$$

Ridge Regression

L_2 regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2 = \frac{1}{m}\sum_i (w^T \cdot x_i - y_i)^2 + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2$ — loss function

Ridge Regression

L_2 regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2 = \frac{1}{m}\sum_i (w^T \cdot x_i - y_i)^2 + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2$ — loss function
- The task is to find $\hat{w} = \arg\min_w (L(w, X_{train}))$

Ridge Regression

L_2 regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2 = \frac{1}{m}\sum_i (w^T \cdot x_i - y_i)^2 + \frac{\alpha}{2}\sum_{i=0}^{n} w_i^2$ — loss function
- The task is to find $\hat{w} = \arg\min_w (L(w, X_{train}))$

Note: L_2 regularization and MAP with the normally distributed weights are the same!

Ridge regression: solution

Theorem

The solution of the problem $\arg\min_w (\sum_{i=1}^{m} (w^T \cdot x^{(i)} - y_i)^2 + \alpha \sum_{i=0}^{n} w_i^2)$ is $\hat{w} = (X^T X + \alpha I_{n+1})^{-1} \cdot X^T \cdot y$, where $X_{i,j} = x_i^j$, $y = (y_1, \ldots, y_m)^T$, I_{n+1} is the identity matrix.

Proof idea

Let's write the problem in vector form $||Xw - y||^2 + \alpha ||w||^2 \to \min_w$. The necessary condition for a minimum in matrix form is:

$$\frac{\partial}{\partial w} \left((Xw - y)^T \cdot (Xw - y) + \alpha w^T w \right) = 0$$

Ridge Regression: Properties

- Regularization prevents model parameters from being too large
- In general, regularization provides better generalization ability
- More resistant to outliers
- A parameter has been added that can be configured using cross-validation

Ridge Regression: Properties

- Regularization prevents model parameters from being too large
- In general, regularization provides better generalization ability
- More resistant to outliers
- A parameter has been added that can be configured using cross-validation

The probabilistic meaning of the α parameter

$\alpha = \frac{1}{\tau^2}$, where τ is the standard deviation of the prior distribution on w

LASSO

L_1-regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + \alpha \sum_{i=0}^{n} |w_i| = \sum_i (w^T \cdot x_i - y_i)^2 + \alpha \sum_{i=0}^{n} |w_i|$ — loss function

LASSO

L_1-regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + \alpha \sum_{i=0}^{n} |w_i| = \sum_i (w^T \cdot x_i - y_i)^2 + \alpha \sum_{i=0}^{n} |w_i|$ — loss function
- The task is to find $\hat{w} = \arg\min_w (L(w, X_{train}))$

LASSO

L_1-regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + \alpha \sum\limits_{i=0}^{n} |w_i| = \sum\limits_{i} (w^T \cdot x_i - y_i)^2 + \alpha \sum\limits_{i=0}^{n} |w_i|$ — loss function
- The task is to find $\hat{w} = \arg\min\limits_{w} (L(w, X_{train}))$

Properties

- This regularization provides feature selection
- No analytical solution

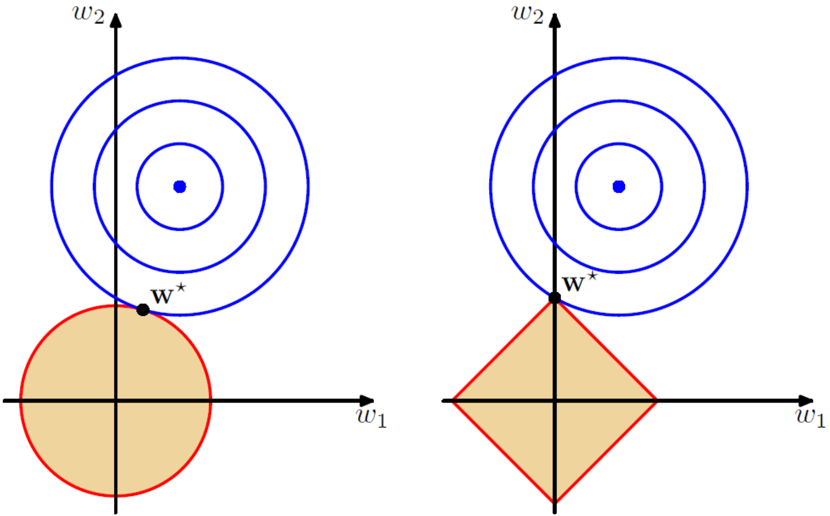
Probabilistic interpretation of LASSO

The probabilistic meaning of the α parameter

The parameter α — is inversely proportional to the standard deviation of the prior distribution by w. In this case, this is the *Laplace* distribution

$$p(w) = \frac{1}{\tau} exp\left(-\frac{||w||}{2\tau}\right)$$

Intuition of feature selection under L_1-regularization



Elastic Net

L_1-regularization and L_2-regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + r\alpha \sum_{i=0}^{n} |w_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^{n} w_i^2 = \sum_i (w^T \cdot x_i - y_i)^2 + r\alpha \sum_{i=0}^{n} |w_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^{n} w_i^2$ — loss function

Elastic Net

L_1-regularization and L_2-regularization

- $L(w, X_{train}) = MSE(w, X_{train}) + r\alpha \sum_{i=0}^{n} |w_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^{n} w_i^2 = \sum_i (w^T \cdot x_i - y_i)^2 + r\alpha \sum_{i=0}^{n} |w_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^{n} w_i^2$ — loss function
- The task is to find $\hat{w} = \arg\min_w (L(w, X_{train}))$

Properties

- No analytical solution
- Combines the positive properties of Ridge regression and LASSO.

Quality Metrics for the Regression Problem

Motivation

- The formulation of a machine learning problem usually begins with the definition of a metric and fixing a test dataset on which this metric will be calculated

Quality Metrics for the Regression Problem

Motivation

- The formulation of a machine learning problem usually begins with the definition of a metric and fixing a test dataset on which this metric will be calculated
- An incorrectly chosen metric can make it difficult to use the machine learning model in real life and nullify the efforts of the team developing the machine learning algorithm

Quality Metrics for the Regression Problem

Motivation

- The formulation of a machine learning problem usually begins with the definition of a metric and fixing a test dataset on which this metric will be calculated
- An incorrectly chosen metric can make it difficult to use the machine learning model in real life and nullify the efforts of the team developing the machine learning algorithm
- As a rule, the customer does not think in terms of metrics and can only explain the problem he wants to solve in business language

Quality Metrics for the Regression Problem

Motivation

- The formulation of a machine learning problem usually begins with the definition of a metric and fixing a test dataset on which this metric will be calculated
- An incorrectly chosen metric can make it difficult to use the machine learning model in real life and nullify the efforts of the team developing the machine learning algorithm
- As a rule, the customer does not think in terms of metrics and can only explain the problem he wants to solve in business language
- Understanding the impact of the choice of a particular metric on the customer's business is the key to successful problem setting

Quality Metrics for the Regression Problem[1]

Mean Square Error

$$MSE = \frac{1}{m}\sum_{i}(y_i - a(x_i))^2$$

[1]**Note**: unless otherwise specified, metrics are applied on top of the test dataset.

Quality Metrics for the Regression Problem[1]

Mean Square Error

$$MSE = \frac{1}{m}\sum_i (y_i - a(x_i))^2$$

Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{m}\sum_i (y_i - a(x_i))^2}$$

[1] **Note**: unless otherwise specified, metrics are applied on top of the test dataset.

Quality Metrics for the Regression Problem[1]

Mean Square Error

$$MSE = \frac{1}{m}\sum_i (y_i - a(x_i))^2$$

Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{m}\sum_i (y_i - a(x_i))^2}$$

Mean Absolute Error

$$MAE = \frac{1}{m}\sum_i |y_i - a(x_i)|$$

[1] **Note**: unless otherwise specified, metrics are applied on top of the test dataset.

Quality Metrics for the Regression Problem

Max Error

$$ME = max(|y_i - a(x_i)|)$$

Quality Metrics for the Regression Problem

Max Error

$$ME = max(|y_i - a(x_i)|)$$

Mean Squared Logarithmic Error

$$MSLE = \frac{1}{m}\sum_i (\ln y_i - \ln a(x_i))^2$$

Quality Metrics for the Regression Problem

Max Error

$$ME = max(|y_i - a(x_i)|)$$

Mean Squared Logarithmic Error

$$MSLE = \frac{1}{m}\sum_i (\ln y_i - \ln a(x_i))^2$$

R^2 score (also known as Coefficient of Determination)

$$R^2 = 1 - \frac{\sum_i (y_i - a(x_i))^2}{\sum_i (y_i - \bar{y})^2},$$

where $\bar{y} = \frac{1}{m}\sum_i y_i$.

Conclusion

- Linear regression — simple, well-interpreted model, but not robust to outliers

Conclusion

- Linear regression — simple, well-interpreted model, but not robust to outliers
- Linear regression variants have a clear probabilistic interpretation

Conclusion

- Linear regression — simple, well-interpreted model, but not robust to outliers
- Linear regression variants have a clear probabilistic interpretation
- Regularization is a great way to deal with the overfitting and data noise

Conclusion

- Linear regression — simple, well-interpreted model, but not robust to outliers
- Linear regression variants have a clear probabilistic interpretation
- Regularization is a great way to deal with the overfitting and data noise
- Ridge Regression == L_2-regularization == adding weighted square (L_2) norm of w to MSE of linear regression

Conclusion

- Linear regression — simple, well-interpreted model, but not robust to outliers
- Linear regression variants have a clear probabilistic interpretation
- Regularization is a great way to deal with the overfitting and data noise
- Ridge Regression $==$ L_2-regularization $==$ adding weighted square (L_2) norm of w to MSE of linear regression
- LASSO $==$ L_1-regularization $==$ adding weighted L_1 norm of w to MSE of linear regression

Conclusion

- Linear regression — simple, well-interpreted model, but not robust to outliers
- Linear regression variants have a clear probabilistic interpretation
- Regularization is a great way to deal with the overfitting and data noise
- Ridge Regression == L_2-regularization == adding weighted square (L_2) norm of w to MSE of linear regression
- LASSO == L_1-regularization == adding weighted L_1 norm of w to MSE of linear regression
- Elastic Net == $L_1 + L_2$-regularization == adding weighted L_1 and squared L_2 norm of w to MSE of linear regression

Thank you!