Probability Theorems and Metrics

Basics of Machine Learning

Jul 20, 2023
Thu 4 PM

Kwangwoon University MI:RU
Artificial Intelligence Study

Intro

In this course, you will learn

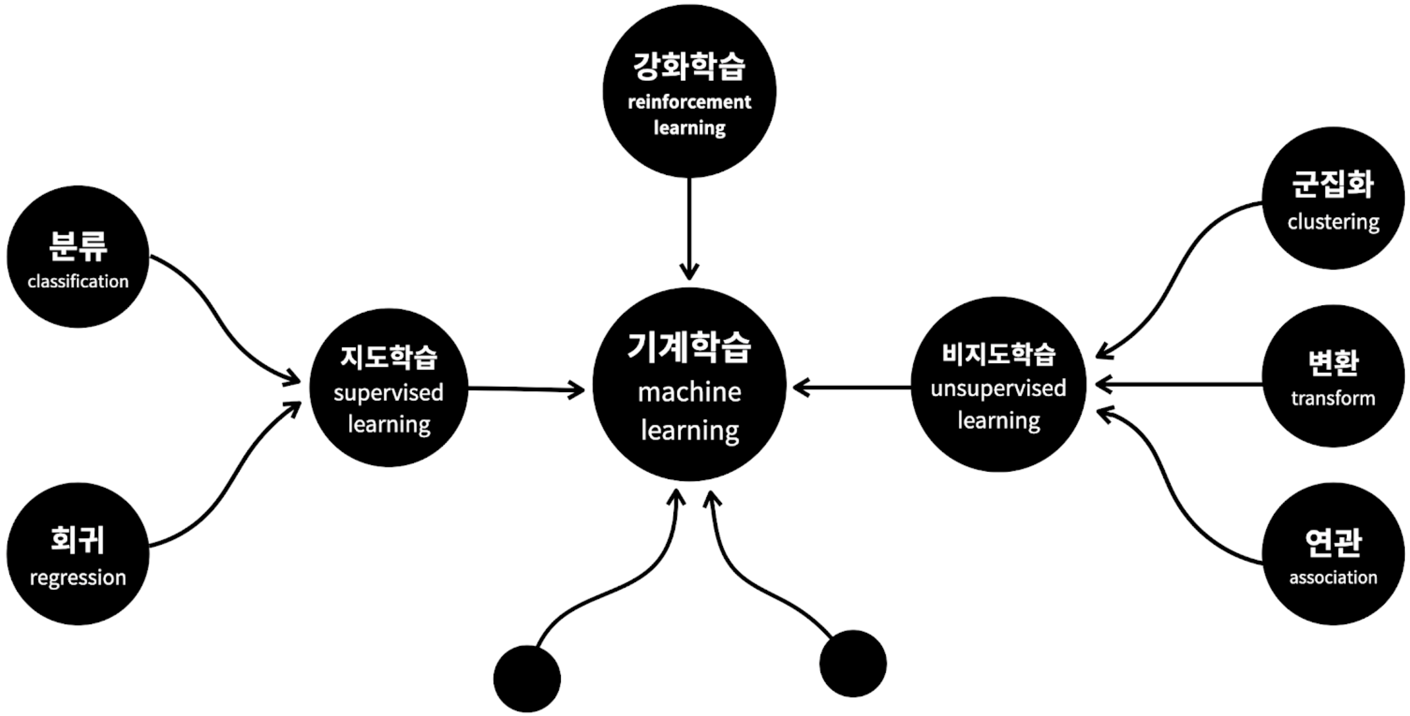
Part 1 – Metrics for Performance check

- Performance Metrics
  - Regression
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
  - Classification
    - Confusion Matrix
      - Accuracy
      - Precision
      - Sensitivity
      - Specificity
    - F1 Score
    - ROC and AUC

Part 2 – Probability Theorems

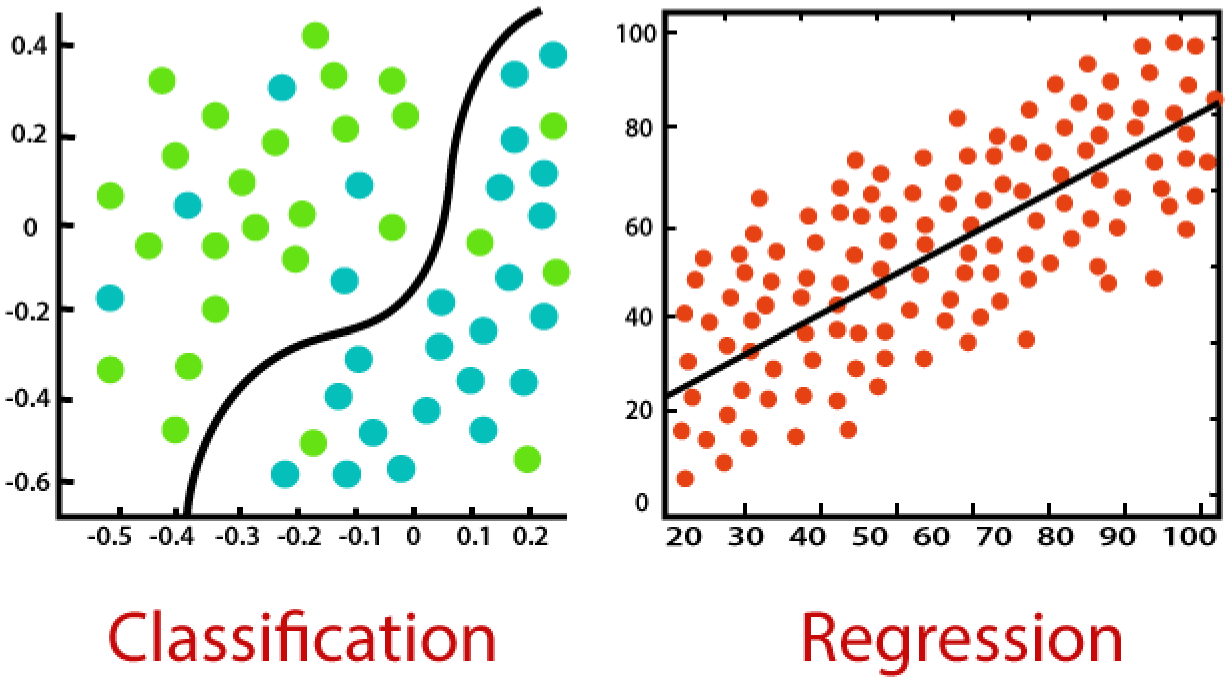
- Concept of Likelihood
- Bayes Theorem

Before we get started…



Before we get started…

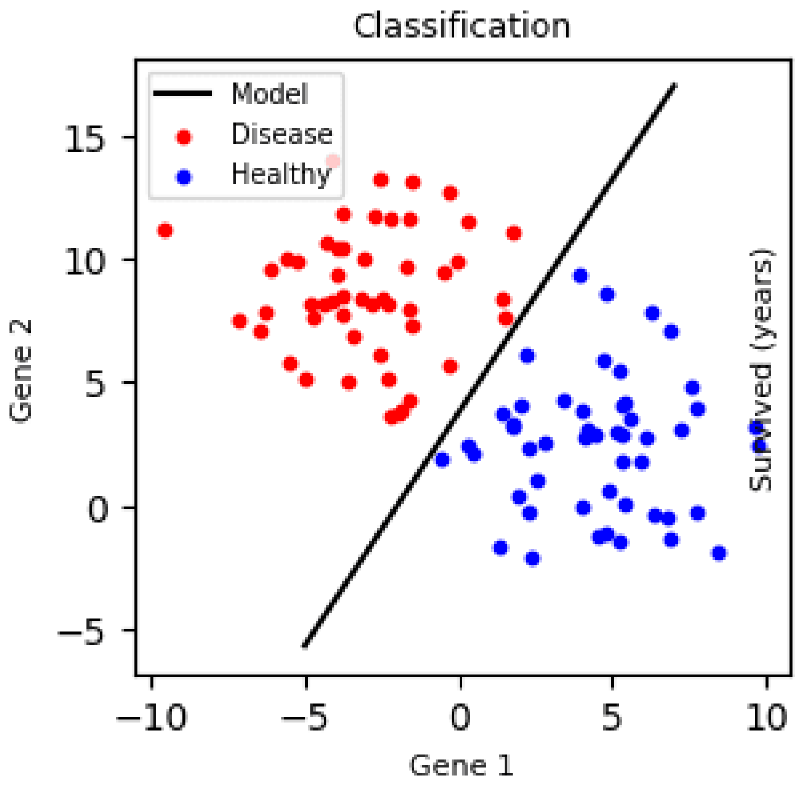
Classification and Regression



[1] Regression vs Classification in Machine Learning

Before we get started…

Classification

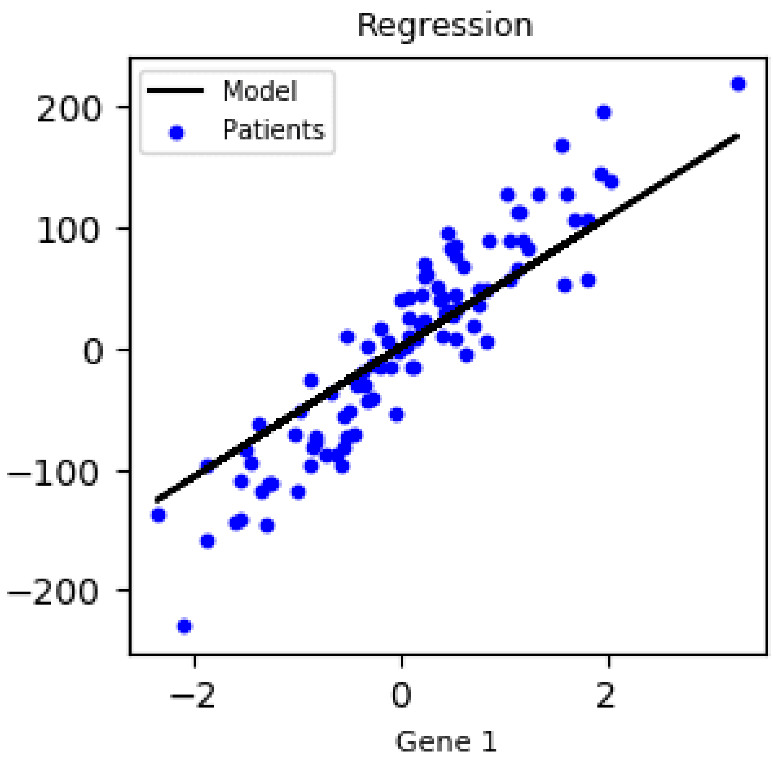


Discrete Problem!(0 or 1)

[2] Classification

Before we get started…

Classification



Continuous Problem!

[2] Classification

Before we get started…

Garbage In Garbage Out

Performance Metrics

Regression

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

[3] MAE, MSE and RMSE

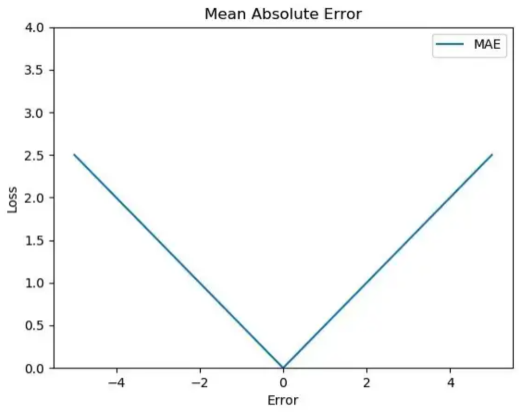
Performance Metrics

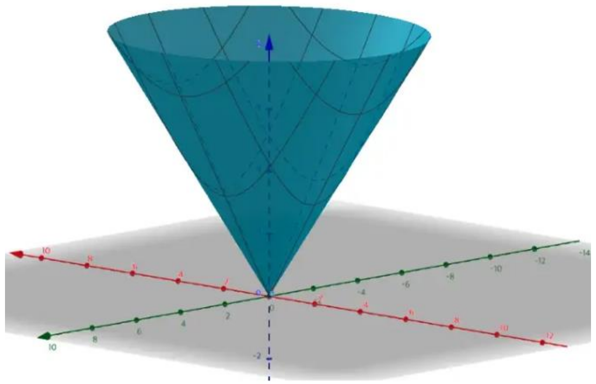
Regression

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$





Common Problems

Loss is positive
Scale dependant

[3] MAE, MSE and RMSE

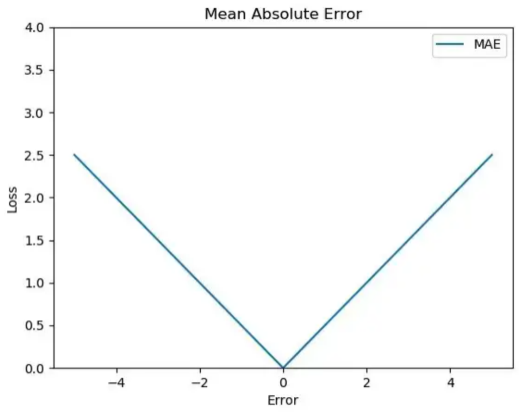
Performance Metrics

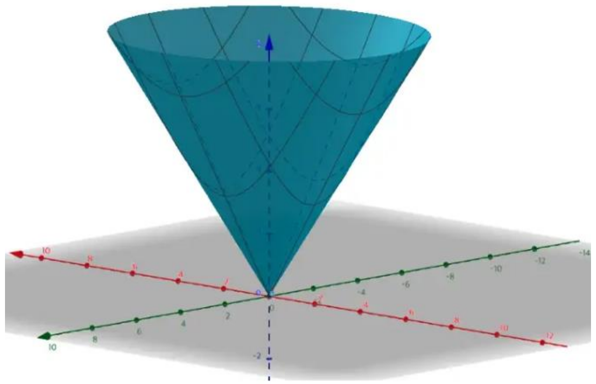
Regression

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$





	MAE	MSE	RMSE
PROS	Intuitive	Good for big errors	Good for big errors
CONS	Not differentiable at 0	Hard to deal with large value, Not robust	Not Intuitive

[3] MAE, MSE and RMSE

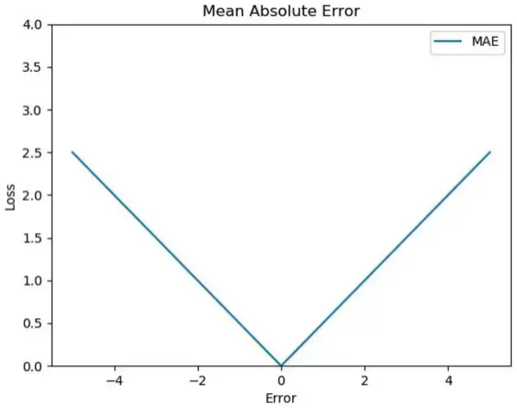
Performance Metrics

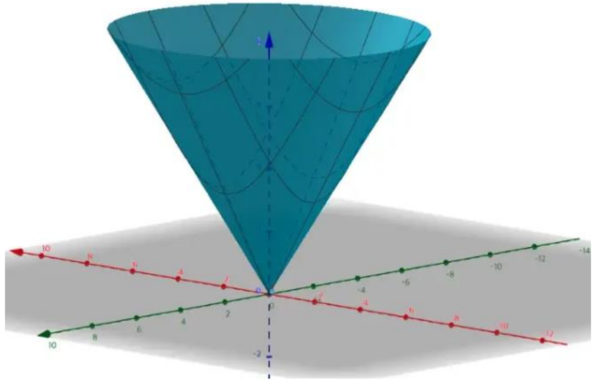
Regression

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$





Strategy

Use RMSE as Loss function
And
Use MAE for performance check only!

[3] MAE, MSE and RMSE

Performance Metrics

Classification

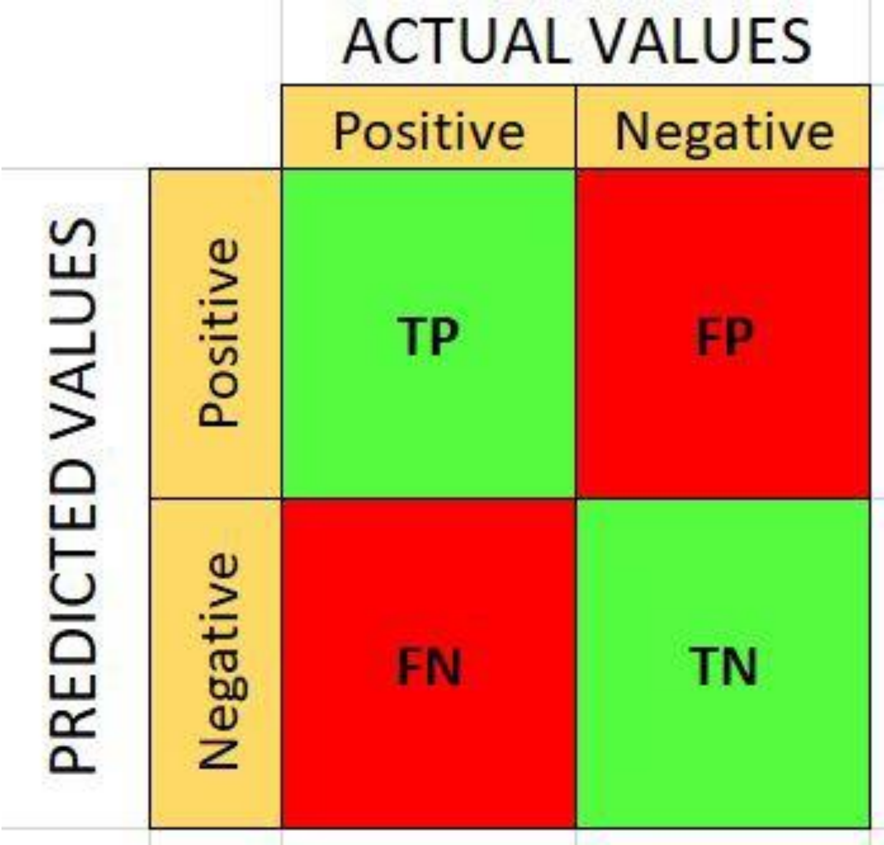
		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive **+**	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

Confusion matrix

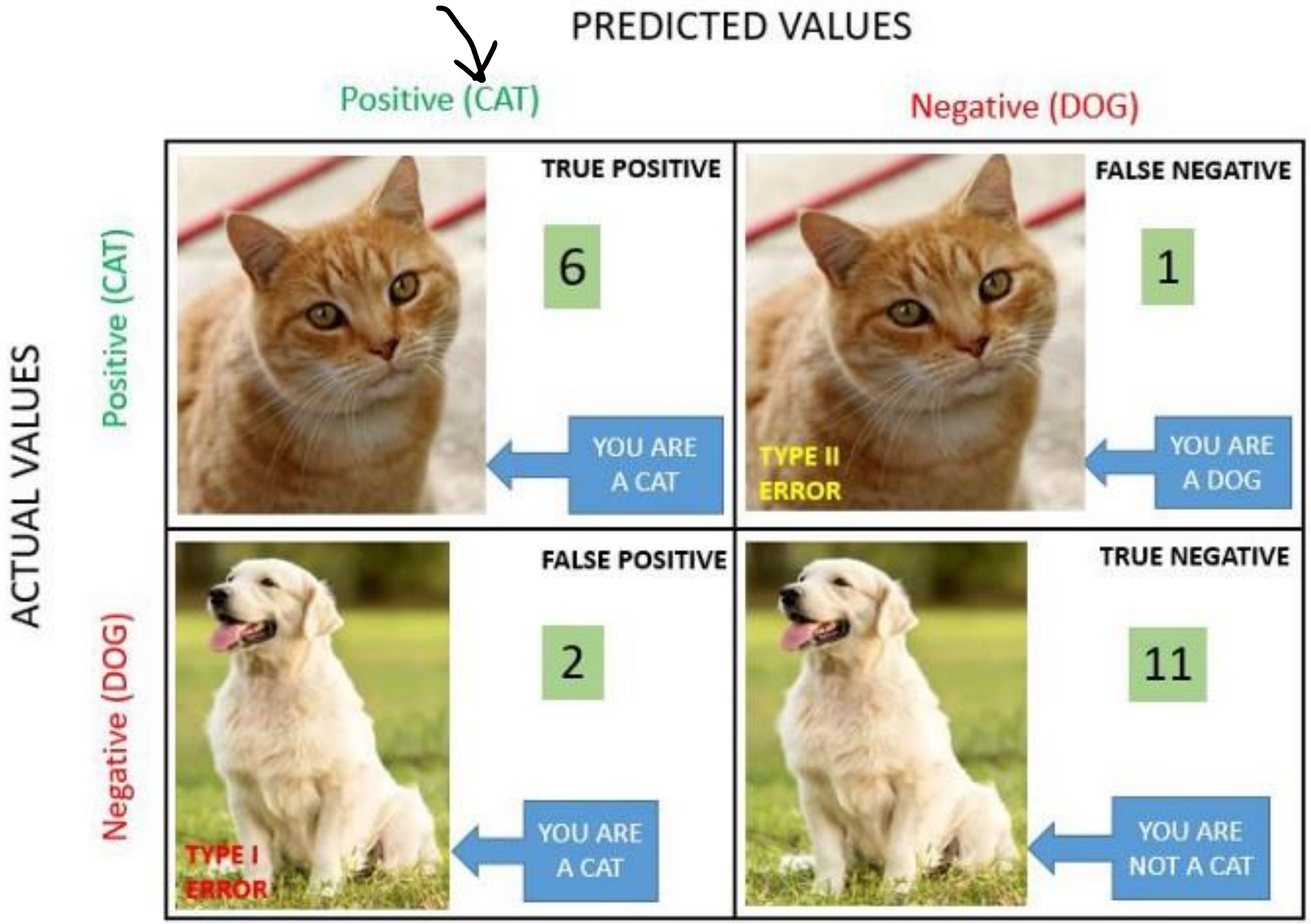
[4] Confusion matrix

Performance Metrics

Classification



Ground truth!



A simple example
- Cat and dog

[4] Confusion matrix

Performance Metrics

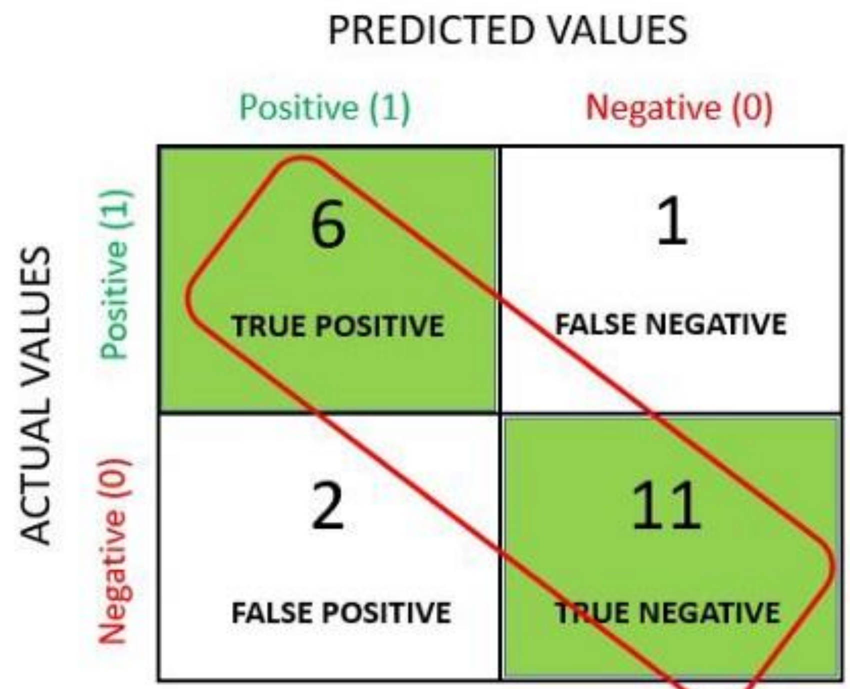
Classification

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

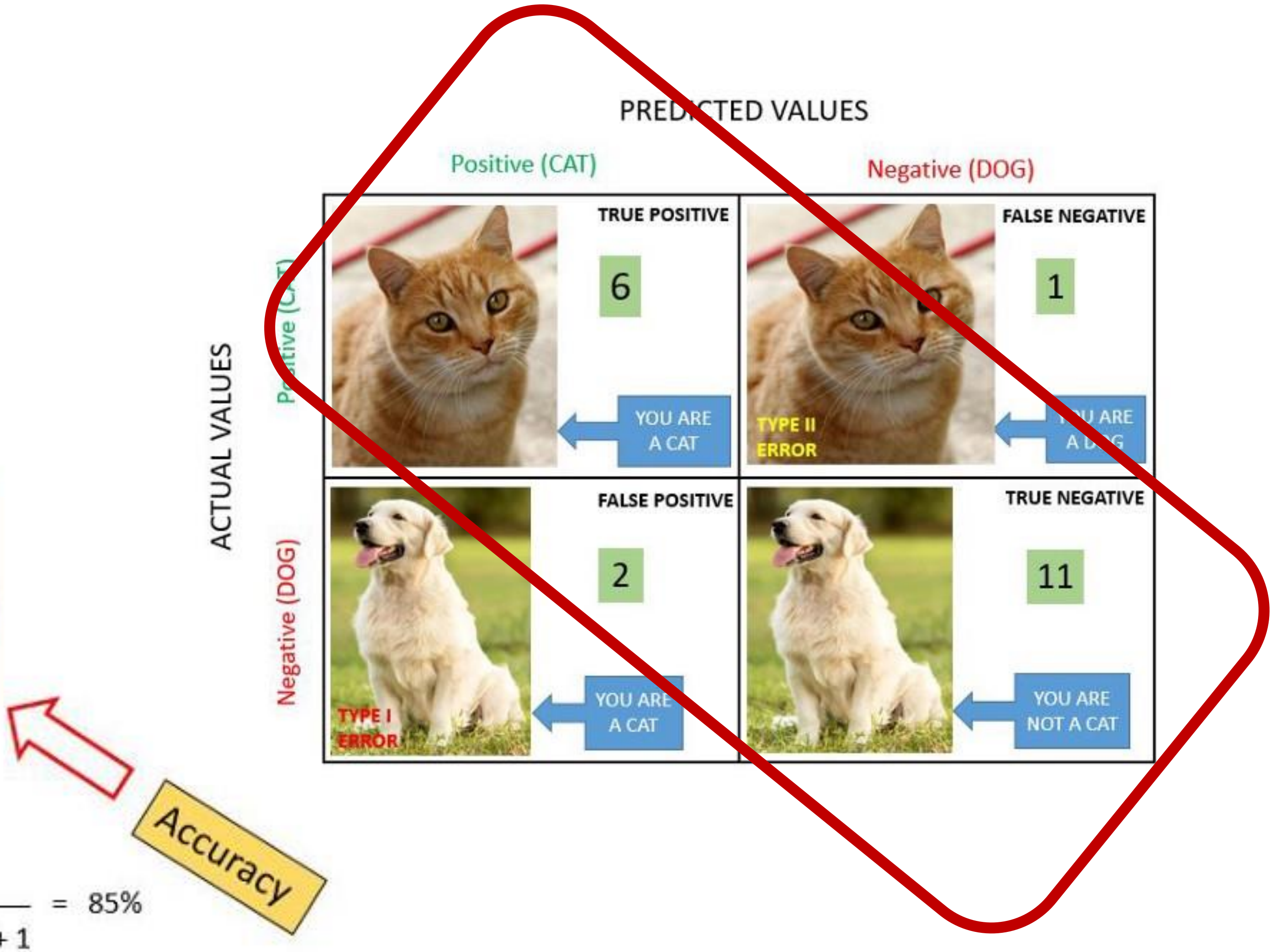
[4] Confusion matrix

Performance Metrics

Accuracy (ACC), 정확도



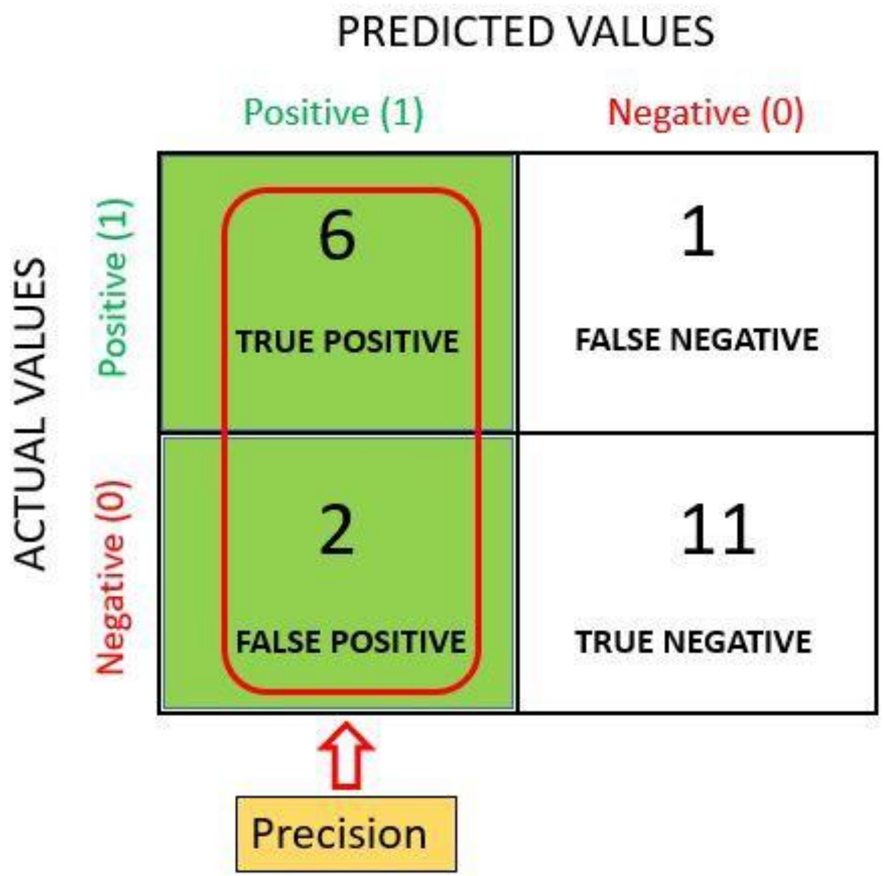
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{6 + 11}{6 + 11 + 2 + 1} = 85\%$$

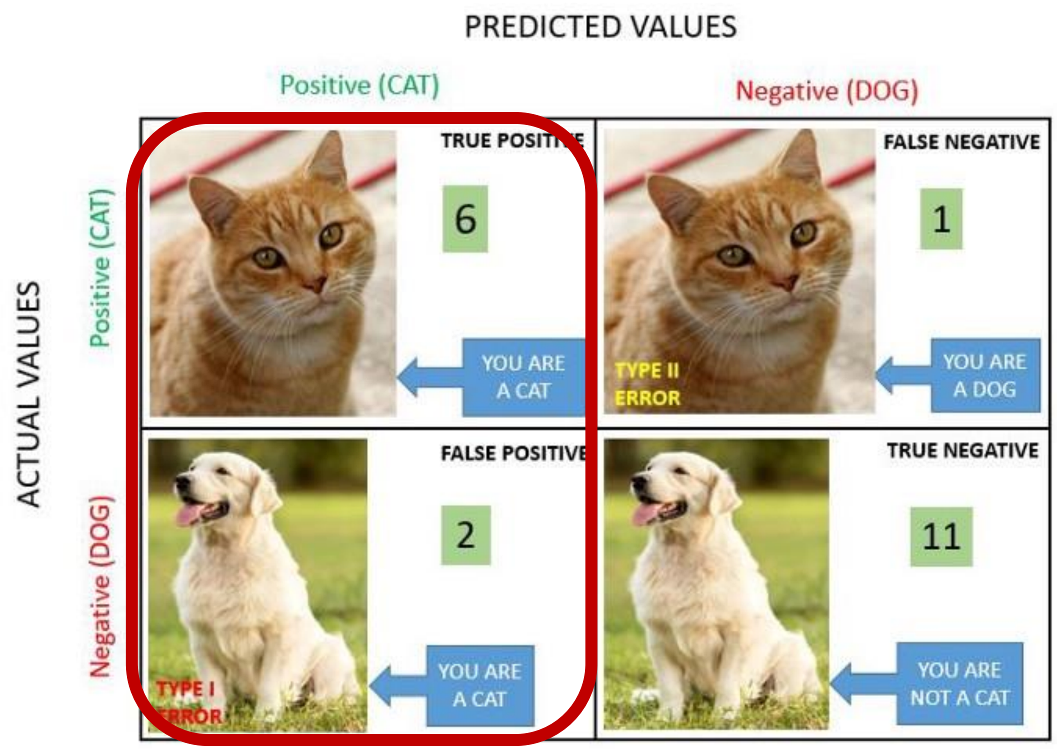


[4] Confusion matrix

Performance Metrics

Precision, 정밀도
Positive Predictive Value (PPV)



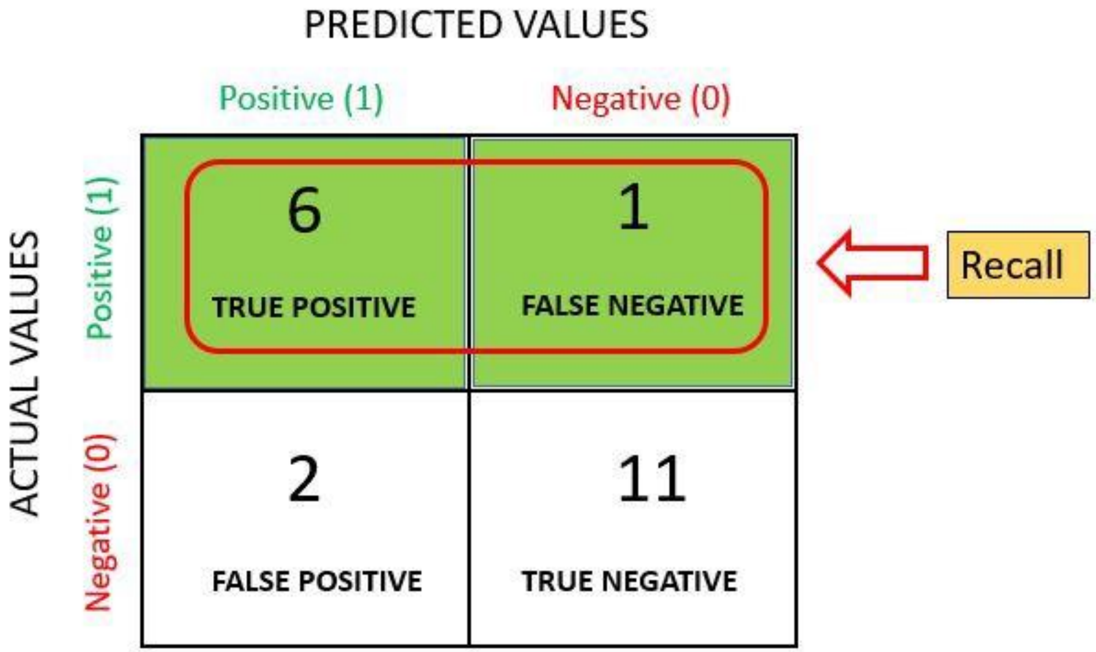


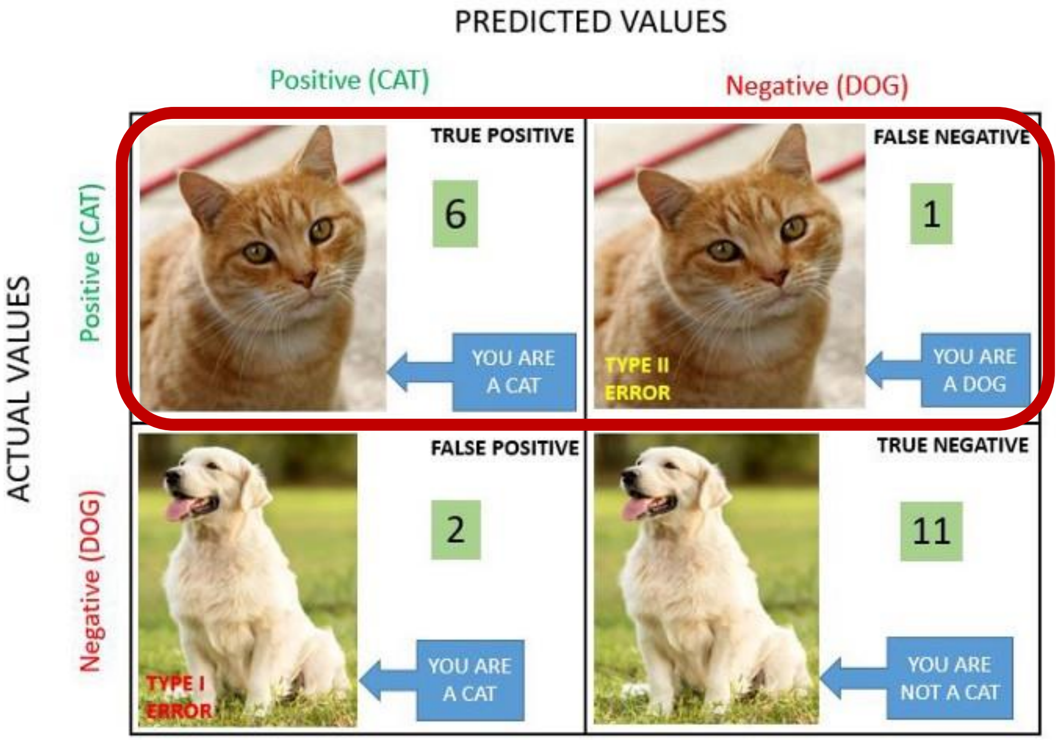
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{Predictions Actually Positive}}{\text{Total Predicted positive}} = \frac{6}{6+2} = 0.75$$

[4] Confusion matrix

Performance Metrics

Sensitivity(Recall), 민감도
True Positive Rate (TPR)



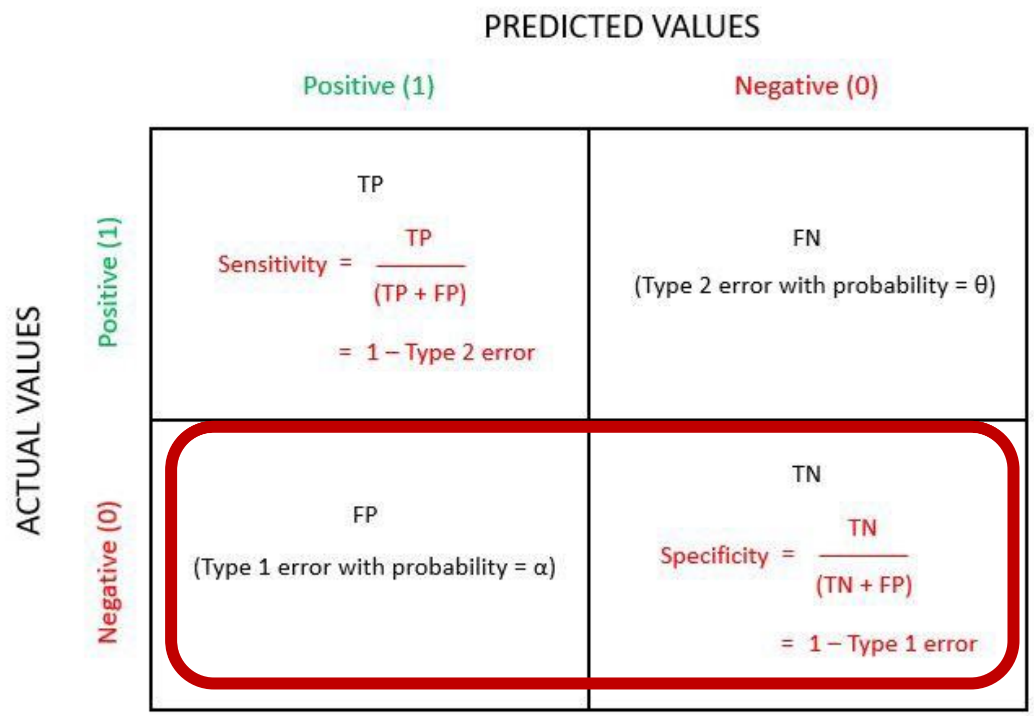


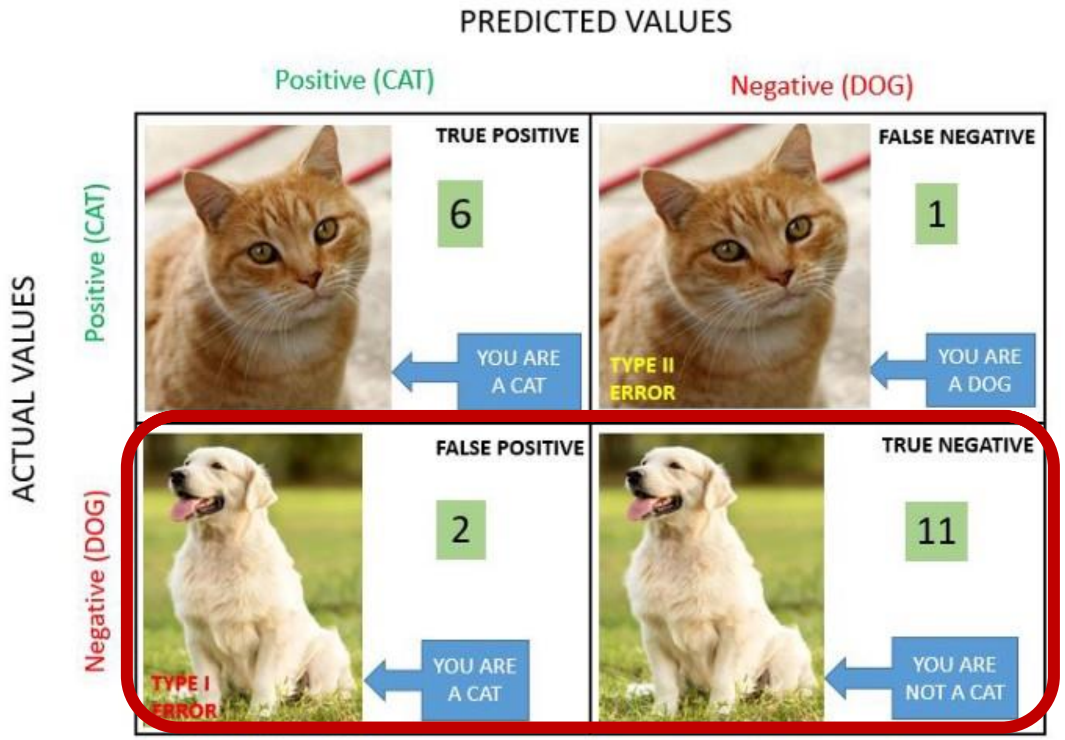
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Predictions Actually Positive}}{\text{Total Actual positive}} = \frac{6}{6+1} = 0.85$$

[4] Confusion matrix

Performance Metrics

Specificity, 특이도
True Negative Rate (TNR)



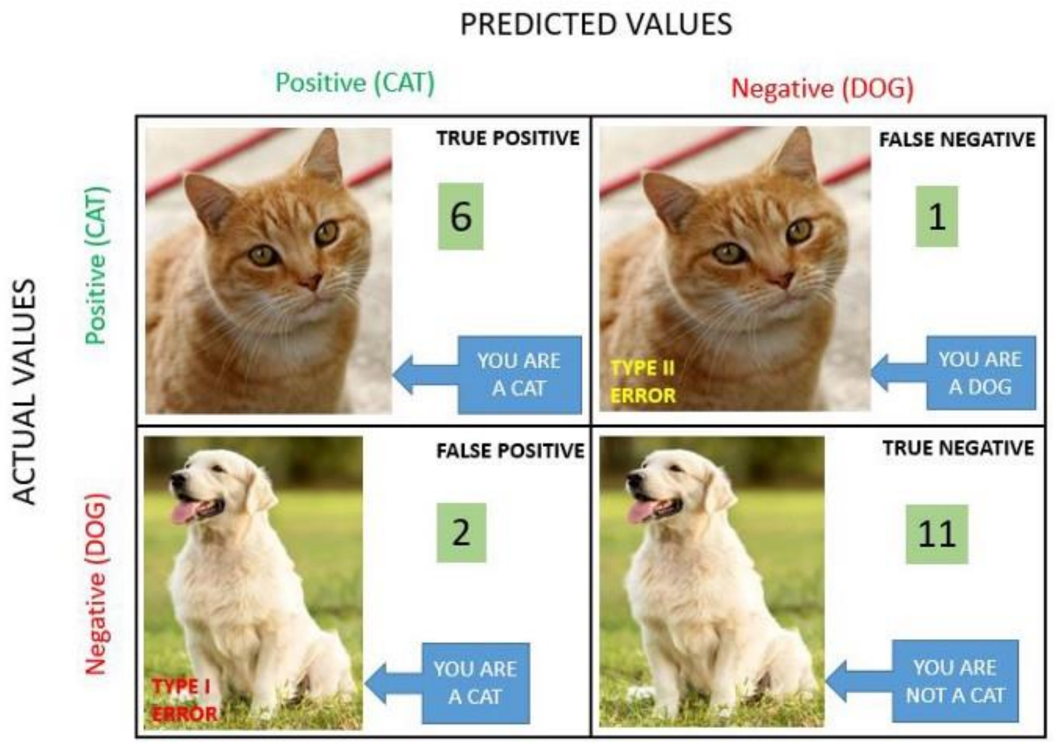


$$Specificity = \frac{11}{11+2} = \frac{11}{13} \approx 85\%$$

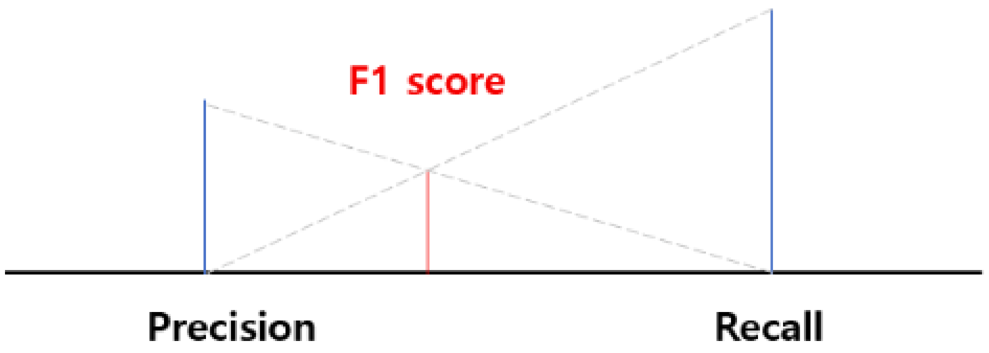
[4] Confusion matrix

Performance Metrics

F1-score



Harmonic mean

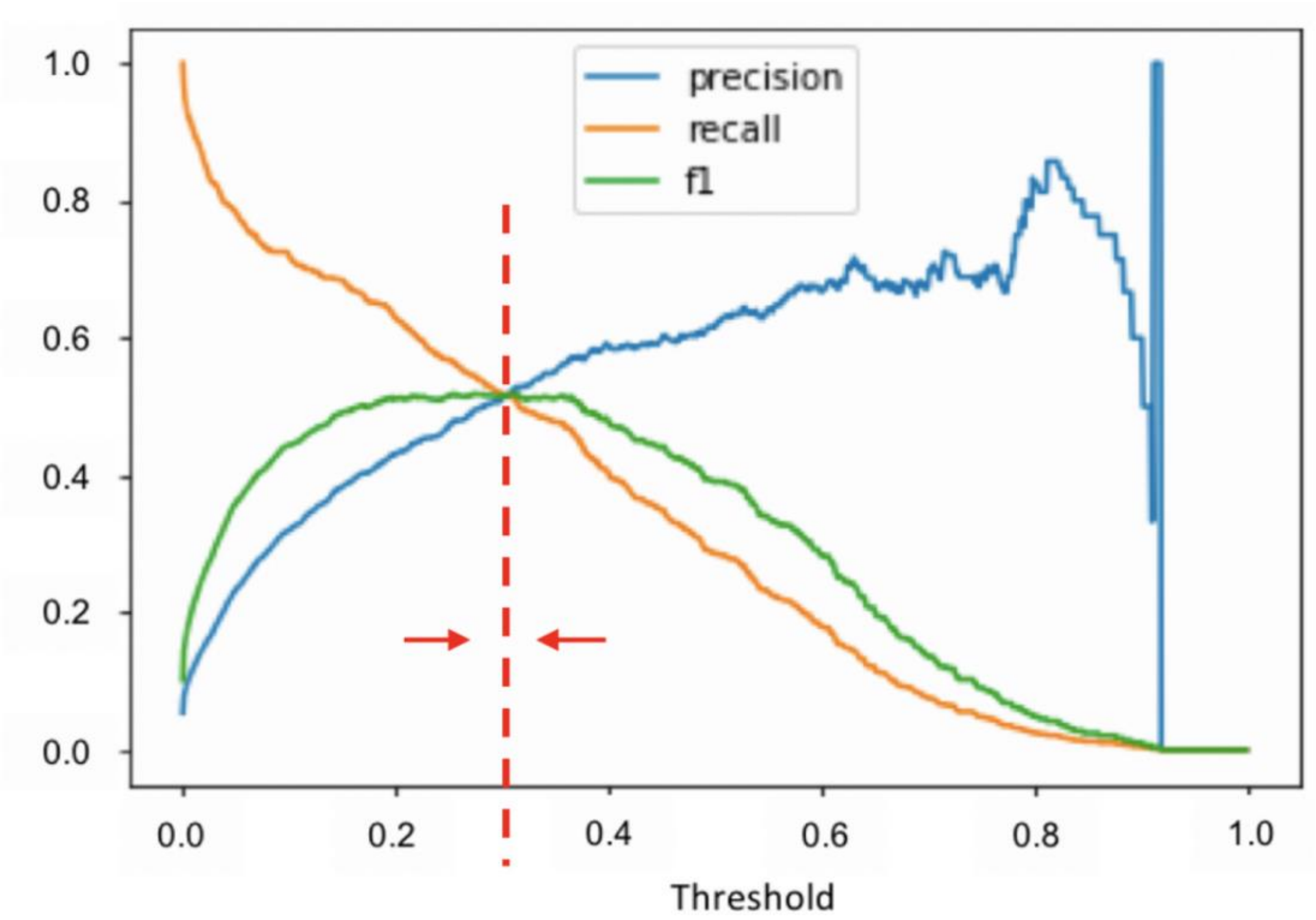


$$F1\_score = 2 \cdot \frac{1}{\frac{1}{Sensitivity} + \frac{1}{Precision}} = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}$$

[4] Confusion matrix
[5] Harmonic mean

Performance Metrics

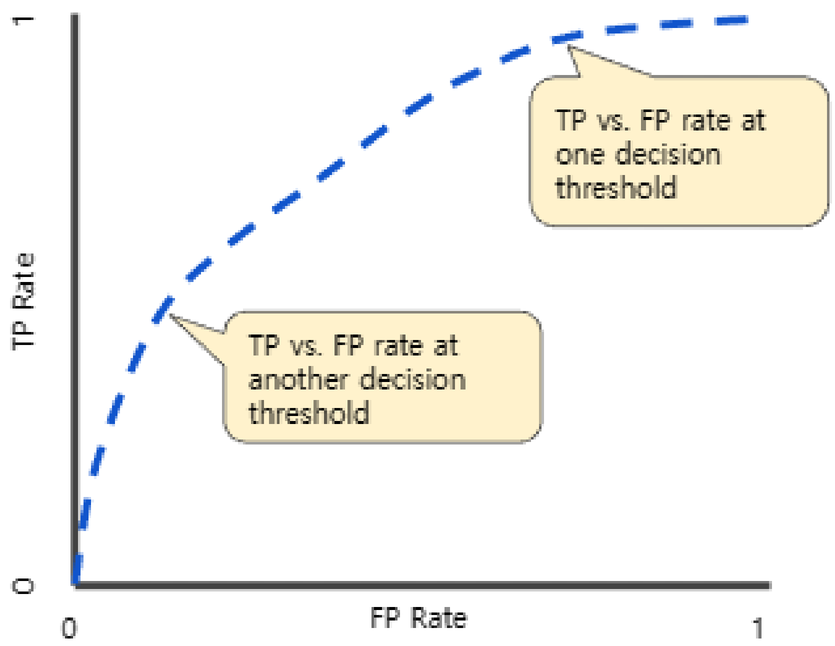
F1-score



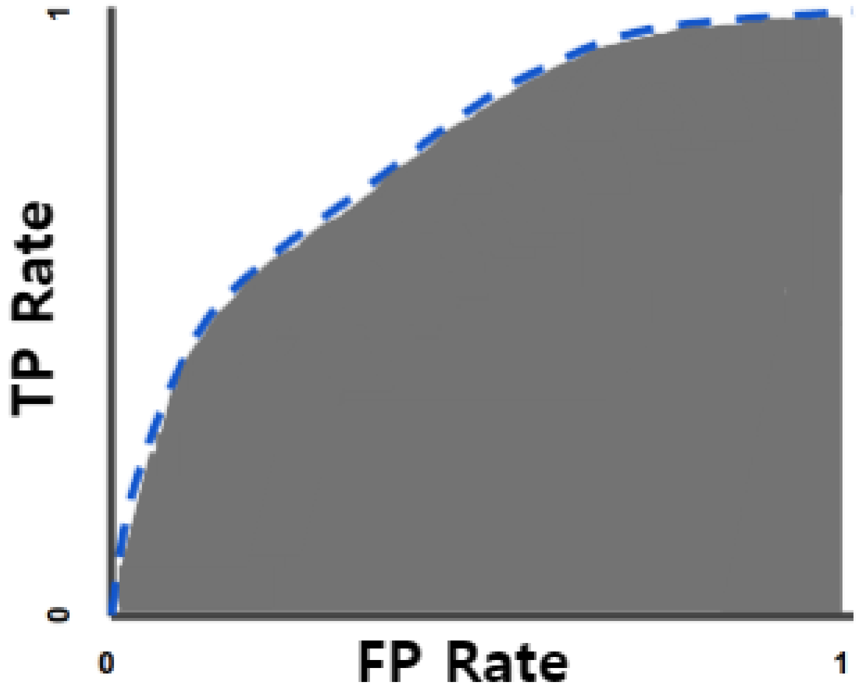
[4] Confusion matrix

Performance Metrics

AUC and ROC



ROC curve (receiver operating characteristic curve)



AUC (Area Under Curve)

[6] AUC and ROC

Performance Metrics

Why/When do we need to perform scaling?

PCA
Clustering (k-NN, K-means, DBSCAN, …)
Deep Neural Network

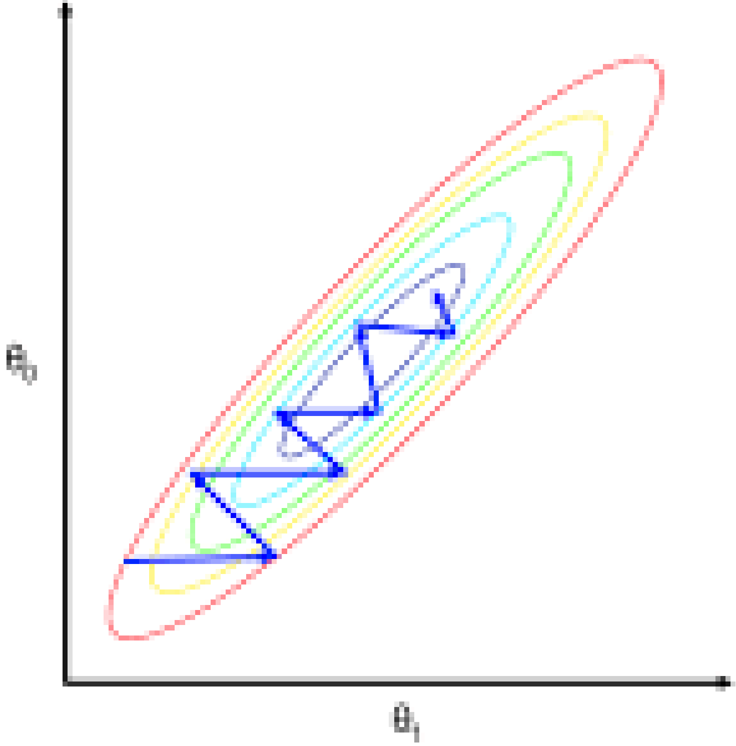
Tree based model
(Decision Tree, Random Forest, Boosting, …)

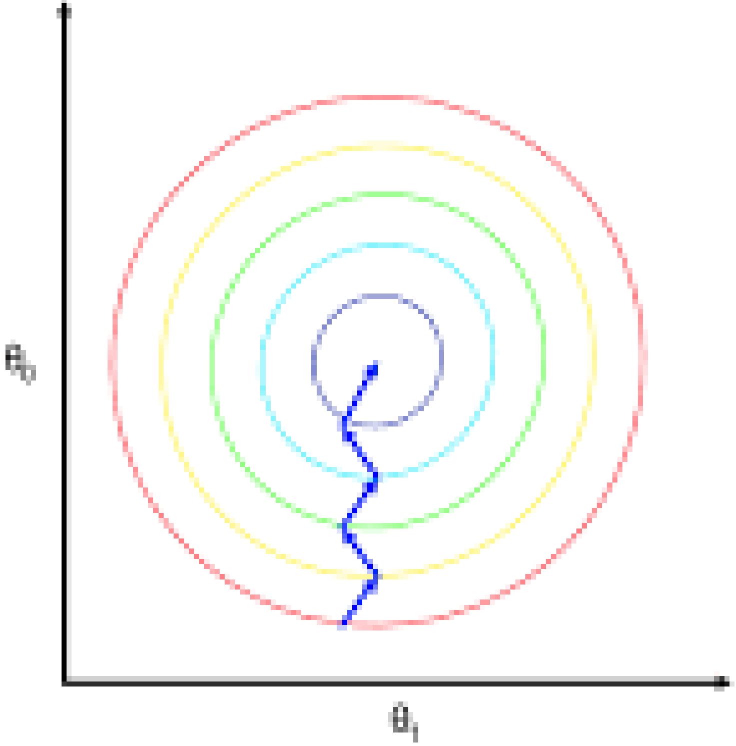
Distance-dependent
Need to be scaled!

Distance-independent
Doesn't need to be scaled!

Performance Metrics

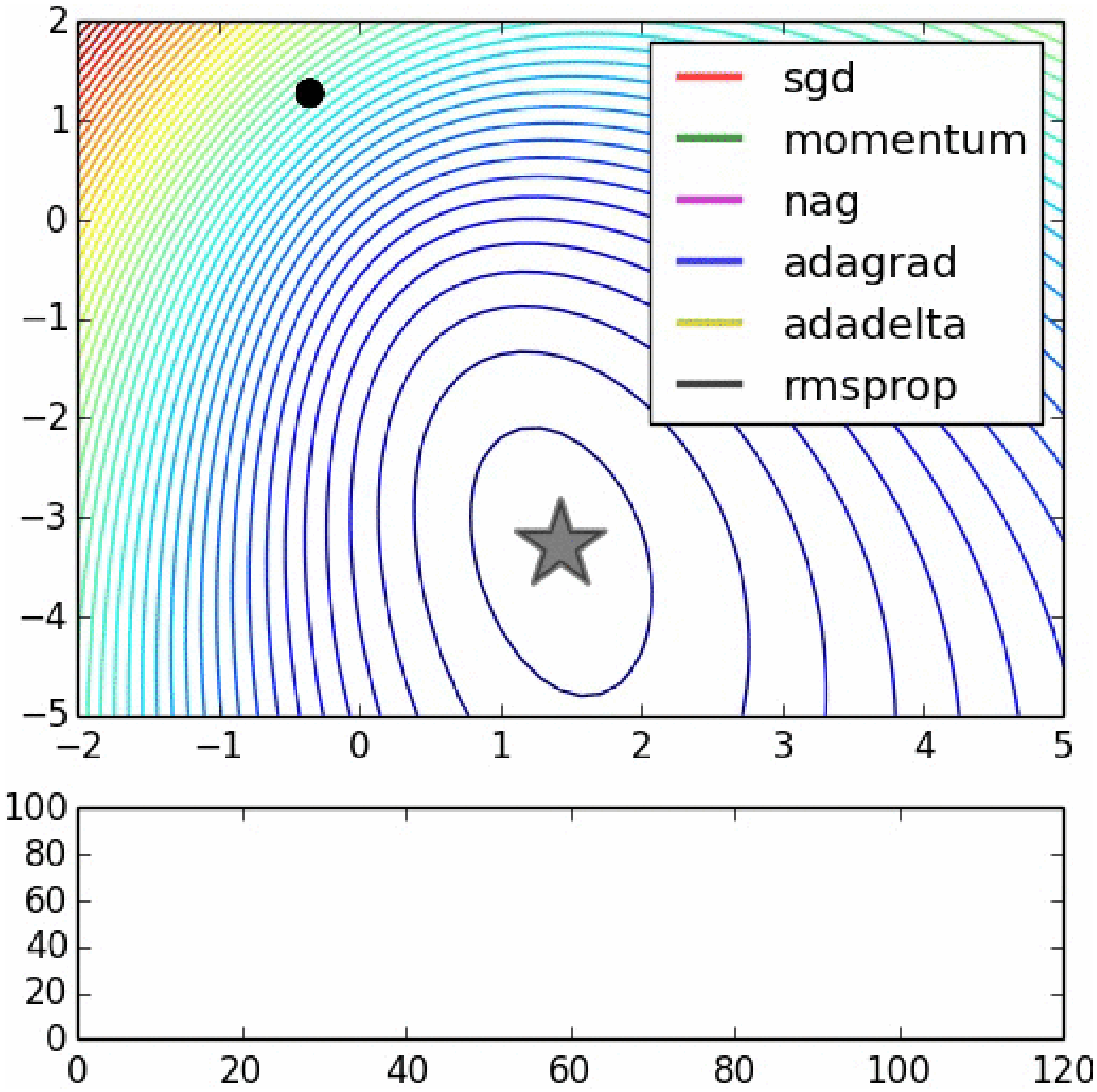
Data Scaling





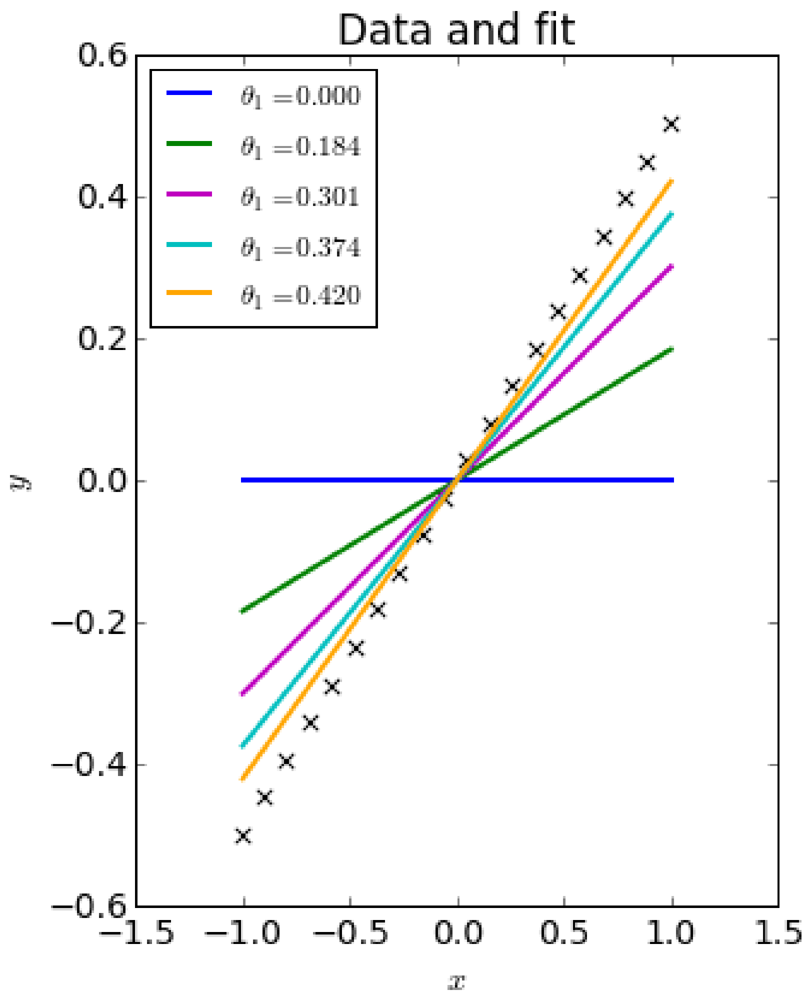
Performance Metrics

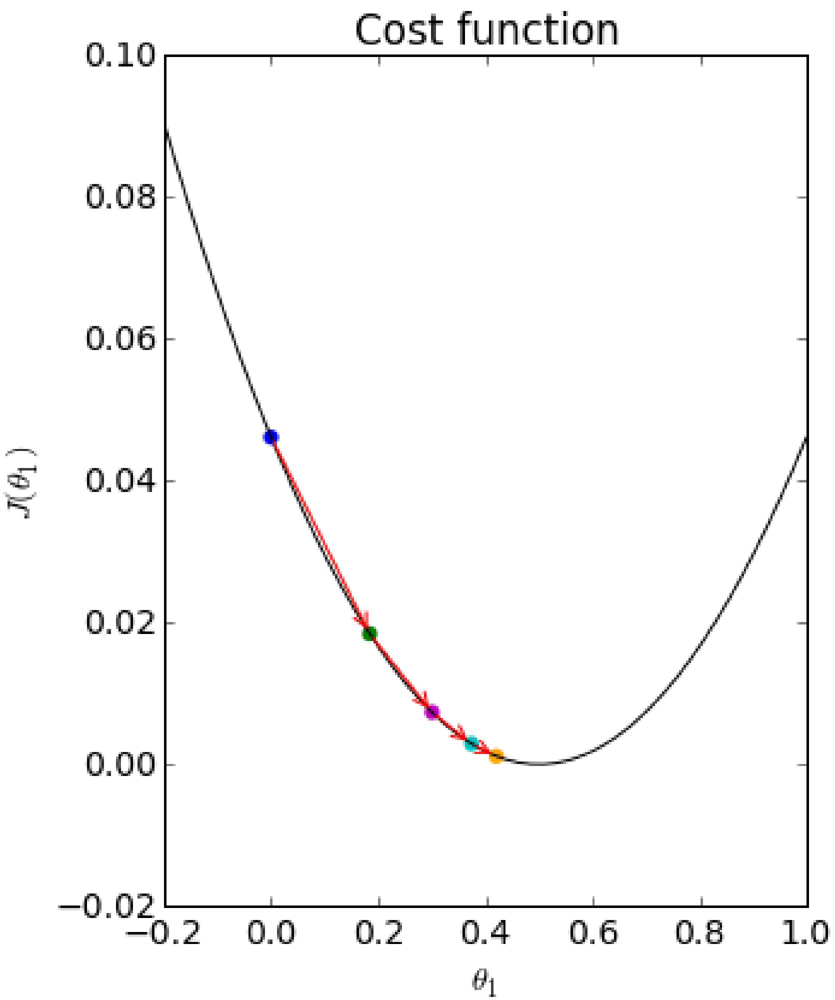
Data Scaling



Performance Metrics

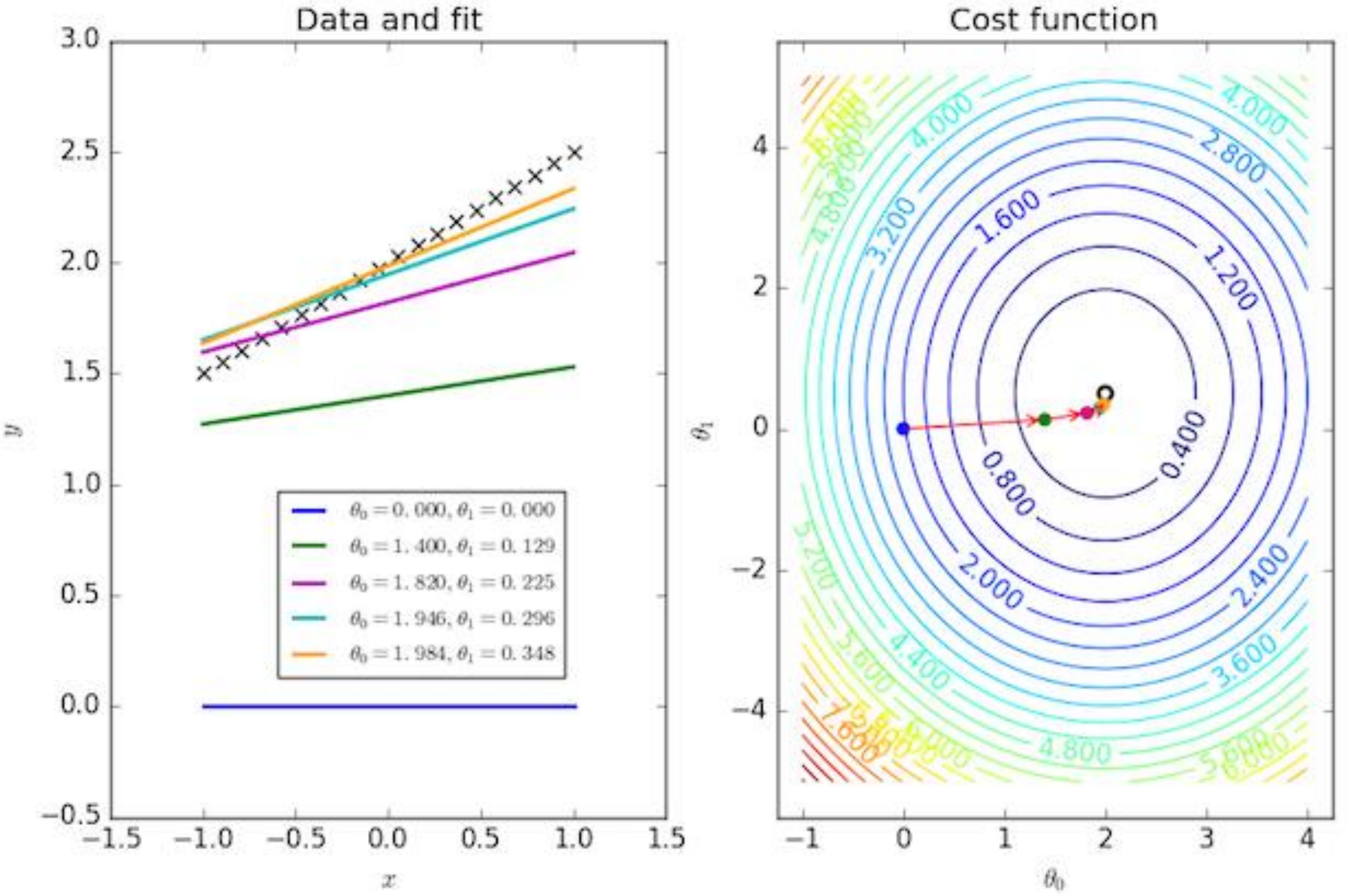
Data Scaling





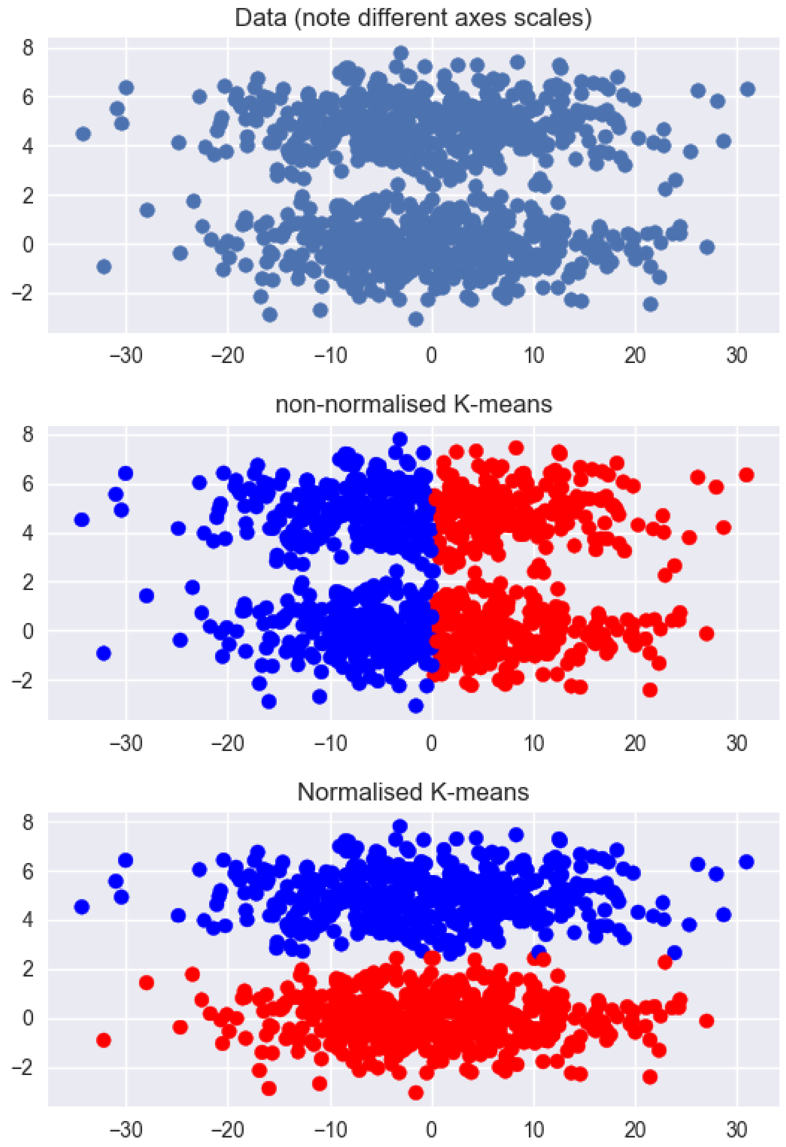
Performance Metrics

Data Scaling



Performance Metrics

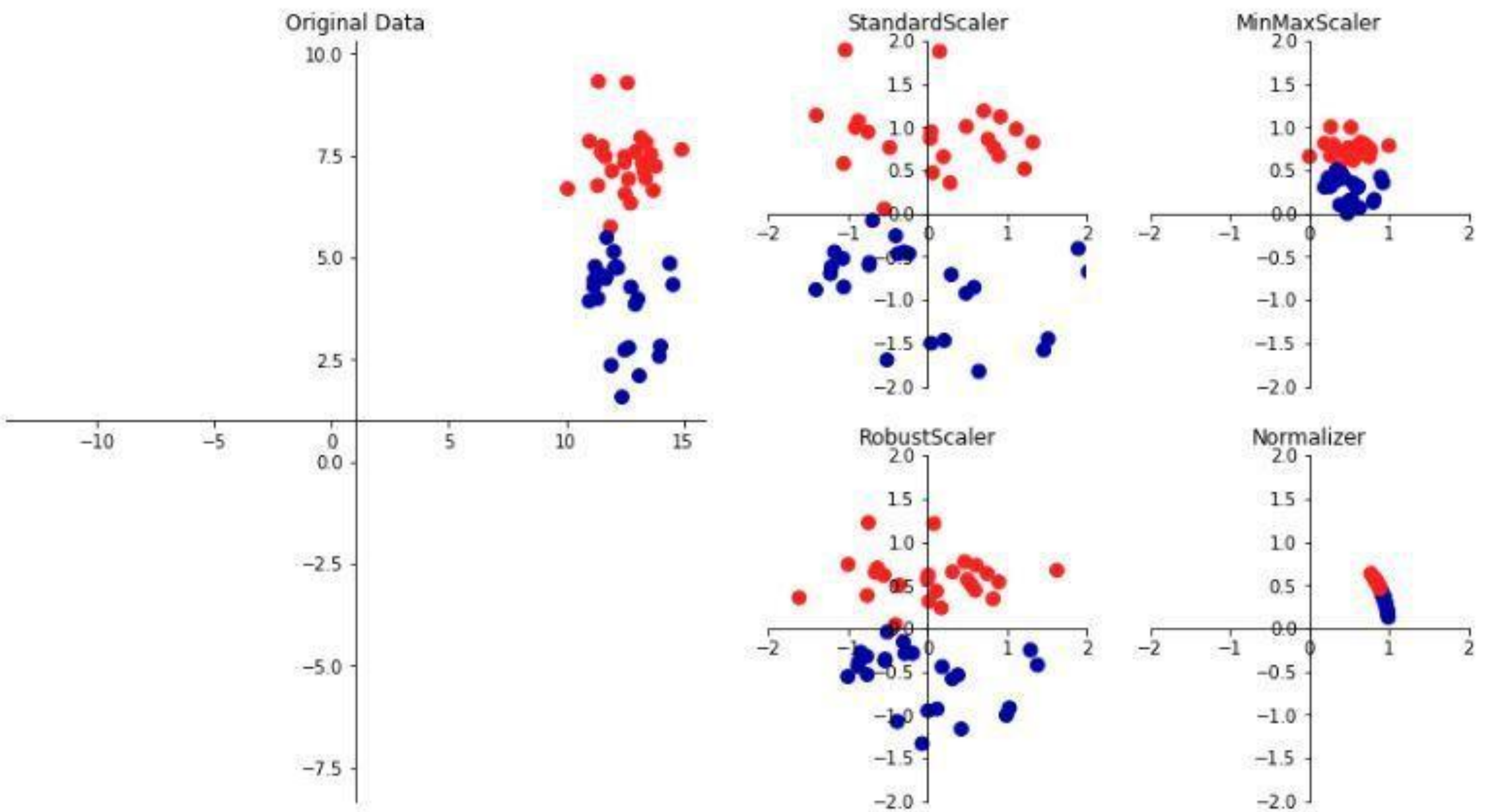
Data Scaling



Performance Metrics

Normalization

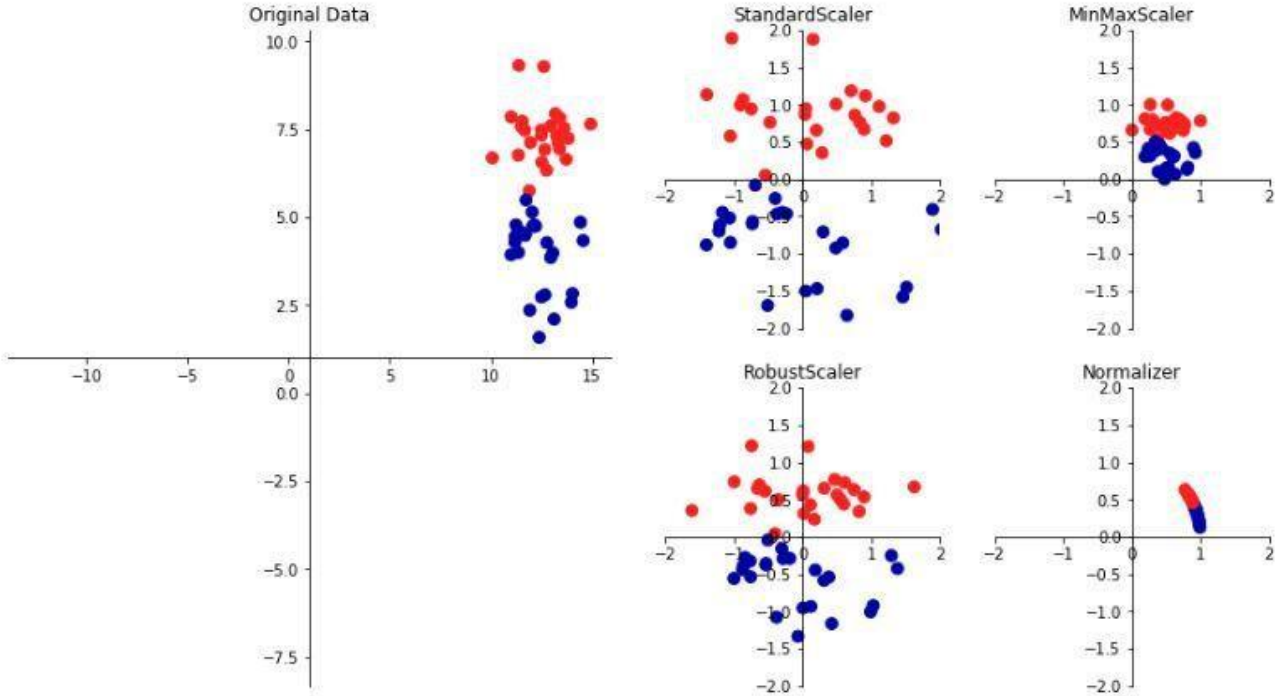
```
mglearn.plots.plot_scaling()
```



Performance Metrics

Normalization

```
mglearn.plots.plot_scaling()
```



$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Minmax scaling

$$z = \frac{x - \mu}{\sigma}$$

Standard scaling

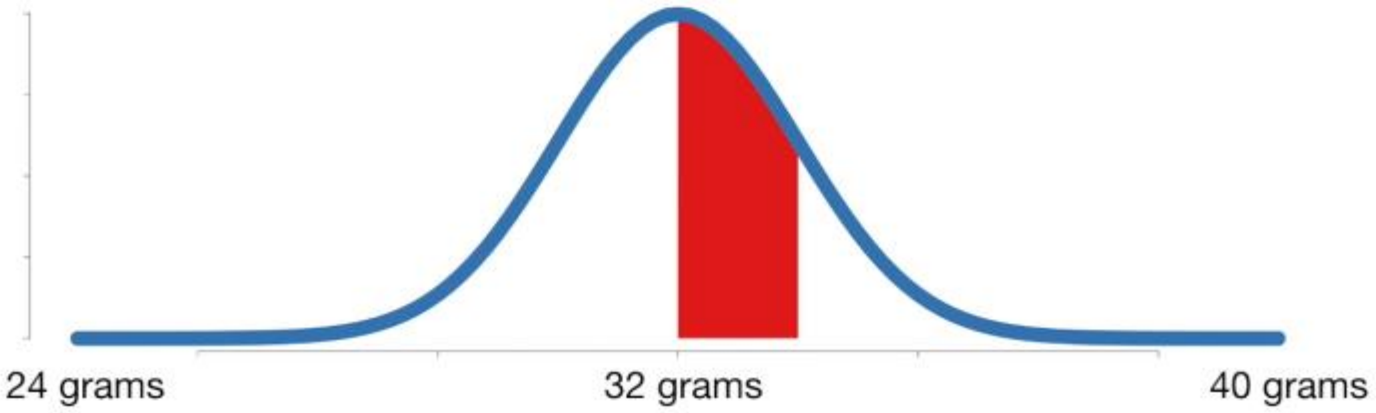
Concept of Likelihood

Probability vs Likelihood

$$Probability = P(X|D)$$

X: Observed
value
D: Distribution

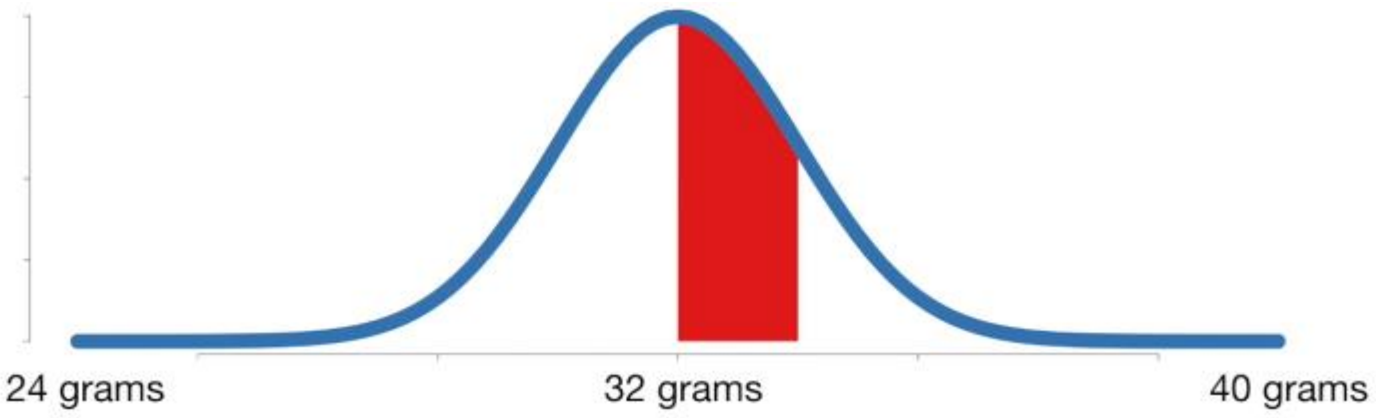
$$P(weight\ between\ 32\ and\ 34\ grams|mean = 32\ and\ stdev = 2.5)$$



[7] Probability, Likelihood and Likelihood Maximization

Concept of Likelihood

Probability vs Likelihood



$$P(weight\ between\ 32\ and\ 34\ grams|mean = 32\ and\ stdev = 2.5)$$

Probability
Area under distribution (probability that SOMETHING will be observed)
with 'specified distribution'

== Distribution is fixed
& Observation is variable!

[7] Probability, Likelihood and Likelihood Maximization

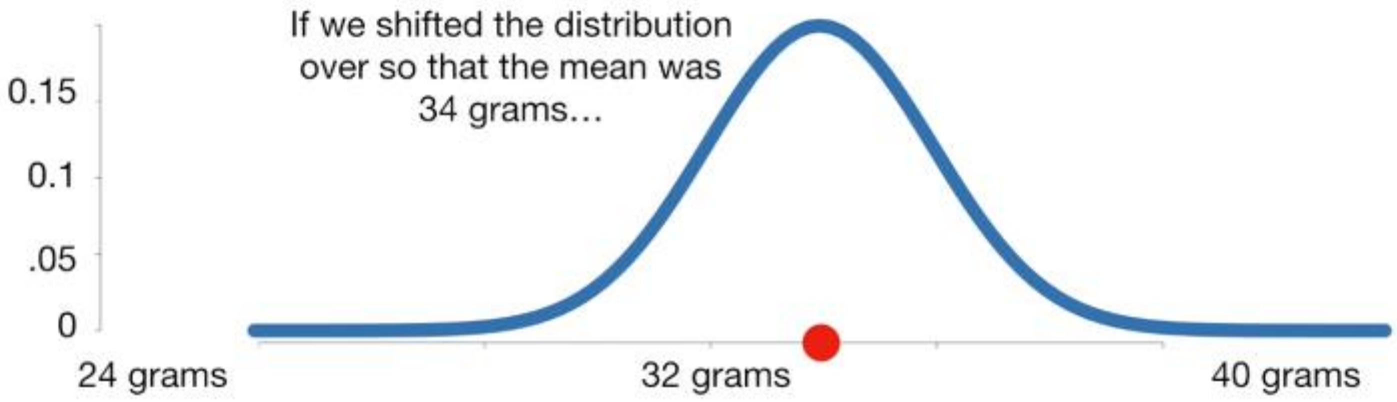
Concept of Likelihood

Probability vs Likelihood

$$Likelihood = L(D|X)$$

D: Distribution
X: Observed value

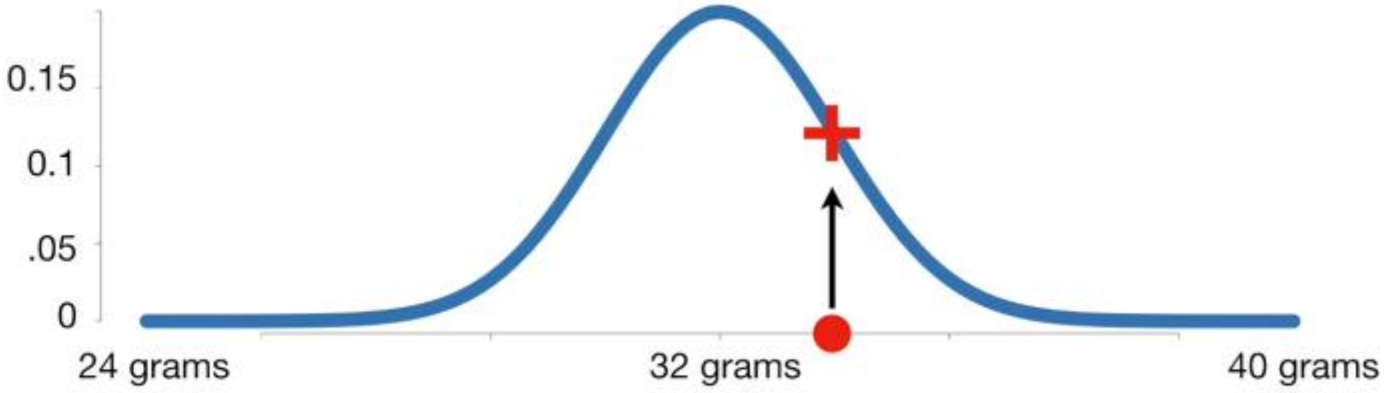
$$L(mean = 34\ and\ stdev = 2.5|weight = 34)$$



[7] Probability, Likelihood and Likelihood Maximization

Concept of Likelihood

Probability vs Likelihood



$$L(mean = 32\ and\ stdev = 2.5 | weight = 34)$$

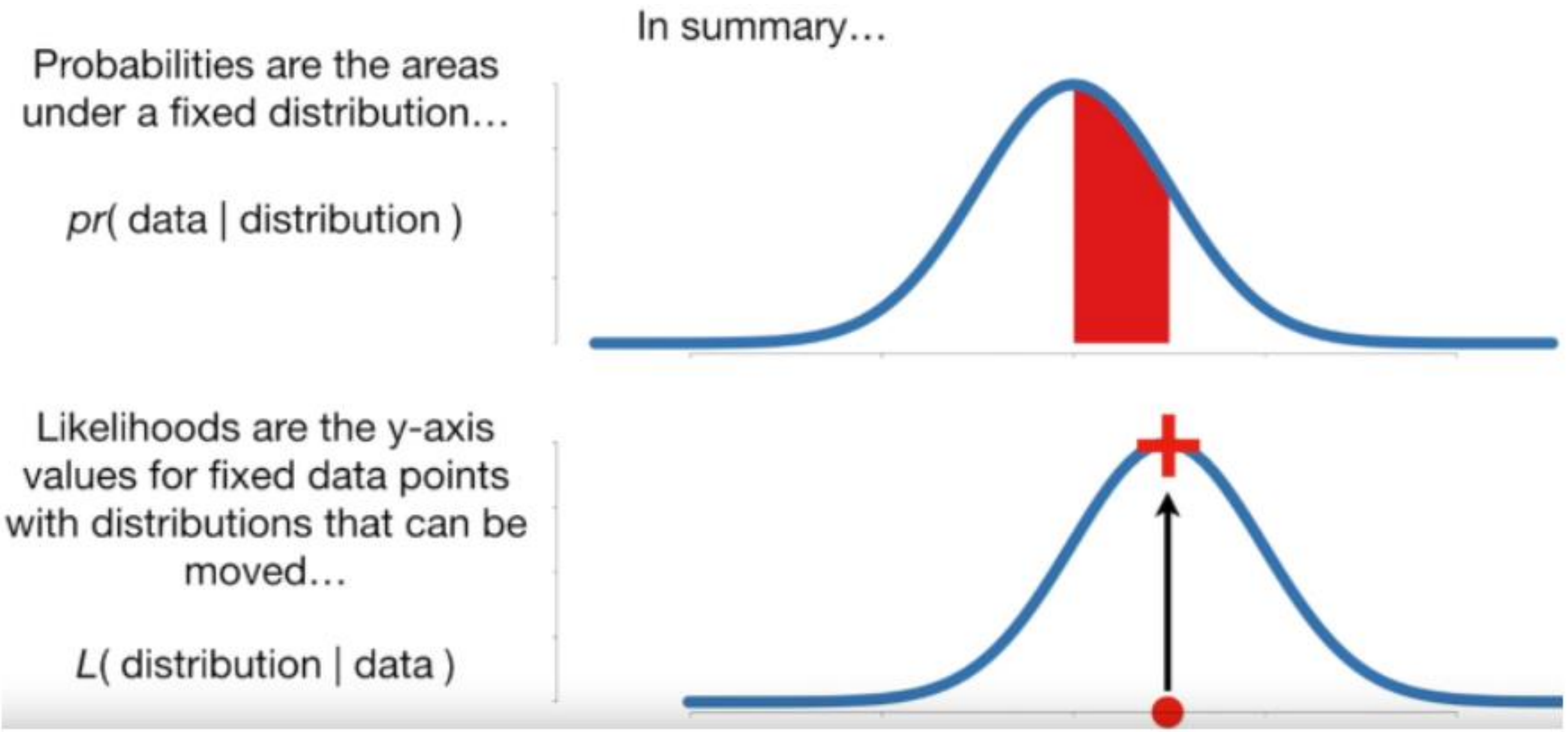
Likelihood
A probability that the value sampled from a given observation came from that probability distribution

== Observation is fixed
& Distribution is variable!

[7] Probability, Likelihood and Likelihood Maximization

Concept of Likelihood

Summary



Probability: Observation given Distribution (Distribution is fixed)
Likelihood: Distribution given Observation (Data is fixed)

[8] An animation for explanation of Likelihood

[7] Probability, Likelihood and Likelihood Maximization

Bayes Theorem

Conditional Probabilistic approach

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Prior Probability of $\boldsymbol{H}$

$$P(H)$$

Don't have any information about E

Posterior Probability given $\boldsymbol{E}$

$$P(H|E)$$

Conditional Probability

$$P(E|H)$$

$L(H|E) = P(E|H)$ is a likelihood of H given E

H: Hypothesis (가설, 사건)
E: Evidence (새로운 정보)

[9] Bayes' theorem

Bayes Theorem

Conditional Probabilistic approach

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Let's assume that we want to know H,
but what we have is only E and $P(E|H)$.

We can get $P(H|E)$ using Bayes Theorem!

H: Hypothesis (가설, 사건)
E: Evidence (새로운 정보)

[9] Bayes' theorem

Bayes Theorem

Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Incidence of disease A(발병률): 0.1% (0.001)
Probability of detecting the disease when the disease actually exists(민감도): 99% (0.99)
Probability of not detecting the disease when the disease is not present(특이도): 98% (0.98)

What is $P(H|E) = ?$

H: Actually having a disease
E: Determined to have a disease

[10] Test reagent problem

Bayes Theorem

Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Incidence of disease A(발병률): 0.1% (0.001)
Probability of detecting the disease when the disease actually exists(민감도): 99% (0.99)
Probability of not detecting the disease when the disease is not present(특이도): 98% (0.98)

$P(H) = 0.001 = Incidence\ of\ getting\ disease\ A$
$P(E|H) = 0.99 = Actually\ having\ a\ disease, determined\ to\ have\ a\ disease$(True Positive)
$P(E^c|H^c) = 0.98 = Actually\ not\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$(True Negative)

H: Actually having a disease
E: Determined to have a disease

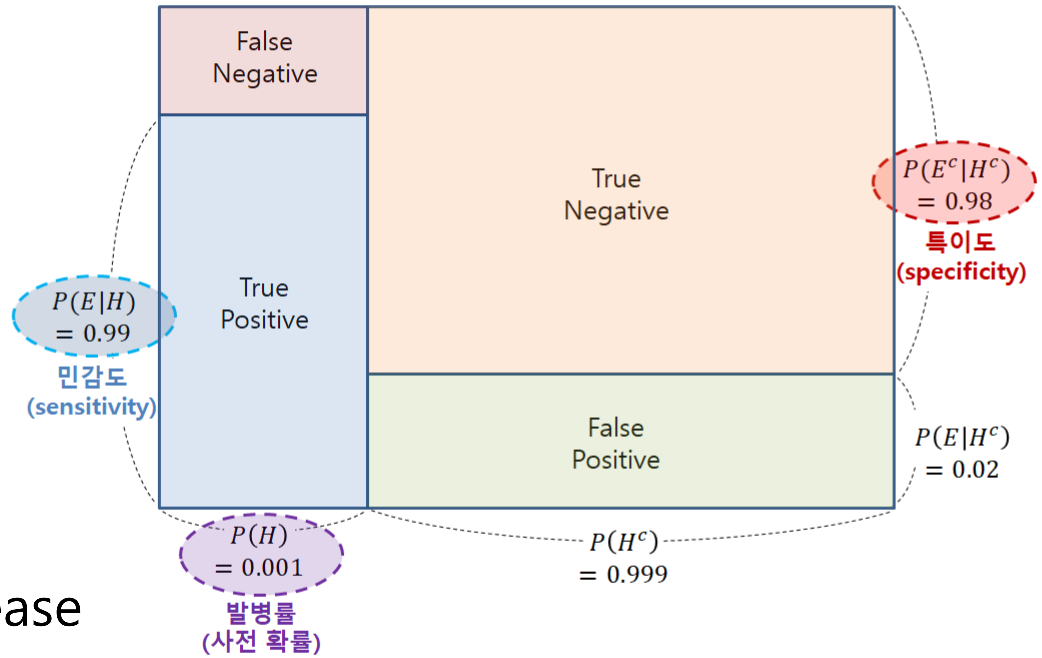
[10] Test reagent problem

Bayes Theorem

Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

$P(H) = 0.001 = Incidence\ of\ getting\ disease\ A$
$P(E|H) = 0.99 = Actually\ having\ a\ disease, determined\ to\ have\ a\ disease$(True Positive)
$P(E^c|H^c) = 0.98 = Actually\ not\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$(True Negative)



H: Actually having a disease
E: Determined to have a disease

[10] Test reagent problem

Bayes Theorem

Example

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)}$$

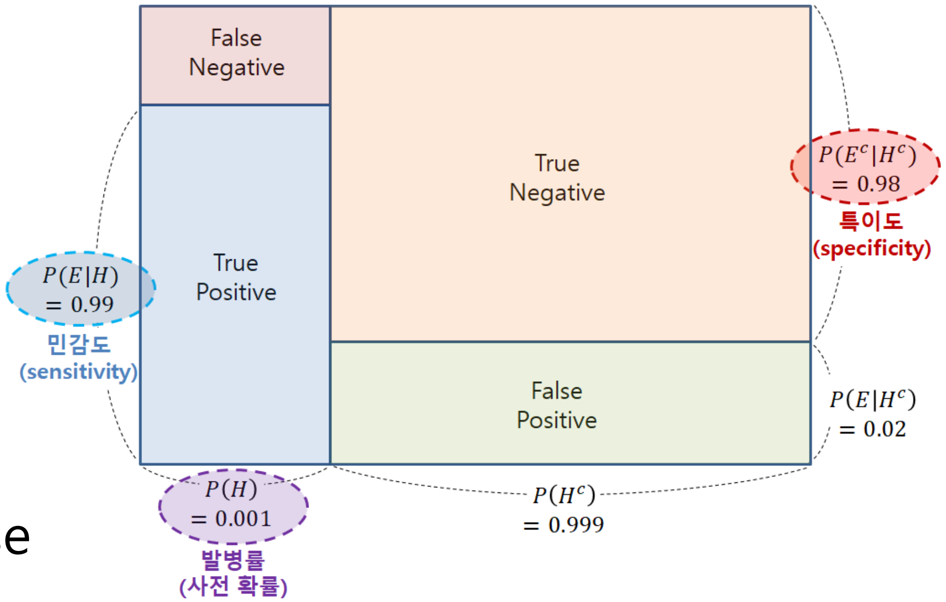
$P(H) = 0.001 = Incidence\ of\ getting\ disease\ A$

$P(H^c) = 0.999 = Incidence\ of\ not\ getting\ disease\ A$

$P(E|H) = 0.99 = Actually\ having\ a\ disease, determined\ to\ have\ a\ disease$(True Positive)

$P(E^c|H^c) = 0.98 = Actually\ not\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$(True Negative)

$P(E|H^c) = 0.02 = Actually\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$ (False Positive)



H: Actually having a disease

E: Determined to have a disease

[10] Test reagent problem

Bayes Theorem

Example

$P(H) = 0.001 = Incidence\ of\ getting\ disease\ A$
$P(H^c) = 0.999 = Incidence\ of\ not\ getting\ disease\ A$
$P(E|H) = 0.99 = Actually\ having\ a\ disease, determined\ to\ have\ a\ disease$(True Positive)
$P(E^c|H^c) = 0.98 = Actually\ not\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$(True Negative)
$P(E|H^c) = 0.02 = Actually\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$ (False Positive)

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)} = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.02 \times 0.999}$$
$$\approx 0.047 = 4.7\% = Determined\ to\ have\ a\ disease, actually\ having\ a\ disease$$
$$\neq P(H) \times P(E) = 0.001998\%$$

H: Actually having a disease
E: Determined to have a disease

[10] Test reagent problem

Bayes Theorem

Practice

If a person who has already tested positive is tested again and tested positive again, what is the probability that this person will actually get the disease?

$P(H) = 0.047 = Incidence\ of\ getting\ disease\ A$ – Posterior changed to Prior!
$P(H^c) = 0.953 = Incidence\ of\ not\ getting\ disease\ A$
$P(E|H) = 0.99 = Actually\ having\ a\ disease, determined\ to\ have\ a\ disease$(True Positive)
$P(E^c|H^c) = 0.98 = Actually\ not\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$(True Negative)
$P(E|H^c) = 0.02 = Actually\ having\ a\ disease, determined\ not\ to\ have\ a\ disease$ (False Positive)

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)} = \frac{0.99 \times 0.047}{0.99 \times 0.047 + 0.02 \times 0.953}$$
$$\approx 0.709 \approx 71\%$$

H: Actually having a disease
E: Determined to have a disease

[10] Test reagent problem

Exercise

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-max scaling

$$z = \frac{x_i - \mu}{\sigma}$$

Standard scaling

Data: [-1, 1, 3, 5, 7, 9]

Hint
Mean: 4
Variance(sigma^2): 12

Calculate the scaled result

Exercise

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-max scaling

$$z = \frac{x_i - \mu}{\sigma}$$

Standard scaling

Data: [-1, 1, 3, 5, 7, 9]

Hint
Mean: 4
Variance(sigma^2): 12

Calculate the scaled result

Exercise

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-max scaling

Data: [-1, 1, 3, 5, 7, 9]

$$z = \frac{x_i - \mu}{\sigma}$$

Standard scaling

Hint
Mean: 4
Variance(sigma^2): 12

Data: [-1, 1, 3, 5, 7, 9]