

SUPERVISED LEARNING - CLASSIFICATION

JENS BAETENS

GLOSSARY

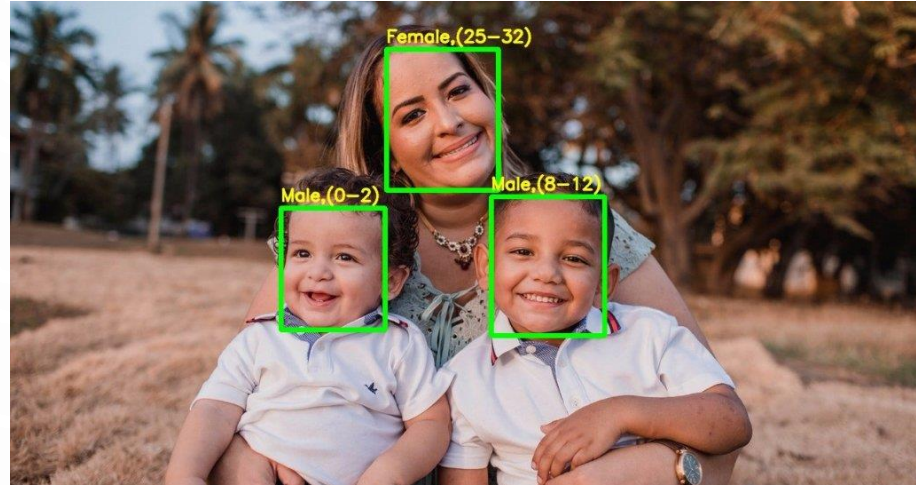
- Supervised
- Unsupervised
- Reinforcement Learning
- Regression
- Overfitting
- Underfitting
- Learning Rate
- Loss Function
- Feature Engineering
- Normalisation
- Regularisation
- Trainen van een model

WAT IS CLASSIFICATIE?

Supervised learning

Input omzetten naar klasse

Classifier genoemd



WAT IS CLASSIFICATIE?

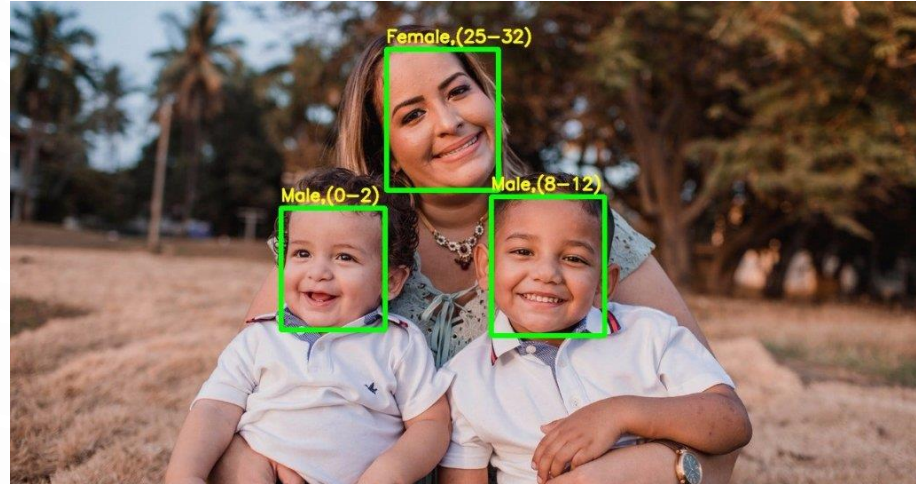
Gezichtsherkenning

Geschriftherkenning

Spam detectie

Kwaliteitscontroles

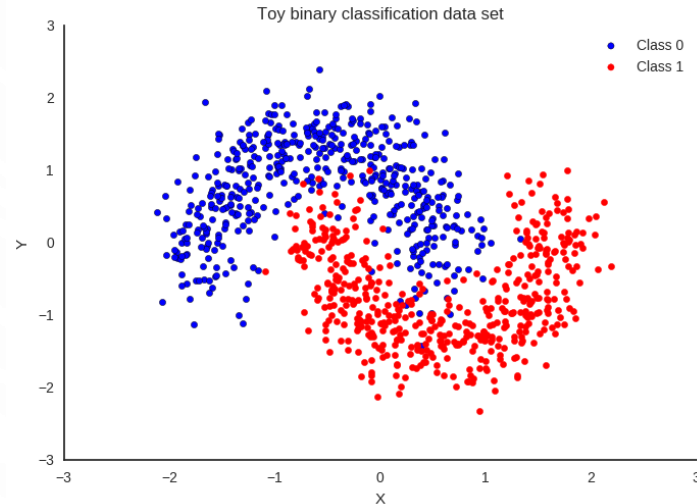
Medische diagnoses



TYPES CLASSIFIERS - BINARY

Twee verschillende klassen

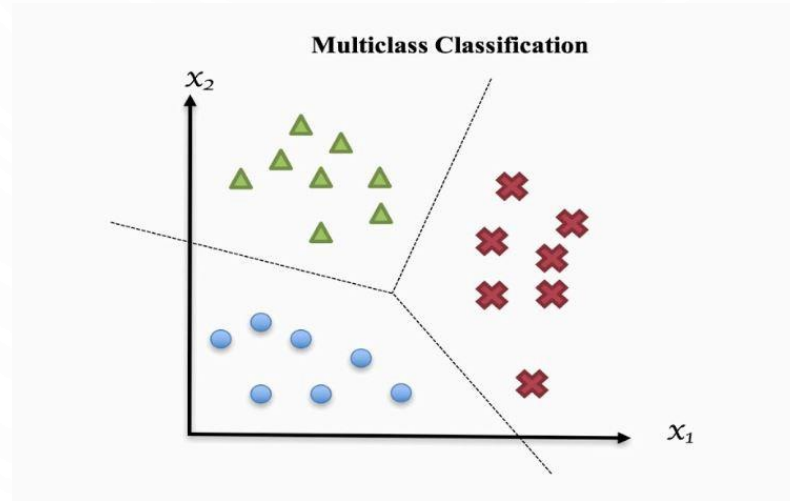
Voorbeeld: Goede of slechte kwaliteit, man of vrouw, Goed- of kwaadaardig



TYPES CLASSIFIERS - MULTICLASS

$N > 2$ verschillende klassen (maar 1 mogelijk voor elke input)

Voorbeeld: Gezichtsherkenning (1 klasse per persoon), Hondenrasherkenning, ...



TYPES CLASSIFIERS - MULTILABEL

$N > 2$ verschillende klassen maar meerdere mogelijk per input

Voorbeeld: Beeldherkenning, Meerdere genres mogelijk voor een film, ...

Binary
Classification



- Spam
- Not spam

Multiclass
Classification



- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label
Classification

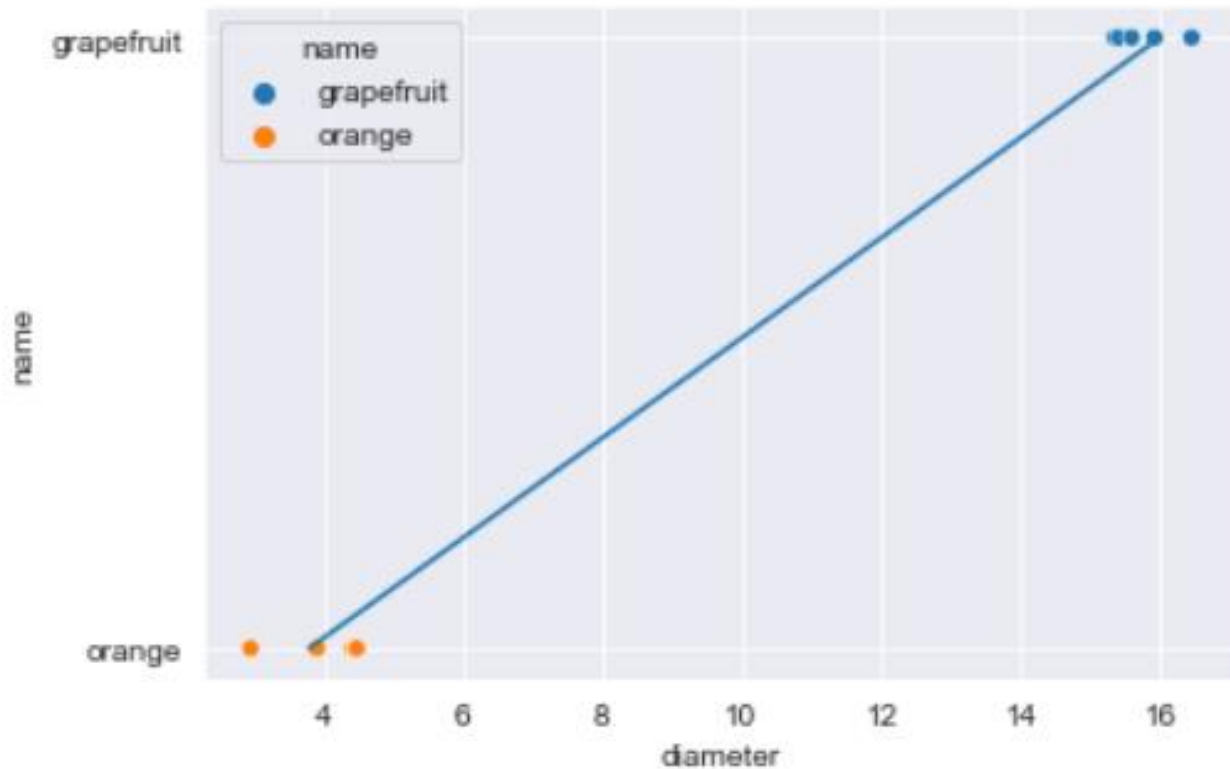


- Dog
- Cat
- Horse
- Fish
- Bird
- ...

KAN HET MET LINEAIRE REGRESSIE?

		name	diameter	weight	red	green	blue
grapefruit	9995	grapefruit	15.35	253.89	149	77	20
	9996	grapefruit	15.41	254.67	148	68	7
	9997	grapefruit	15.59	256.50	168	82	20
	9998	grapefruit	15.92	260.14	142	72	11
	9999	grapefruit	16.45	261.51	152	74	2
orange	0	orange	2.96	86.76	172	85	2
	1	orange	3.91	88.05	166	78	3
	2	orange	4.42	95.17	156	81	2
	3	orange	4.47	95.60	163	81	4
	4	orange	4.48	95.76	161	72	9

KAN HET MET LINEAIRE REGRESSIE?

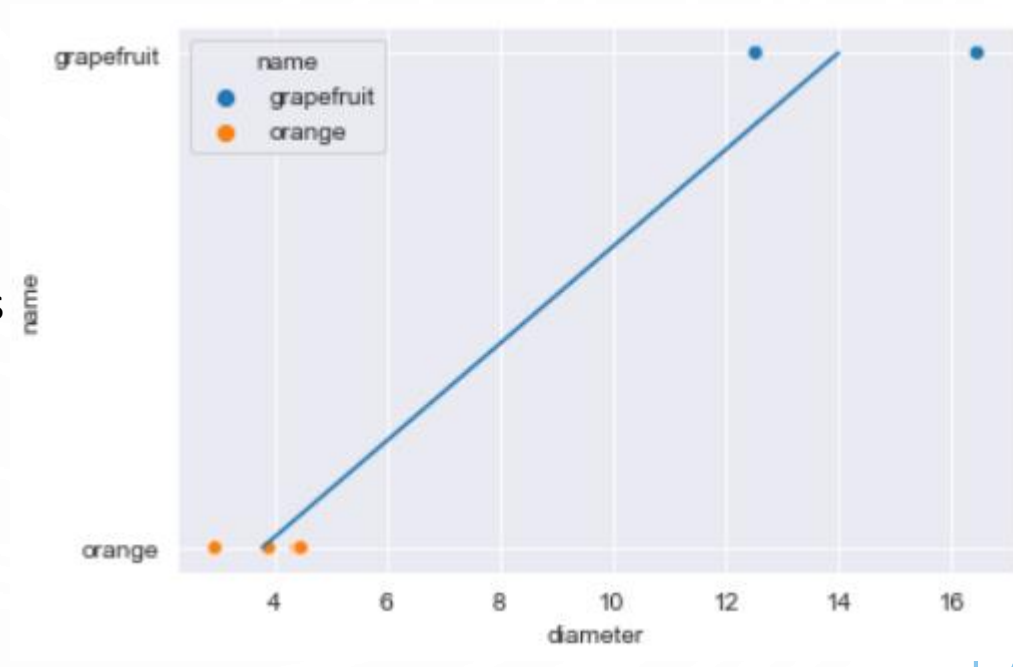


KAN HET MET LINEAIRE REGRESSIE?

Gevoelig voor outliers

Zeer breed “fuzzy” middenstuk

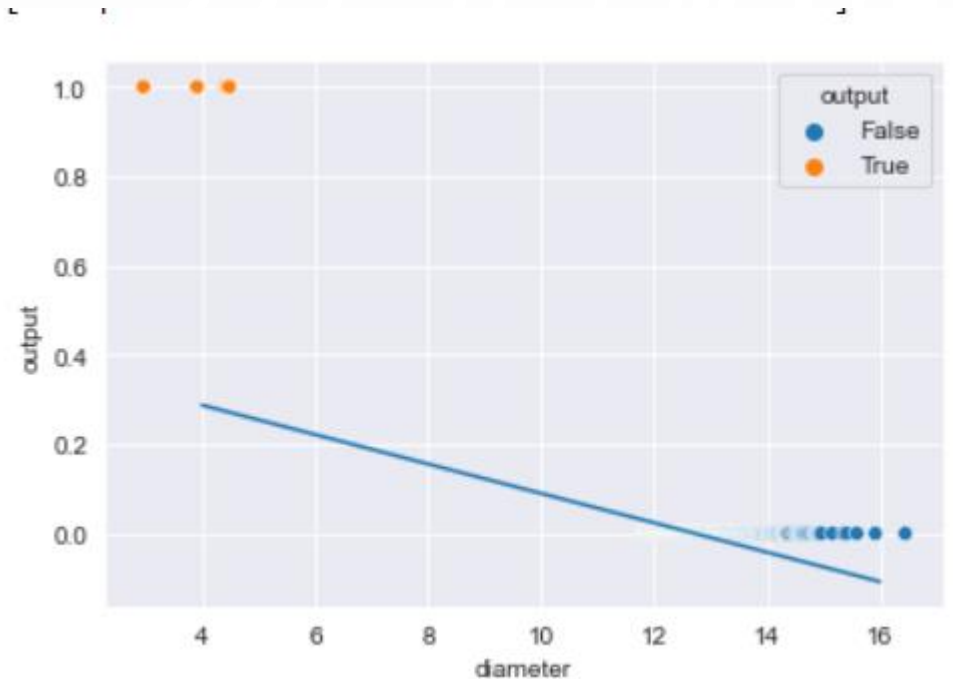
Komt niet overeen met een kans



KAN HET MET LINEAIRE REGRESSIE?

Ongebalanceerde klassen

=> Geen lineaire regressie mogelijk



CLASSIFICATIE – LOGISTIC REGRESSION

JENS BAETENS

LOGISTIC REGRESSION

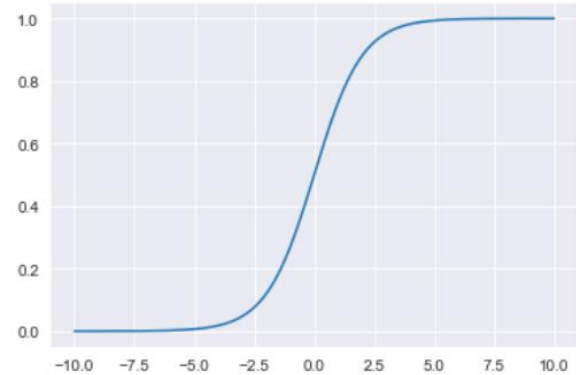
Logistische functie (sigmoid)

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

Geeft een waarde terug tussen 0 en 1
- De kans het tot de klasse hoort

$$f_{\mathbf{w}}(x) = \frac{1}{1+e^{-\mathbf{w}^T x}}$$

$$\mathbf{w}^T x = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_N x_N$$

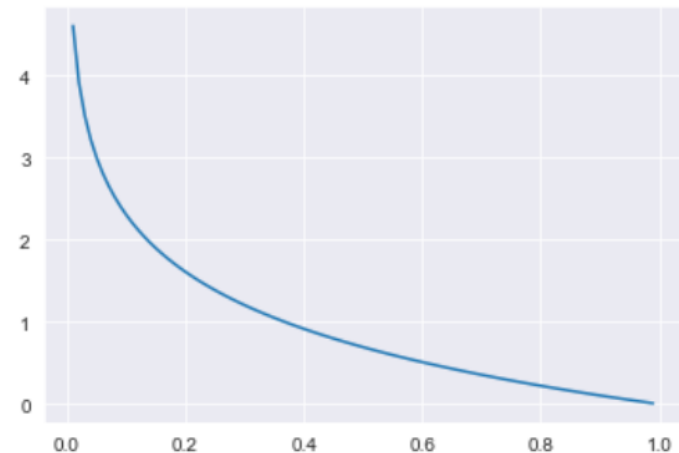


LOGISTIC REGRESSION

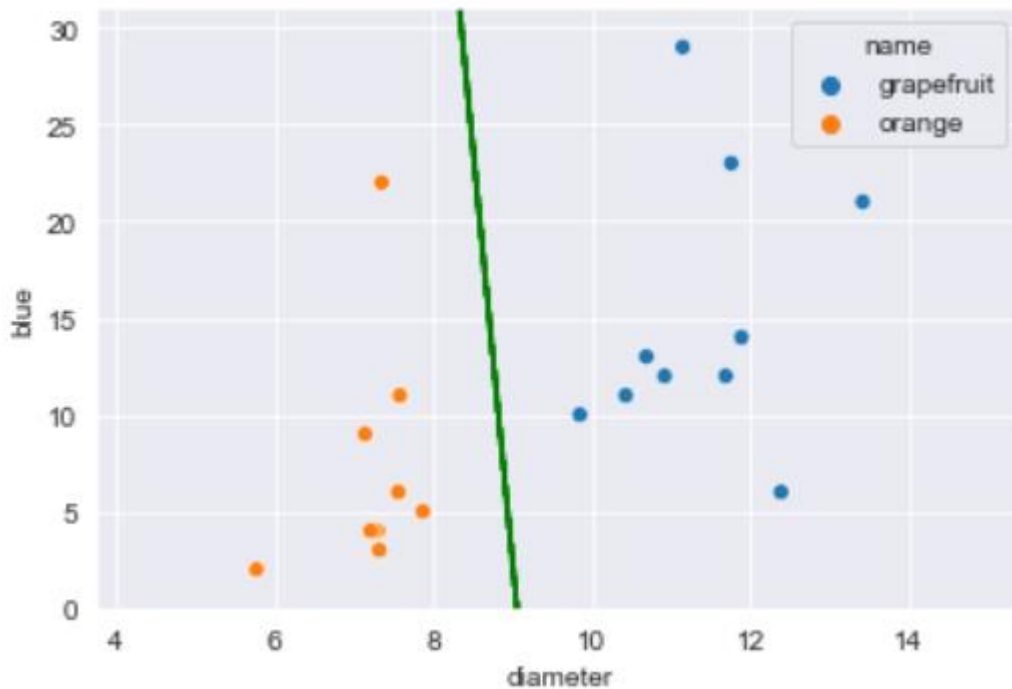
$$L(\mathbf{w}) = \begin{cases} -\ln(f_{\mathbf{w}}(x)) & \text{als } y = 1 \\ -\ln(1 - f_{\mathbf{w}}(x)) & \text{als } y = 0 \end{cases}$$

$$L(\mathbf{w}) = -\frac{1}{N} \left[\sum_{i=1}^N y_i \ln(f_{\mathbf{w}}(x_i)) + (1 - y_i) \ln(1 - f_{\mathbf{w}}(x_i)) \right]$$

Minimalisatie dmv Gradient Descent



LOGISTIC REGRESSION



LOGISTIC REGRESSION – HIGHER ORDER FEATURES



A scatter plot showing the relationship between 'diameter' (x-axis, ranging from 4 to 12) and 'blue' (y-axis, ranging from 0 to 35). The data points are categorized by 'name' (grapefruit, represented by blue dots, and orange, represented by orange dots). A green line represents a linear decision boundary, separating the two classes. The line is nearly vertical, positioned at approximately diameter = 11.5. Most orange points are located to the left of this line (diameter < 11.5), while most grapefruit points are to the right (diameter > 11.5). There is a significant overlap between the two classes in the region where diameter is between 6 and 11 and blue is between 0 and 20.



LOGISTIC REGRESSION – HIGHER ORDER FEATURES



A scatter plot showing the relationship between 'diameter' (x-axis, ranging from 4 to 14) and 'blue' (y-axis, ranging from 0 to 35). The data points are categorized by 'name' (grapefruit, orange). The plot illustrates a non-linear decision boundary (green line) separating the two classes. The boundary is roughly elliptical, centered around diameter 8 and blue 10. The legend indicates that blue dots represent 'grapefruit' and orange dots represent 'orange'.

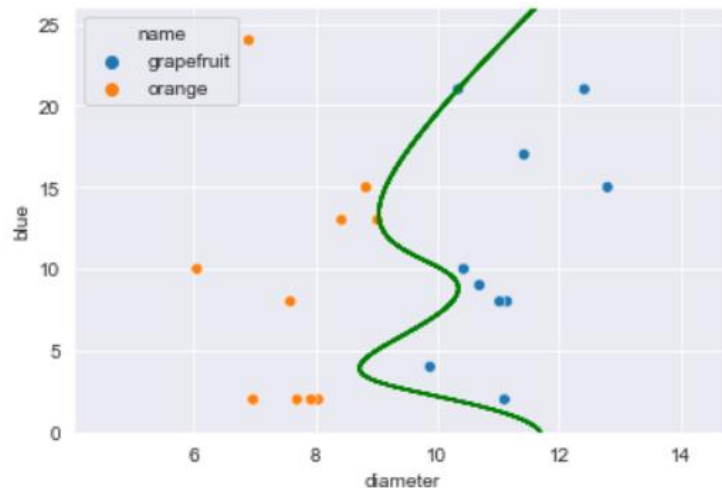


LOGISTIC REGRESSION – REGULARISATIE

Regularisatie via C-parameter

Inverse regularisatie sterkte

Hoge waarde = weinig regularisatie



```
model = LogisticRegression(C=10) # C= inverse regularisatiesterkte  
model.fit(X, df_trimmed.output)
```

LOGISTIC REGRESSION – EVALUATIE

Accuraatheid

Precisie $\frac{TP}{TP+FP}$

Specificiteit $\frac{TN}{TN+FP}$

Recall $\frac{TP}{TP+FN}$

F1-Score $2 \frac{Precision * Recall}{Precision + Recall} = \frac{2TP}{2TP+FP+FN}$

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

VOORBEELD CONFUSION MATRIX

Accuraatheid = $9/12$

Sensitiviteit/Recall = $6/8$

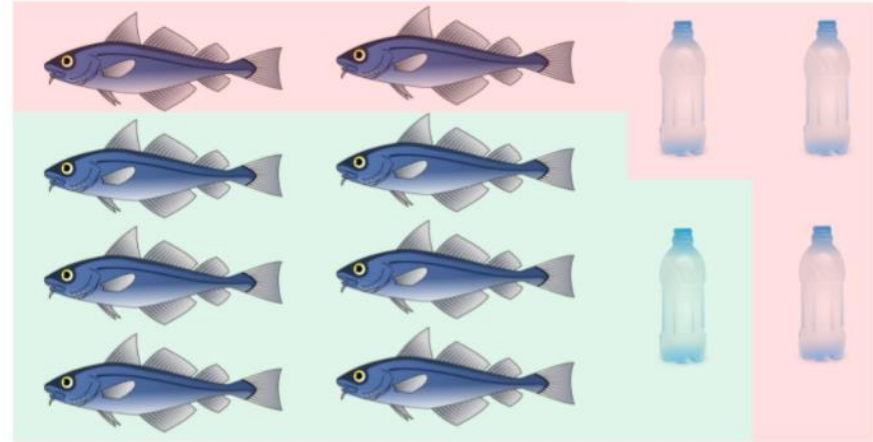
- Weinig positieve samples gemist

Specificiteit = $3/4$

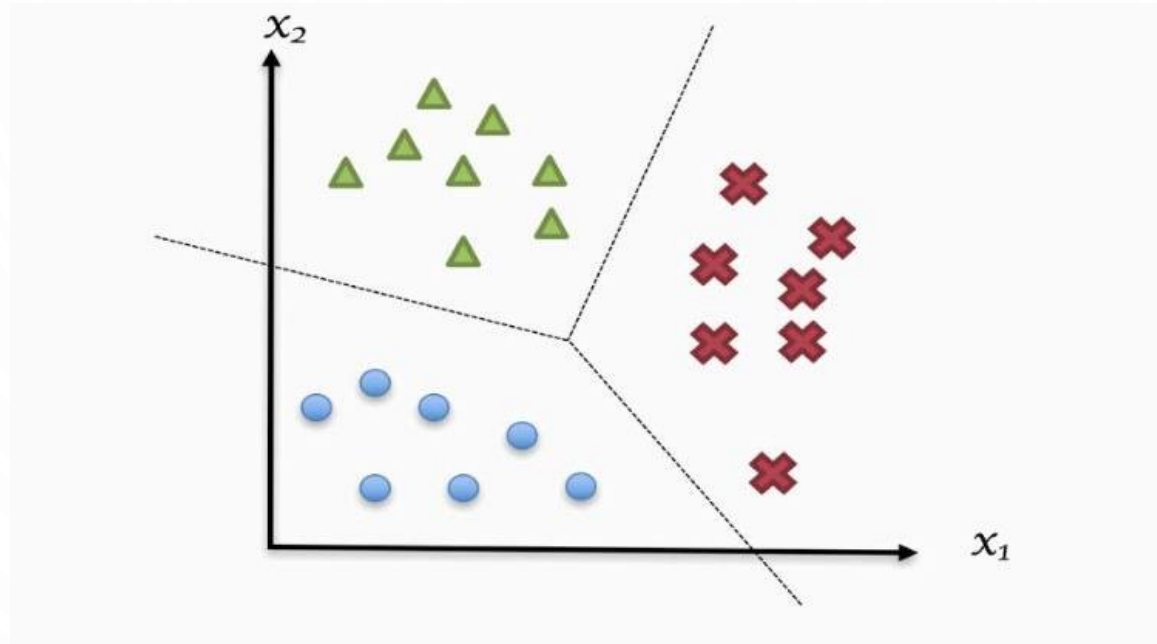
- Weinig negatieve samples gemist

Precision = $6/7$

- Weinig negatieve samples als positieve geclassificeerd

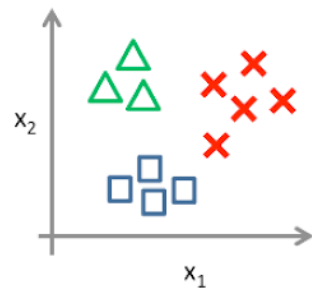


LOGISTIC REGRESSION – MULTICLASS

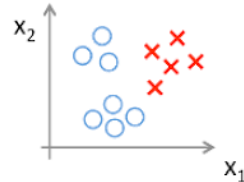
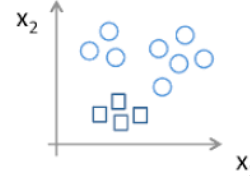
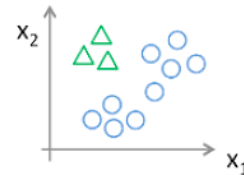


LOGISTIC REGRESSION – ONE VS ALL

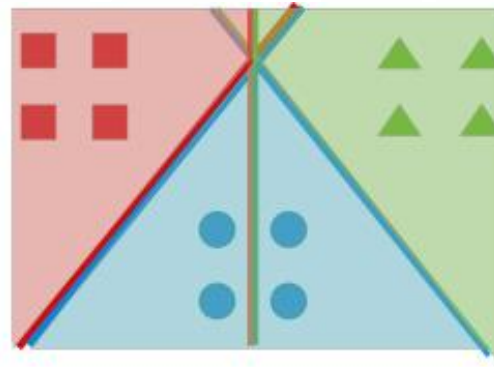
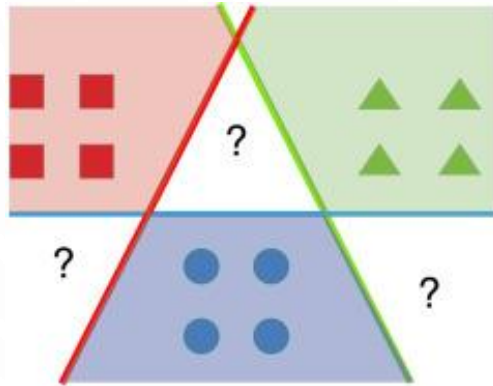
One-vs-all (one-vs-rest):



Class 1: Green
Class 2: Blue
Class 3: Red



LOGISTIC REGRESSION – ONE VS ONE



LOGISTIC REGRESSION – MULTICLASS EVALUATION

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

LOGISTIC REGRESSION – MULTICLASS EVALUATION

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

Class	Precision	Recall	F1-score
Apple	0.29	0.64	0.40
Orange	0.33	0.17	0.22
Mango	0.17	0.08	0.11

LOGISTIC REGRESSION – MULTICLASS EVALUATIE

Micro – F1: Globale waarden

Macro – F1: Gemiddelde F1 – scores

Weighted F1: Gew. Gemiddelde
- Gewichten = # samples

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

Class	Precision	Recall	F1-score
Apple	0.29	0.64	0.40
Orange	0.33	0.17	0.22
Mango	0.17	0.08	0.11

GLOSSARY

- Classificatie
- Binary classifier
- Multi-class classifier
- Multi-label classifier
- True/False Positive/Negative
- Accuraatheid / Specificiteit / ...
- One-vs-All
- One-vs-One
- Confusion matrix