# Machine Learning Process #2

Madson Luiz Dantas Dias e Lucas Silva de Sousa

Huawei / IFCE

March 8, 2021

# Agenda

1. Model Validation
   - Training and Generalization Error
   - Underfitting and Overfitting
   - Model capacity
   - Bias and Variance
   - Model Complexity and Error
2. Model Learning Performance Evaluation
   - Classification
   - Regression

# Agenda

1. Model Validation
   - Training and Generalization Error
   - Underfitting and Overfitting
   - Model capacity
   - Bias and Variance
   - Model Complexity and Error
2. Model Learning Performance Evaluation
   - Classification
   - Regression

# Model Validation

**Model validation**

# Model Validation

**Model validation**

- The process where a **trained model** is evaluated with a **testing data set**

# Model Validation

**Model validation**

- The process where a **trained model** is evaluated with a **testing data set**
- The main purpose of using the testing data set is to test the **generalization** ability of a trained model

# Model Validation

## Model validation

- The process where a **trained model** is evaluated with a **testing data set**
- The main purpose of using the testing data set is to test the **generalization** ability of a trained model

## Generalization

# Model Validation

**Model validation**

- The process where a **trained model** is evaluated with a **testing data set**
- The main purpose of using the testing data set is to test the **generalization** ability of a trained model

**Generalization**

- The goal of machine learning is that the model obtained after learning should perform well on new samples

# Model Validation

### Model validation

- The process where a **trained model** is evaluated with a **testing data set**
- The main purpose of using the testing data set is to test the **generalization** ability of a trained model

### Generalization

- The goal of machine learning is that the model obtained after learning should perform well on new samples
- The capability of applying a model to new samples is called generalization or robustness

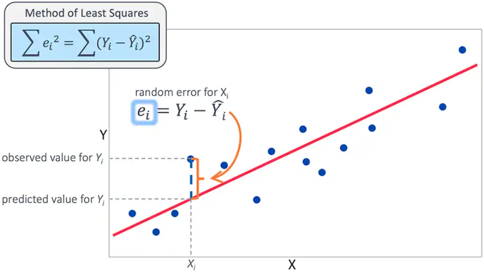
# Agenda

1. Model Validation
   - Training and Generalization Error
   - Underfitting and Overfitting
   - Model capacity
   - Bias and Variance
   - Model Complexity and Error
2. Model Learning Performance Evaluation
   - Classification
   - Regression

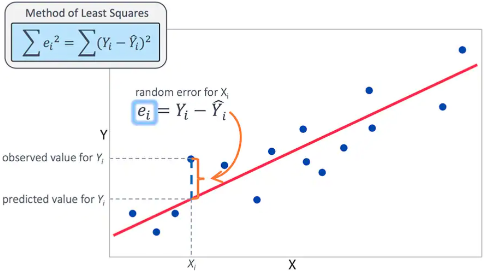
# Model Validation

# Model Validation

## Error



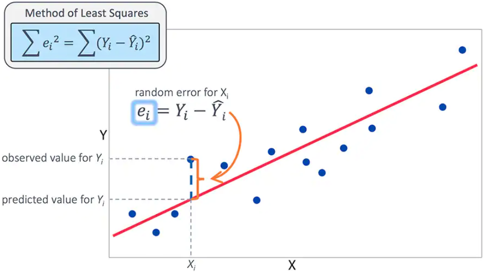
## Error




- Difference between the sample result predicted by the model obtained after learning and the actual sample result.

# Model Validation

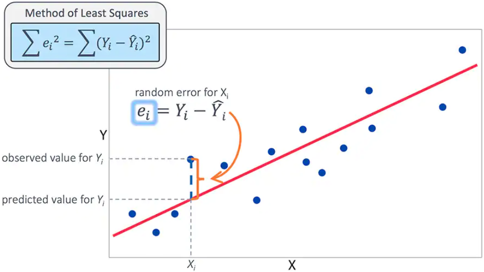
## Error




- Difference between the sample result predicted by the model obtained after learning and the actual sample result.
- **Training error**: error that you get when you run the model on the training data.

# Model Validation

## Error




- Difference between the sample result predicted by the model obtained after learning and the actual sample result.
- **Training error**: error that you get when you run the model on the training data.
- **Generalization error**: error that you get when you run the model on new samples.

# Agenda

# Underfitting and Overfitting

# Underfitting and Overfitting

**Underfitting**

- Occurs when the model or the algorithm does not fit the data well enough

# Underfitting and Overfitting

**Underfitting**

- Occurs when the model or the algorithm does not fit the data well enough

**Overfitting**

- Occurs when the training error of the model obtained after learning is small but the generalization error is large

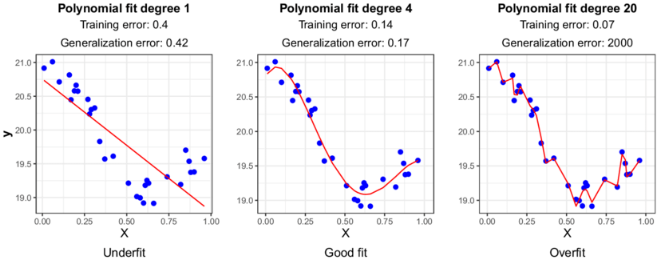
# Underfitting and Overfitting

## Underfitting

- Occurs when the model or the algorithm does not fit the data well enough

## Overfitting

- Occurs when the training error of the model obtained after learning is small but the generalization error is large



# Agenda

# Model capacity

# Model capacity

- Model's capability of fitting functions, which is also called model complexity.

# Model capacity

- Model's capability of fitting functions, which is also called model complexity.
- When the capacity suits the task complexity and the amount of training data provided, the algorithm effect is usually optimal.

# Model capacity

- Model's capability of fitting functions, which is also called model complexity.
- When the capacity suits the task complexity and the amount of training data provided, the algorithm effect is usually optimal.
- Models with insufficient capacity cannot solve complex tasks and underfitting may occur.

# Model capacity

- Model's capability of fitting functions, which is also called model complexity.
- When the capacity suits the task complexity and the amount of training data provided, the algorithm effect is usually optimal.
- Models with insufficient capacity cannot solve complex tasks and underfitting may occur.
- A high-capacity model can solve complex tasks, but overfitting may occur if the capacity is higher than that required by a task.

# Agenda

# Bias and Variance

- Let $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ a data set. We assume there is a relationship between a $\boldsymbol{x}_n$ and a $y_n$ is

$$y_n = f(\boldsymbol{x}) + \epsilon \tag{1}$$

Where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$.

# Bias and Variance

- Let $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ a data set. We assume there is a relationship between a $\boldsymbol{x}_n$ and a $y_n$ is

$$y_n = f(\boldsymbol{x}) + \epsilon \tag{1}$$

  Where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$.
- We will make a model $\hat{f}(\boldsymbol{x})$ of $f(\boldsymbol{x})$ using a modeling technique. So the expected squared error at a point $\boldsymbol{x}$ is

# Bias and Variance

- Let $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ a data set. We assume there is a relationship between a $\boldsymbol{x}_n$ and a $y_n$ is

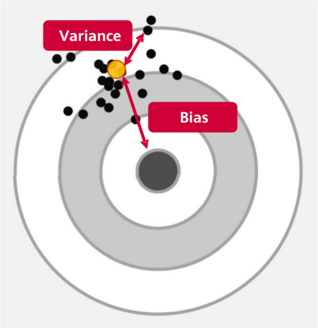
$$y_n = f(\boldsymbol{x}) + \epsilon \tag{1}$$

  Where $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$.
- We will make a model $\hat{f}(\boldsymbol{x})$ of $f(\boldsymbol{x})$ using a modeling technique. So the expected squared error at a point $\boldsymbol{x}$ is

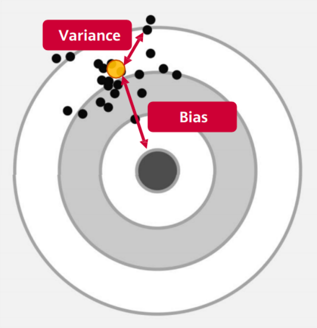
$$\begin{aligned} \mathrm{Error}(\boldsymbol{x}) &= \mathsf{E}\left[(y - \hat{f}(\boldsymbol{x}))^2\right] \qquad (2) \\ &= \underbrace{\left(\mathsf{E}\left[f(\boldsymbol{x})\right] - \hat{f}(\boldsymbol{x})\right)^2}_{\mathrm{bias}^2} + \underbrace{\mathsf{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathsf{E}\left[\hat{f}(\boldsymbol{x})\right]\right)^2\right]}_{\mathrm{variance}} \\ &+ \underbrace{\sigma_e^2}_{\text{irreducible error}} \end{aligned}$$

# Bias and Variance

# Bias and Variance



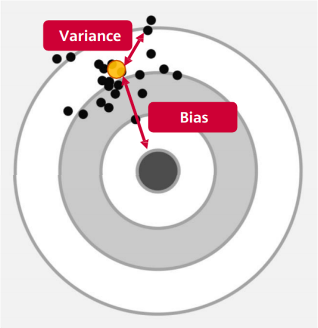
# Bias and Variance




**Variance**

- Offset of the prediction result from the average value

# Bias and Variance




**Variance**

- Offset of the prediction result from the average value

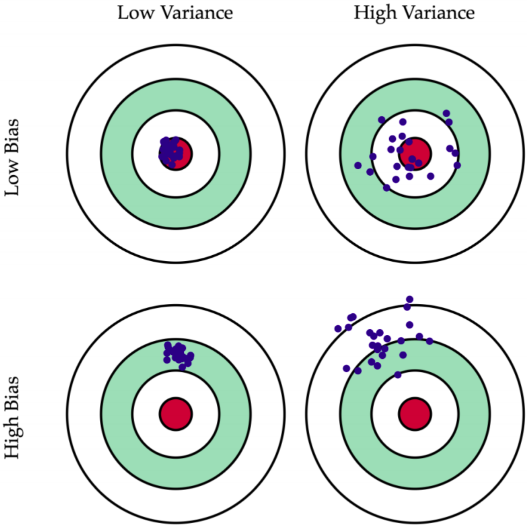
**Bias**

- Difference between the expected prediction value and the correct

# Bias and Variance

# Bias and Variance

| *bias* | *variance* |
|---|---|
| $\downarrow$ | $\downarrow$ |
| $\downarrow$ | $\uparrow$ |
| $\uparrow$ | $\downarrow$ |
| $\uparrow$ | $\uparrow$ |



# Agenda

1. Model Validation
   - Training and Generalization Error
   - Underfitting and Overfitting
   - Model capacity
   - Bias and Variance
   - Model Complexity and Error
2. Model Learning Performance Evaluation
   - Classification
   - Regression

# Model Complexity and Error

# Model Complexity and Error

As the model complexity increases

# Model Complexity and Error

As the model complexity increases

- The training error decreases

# Model Complexity and Error

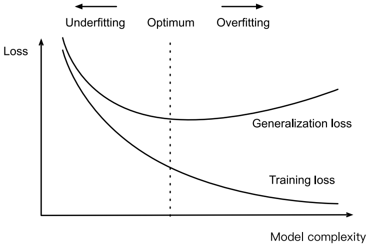
As the model complexity increases

- The training error decreases
- The test error decreases to a certain point and then increases in the reverse direction, forming a convex curve.

# Model Complexity and Error

As the model complexity increases

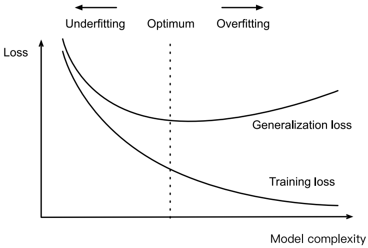
- The training error decreases
- The test error decreases to a certain point and then increases in the reverse direction, forming a convex curve.

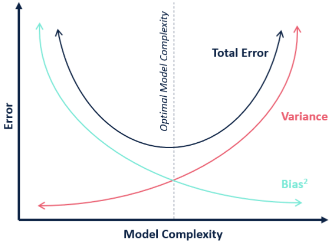


# Model Complexity and Error

As the model complexity increases

- The training error decreases
- The test error decreases to a certain point and then increases in the reverse direction, forming a convex curve.



# Agenda

1. Model Validation
   - Training and Generalization Error
   - Underfitting and Overfitting
   - Model capacity
   - Bias and Variance
   - Model Complexity and Error
2. Model Learning Performance Evaluation
   - Classification
   - Regression

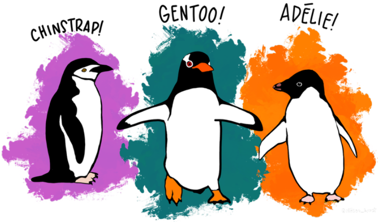
# Agenda

# Classification

- Classification is the process of predicting the class of given data points.

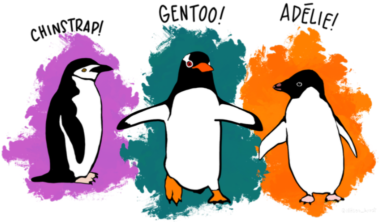
# Classification

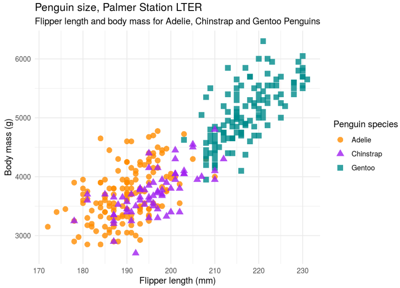
- Classification is the process of predicting the class of given data points.



# Classification

- Classification is the process of predicting the class of given data points.



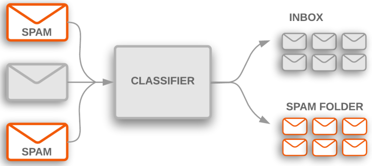
# Classification

# Classification

- Binary classification

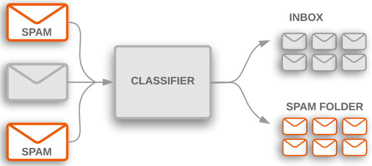
# Classification

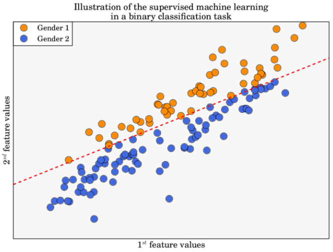
- Binary classification



# Classification

- Binary classification



# Classification

**Terms and definitions**

# Classification

**Terms and definitions**

- Positive ($P$): indicating the number of real positive cases

# Classification

**Terms and definitions**

- Positive ($P$): indicating the number of real positive cases
- Negative ($N$): indicating the number of real negative cases

# Classification

**Terms and definitions**

- Positive ($P$): indicating the number of real positive cases
- Negative ($N$): indicating the number of real negative cases
- True positive ($TP$): indicating the number of correctly classified positive cases

# Classification

**Terms and definitions**

- Positive ($P$): indicating the number of real positive cases
- Negative ($N$): indicating the number of real negative cases
- True positive ($TP$): indicating the number of correctly classified positive cases
- True negative ($TN$): indicating the number of correctly classified negative cases

# Classification

**Terms and definitions**

- Positive ($P$): indicating the number of real positive cases
- Negative ($N$): indicating the number of real negative cases
- True positive ($TP$): indicating the number of correctly classified positive cases
- True negative ($TN$): indicating the number of correctly classified negative cases
- False positive ($FP$): indicating the number of incorrectly classified positive cases

# Classification

**Terms and definitions**

- Positive ($P$): indicating the number of real positive cases
- Negative ($N$): indicating the number of real negative cases
- True positive ($TP$): indicating the number of correctly classified positive cases
- True negative ($TN$): indicating the number of correctly classified negative cases
- False positive ($FP$): indicating the number of incorrectly classified positive cases
- False negative ($FN$): indicating the number of incorrectly classified negative cases

# Classification

**Confusion matrix**

- A table used to describe the performance of a classifier on a set of test data for which the true values are known.

# Classification

## Confusion matrix

- A table used to describe the performance of a classifier on a set of test data for which the true values are known.

| | | **Predicted Class** | | |
|---|---|---|---|---|
| | | **Positive** | **Negative** | |
| **Actual Class** | **Positive** | True Positive (TP) | False Negative (FN) **Type II Error** | **Sensitivity** $\frac{TP}{(TP+FN)}$ |
| | **Negative** | False Positive (FP) **Type I Error** | True Negative (TN) | **Specificity** $\frac{TN}{(TN+FP)}$ |
| | | **Precision** $\frac{TP}{(TP+FP)}$ | **Negative Predictive Value** $\frac{TN}{(TN+FN)}$ | **Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$ |

# Classification

# Classification

## Confusion matrix example and accuracy rate

| | | Predicted Class | |
|---|---|---|---|
| | | **Spam** | **Non-Spam** |
| **Actual Class** | **Spam** | TP=45 | FN=20 |
| | **Non-Spam** | FP=5 | TN=30 |

## Confusion matrix example and accuracy rate

| | | Predicted Class | |
|---|---|---|---|
| | | **Spam** | **Non-Spam** |
| **Actual Class** | **Spam** | TP=45 | FN=20 |
| | **Non-Spam** | FP=5 | TN=30 |

$$\mathrm{ACC}\left(\hat{f}\left(\mathcal{D}_{\mathsf{test}};\mathcal{D}_{\mathsf{train}},\boldsymbol{\theta}\right)\right) = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

# Classification

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

| | |
|---|---|
| Error rate and misclassification rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

| | |
|---|---|
| Error rate and misclassification rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

| | |
|---|---|
| Error rate and misclassification rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Sensitivity or recall | $\frac{TP}{TP+FN}$ |

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

| | |
|---|---|
| Error rate and misclassification rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Sensitivity or recall | $\frac{TP}{TP+FN}$ |
| Specificity and true negative rate | $\frac{TN}{TN+FP}$ |

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

| | |
|---|---|
| Error rate and misclassification rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Sensitivity or recall | $\frac{TP}{TP+FN}$ |
| Specificity and true negative rate | $\frac{TN}{TN+FP}$ |
| $F_1$-score | $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ |

[1] where $\beta$ is a non-negative real number

# Classification

## Other metrics

| | |
|---|---|
| Error rate and misclassification rate | $\frac{FN+FP}{TP+FP+TN+FN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Sensitivity or recall | $\frac{TP}{TP+FN}$ |
| Specificity and true negative rate | $\frac{TN}{TN+FP}$ |
| $F_1$-score | $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ |
| $F_\beta$-score[1] | $(1 - \beta^2) \times \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}$ |

[1] where $\beta$ is a non-negative real number

# Classification

# Classification

**Receiving Operating Characteristic (ROC) curves**

- Shows the sensitivity/specificity trade-off of a classifier for all possible classification thresholds

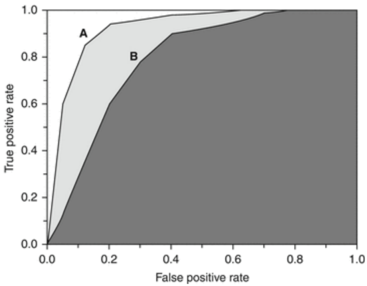
# Classification

**Receiving Operating Characteristic (ROC) curves**

- Shows the sensitivity/specificity trade-off of a classifier for all possible classification thresholds

**Area under a ROC curve**

- Abbreviated as AUC, is a single scalar value that measures the overall performance of a binary classifier



# Classification

# Classification

### Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

# Classification

## Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

- Accuracy rate?

# Classification

**Performance evaluation example**

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

- Accuracy rate? $\frac{150}{200} = 0.75$

# Classification

**Performance evaluation example**

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision?

# Classification

## Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision? $\frac{140}{140+20} = 0.875$

# Classification

### Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision? $\frac{140}{140+20} = 0.875$
- Recall?

# Classification

### Performance evaluation example

- We have trained a machine learning model to identify whether the object in an image is a cat. Now we use 200 pictures to verify the model performance. Among the 200 images, objects in 170 images are cats, while others are not. The identification result of the model is that objects in 160 images are cats, while others are not.

| | yes | no |
|---|---|---|
| yes | 140 | 30 |
| no | 20 | 10 |

- Accuracy rate? $\frac{150}{200} = 0.75$
- Precision? $\frac{140}{140+20} = 0.875$
- Recall? $\frac{140}{140+30} = 0.824$

# Agenda

1. Model Validation
   - Training and Generalization Error
   - Underfitting and Overfitting
   - Model capacity
   - Bias and Variance
   - Model Complexity and Error
2. Model Learning Performance Evaluation
   - Classification
   - Regression

# Regression

# Regression

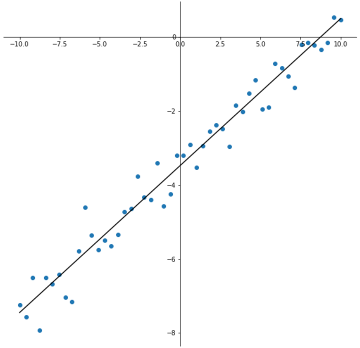
- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.

# Regression

- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common type of regression is linear regression.

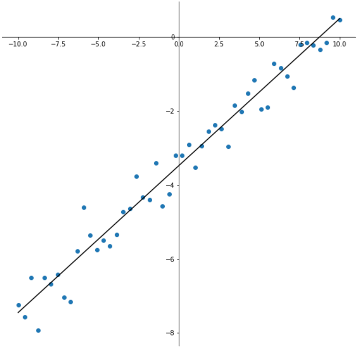
# Regression

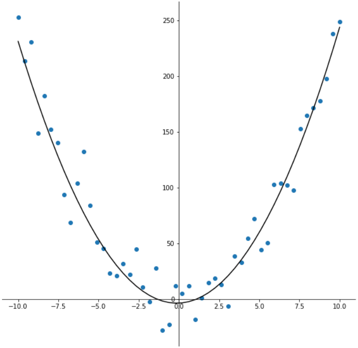
- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common type of regression is linear regression.



# Regression

- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common type of regression is linear regression.



**Regression metrics**

# Regression

**Regression metrics**

| Mean Squared Error (MSE) | $\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$ |
| --- | --- |

# Regression

**Regression metrics**

| | |
|---|---|
| Mean Squared Error (MSE) | $\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$ |
| Mean Absolute Error (MSE) | $\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\lvert y_n - \hat{y}_n \rvert$ |

**Regression metrics**

| | |
|---|---|
| Mean Squared Error (MSE) | $\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$ |
| Mean Absolute Error (MSE) | $\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\lvert y_n - \hat{y}_n\rvert$ |
| R-squared ($R^2$) | $R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y}_n)^2}$ |