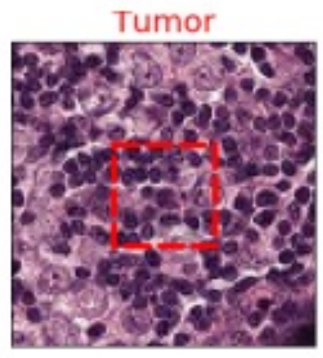
Unabalanced Dataset Problem

Dr. U. Michelucci
umberto.Michelucci@toelt.ai

Problem Description – Classification in Practice (binary)



Classifier

Class: TUMOR

Class: NOT TUMOR

Input Data

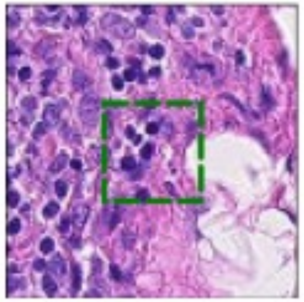
Algorithm

Target
Labels
Classes
...

Image Source: Kolawole, P. (2022). Convolutional Neural Network for Detecting Cancer Tumors in Microscopic Images. Retrieved 19 March 2022, from https://medium.com/swlh/convolutional-neural-network-for-detecting-cancer-tumors-in-microscopic-images-1acab6481d05

Problem Description – Classification in Practice (binary)

No Tumor



Input Data

Classifier

Algorithm

Class: TUMOR

Class: NOT TUMOR

Target
Labels
Classes
...

Image Source: Kolawole, P. (2022). Convolutional Neural Network for Detecting Cancer Tumors in Microscopic Images. Retrieved 19 March 2022, from https://medium.com/swlh/convolutional-neural-network-for-detecting-cancer-tumors-in-microscopic-images-1acab6481d05

Datasets in der Praxis (hospital-acquired infections)

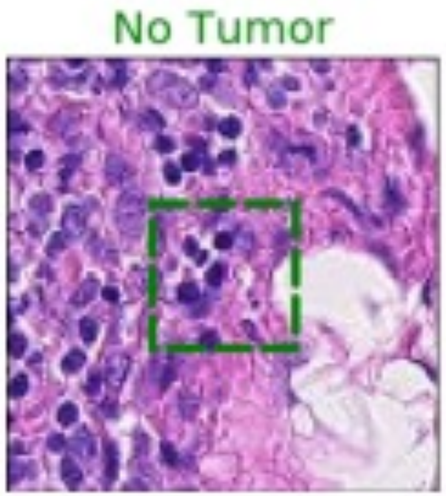
683 patients

Class 0 (Infected): 75 (11% of the total)
Class 1 (healthy): 608 (89% of the total)

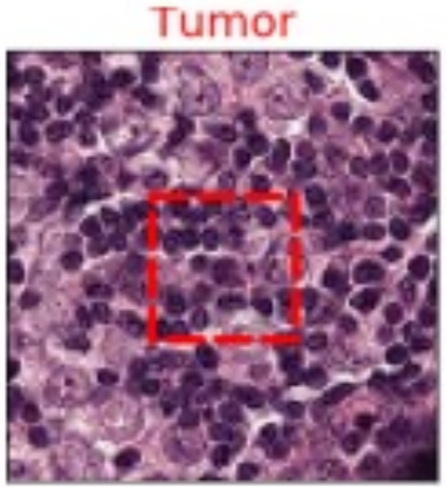
Cohen, Gilles, et al. "Learning from imbalanced data in surveillance of nosocomial infection." Artificial intelligence in medicine 37.1 (2006): 7-18

Datasets in practice (Microsope images)

220,000 training images - 57,458 test images



Class 0 (no Tumor)
55% of images



Class 1 (Tumor)
45% of images

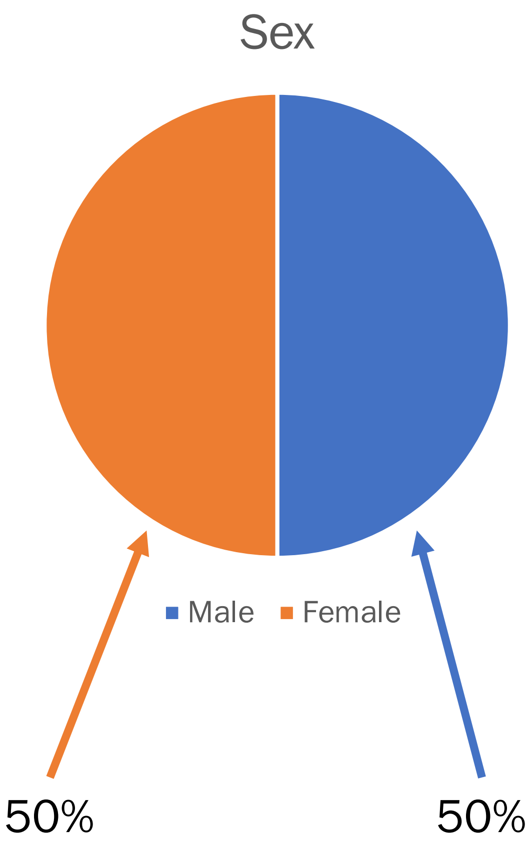
Image Source: Kolawole, P. (2022). Convolutional Neural Network for Detecting Cancer Tumors in Microscopic Images. Retrieved 19 March 2022, from https://medium.com/swlh/convolutional-neural-network-for-detecting-cancer-tumors-in-microscopic-images-1acab6481d05

Unbalanced Dataset

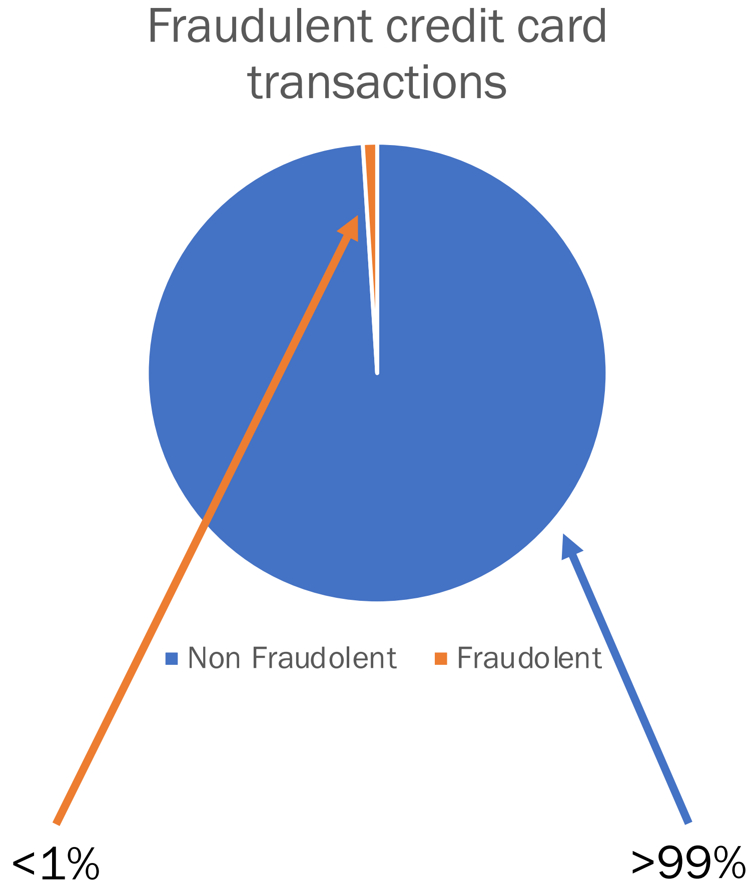
Put simply, an dataset is said unbalanced when the target variable has more observations in a particular class than in the others.

Unbalanced Dataset in Classification - Examples

Target Labels Distribution



Balanced Dataset



Unbalanced Dataset

Quiz

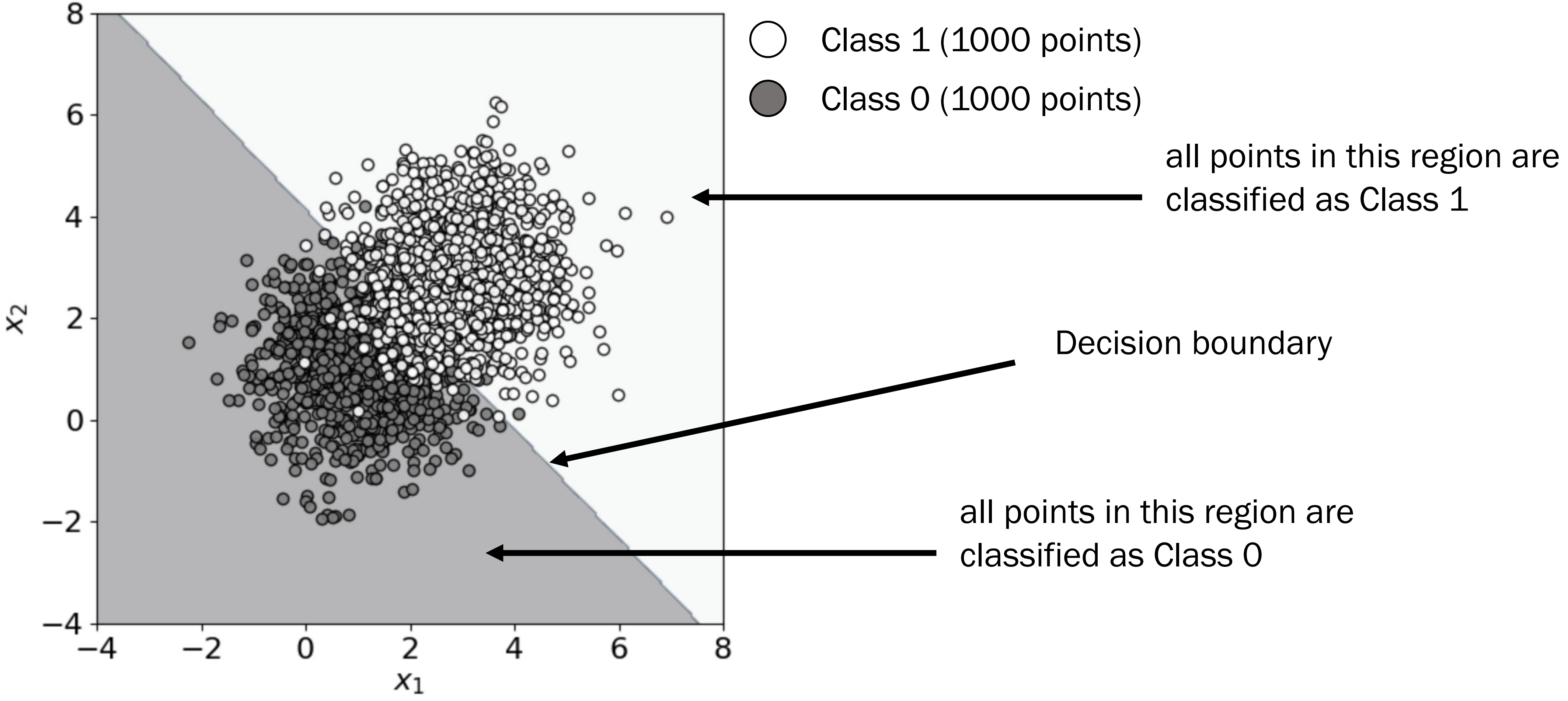
Starting situation

- Dataset: 1000 inputs in Class 1 (e.g. no tumor); 10 in Class 0 (e.g. tumor)
- Trained model achieved: 99% accuracy

Question:

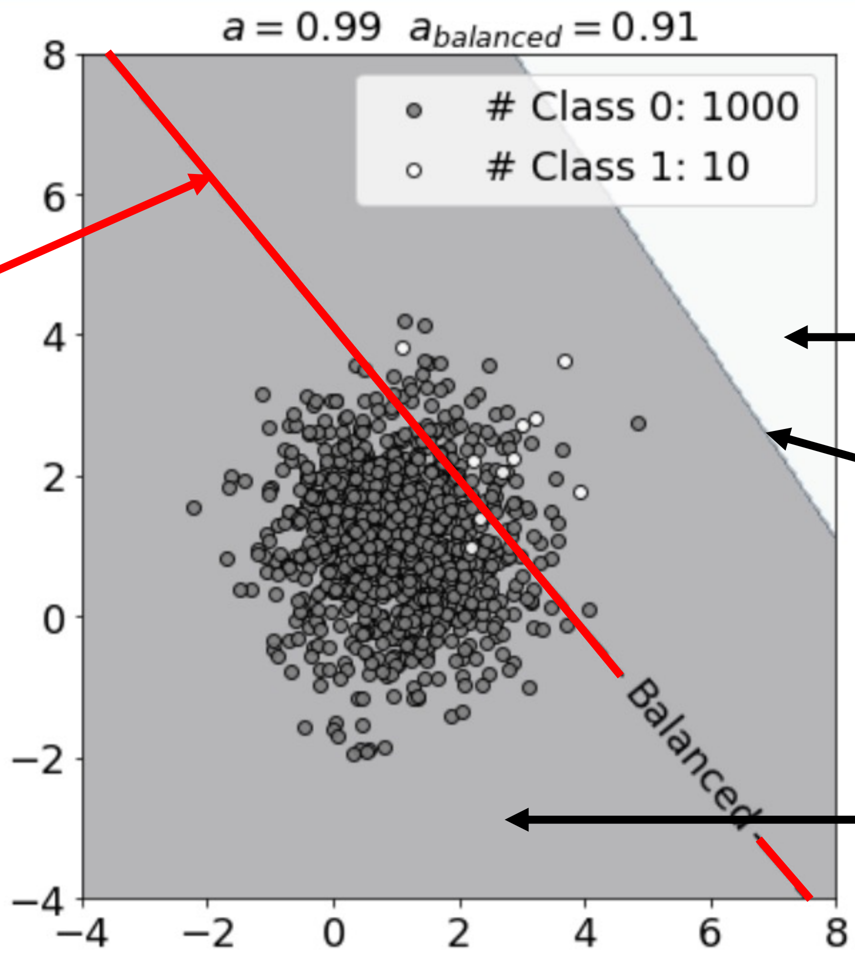
What do you think of the model? Is it good? Ultimately, it achieves 99% accuracy!

Unbalanced Dataset - Consequences



Approx. 91% accuracy is achieved (with linear support vector classifier)

Unbalanced Dataset - Consequences



Class 1 (10 points)

Class 0 (1000 points)

Number of points in Class 1 is 1% Number of points in Class 0

Decision boundary with balanced dataset

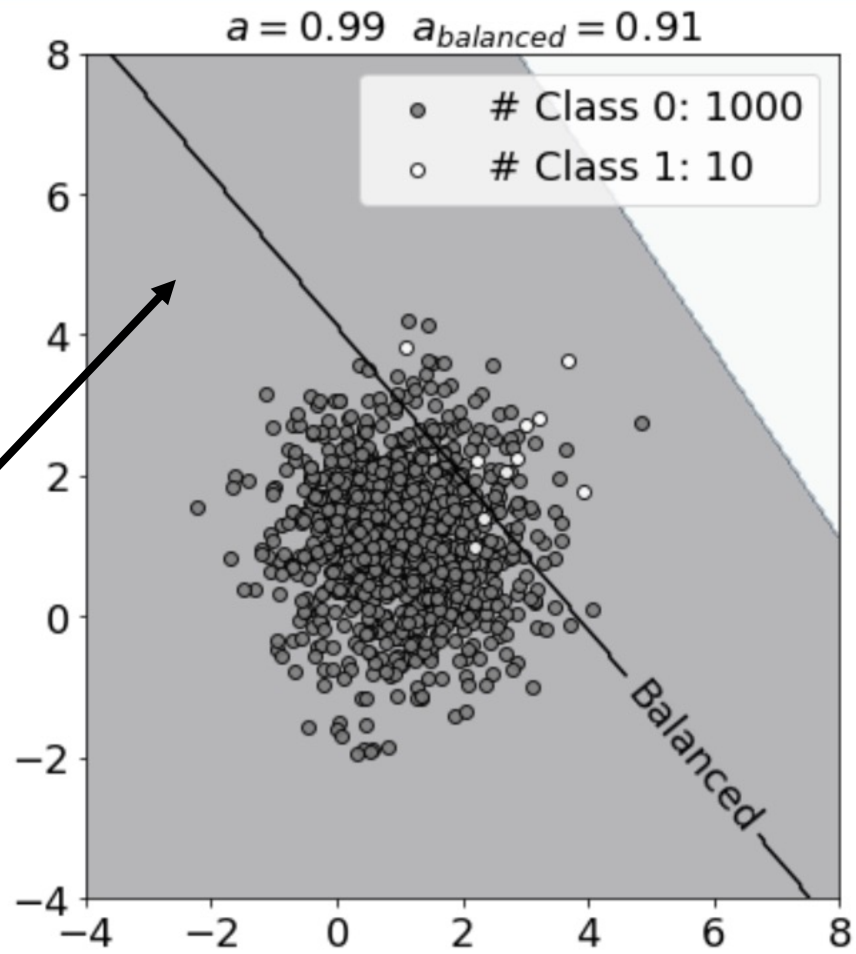
all points in this region are classified as Class 1

Decision boundary

all points in this region are classified as Class 0

Approx. 99% accuracy is achieved (with linear support vector classifier)

Unbalanced Dataset - Consequences



alle Punkte in dieser Region werden als *Class 0* klassifiziert

Number of points in Class 1 is 1% Number of points in Class 0

○ Class 1 (10 points)

● Class 0 (1000 points)

Accuracy $\boldsymbol{a}$

$$a = \frac{number\ of\ correctly\ classified\ points}{total\ number\ of\ points}$$

$$a = \frac{1000}{1000+10} \rightarrow a = \frac{1000/1000}{1000/1000 + {}^{10}/_{1000}}$$

$$\rightarrow a = \frac{1}{1+10/1000} \rightarrow a = \frac{1}{1+0.01}$$

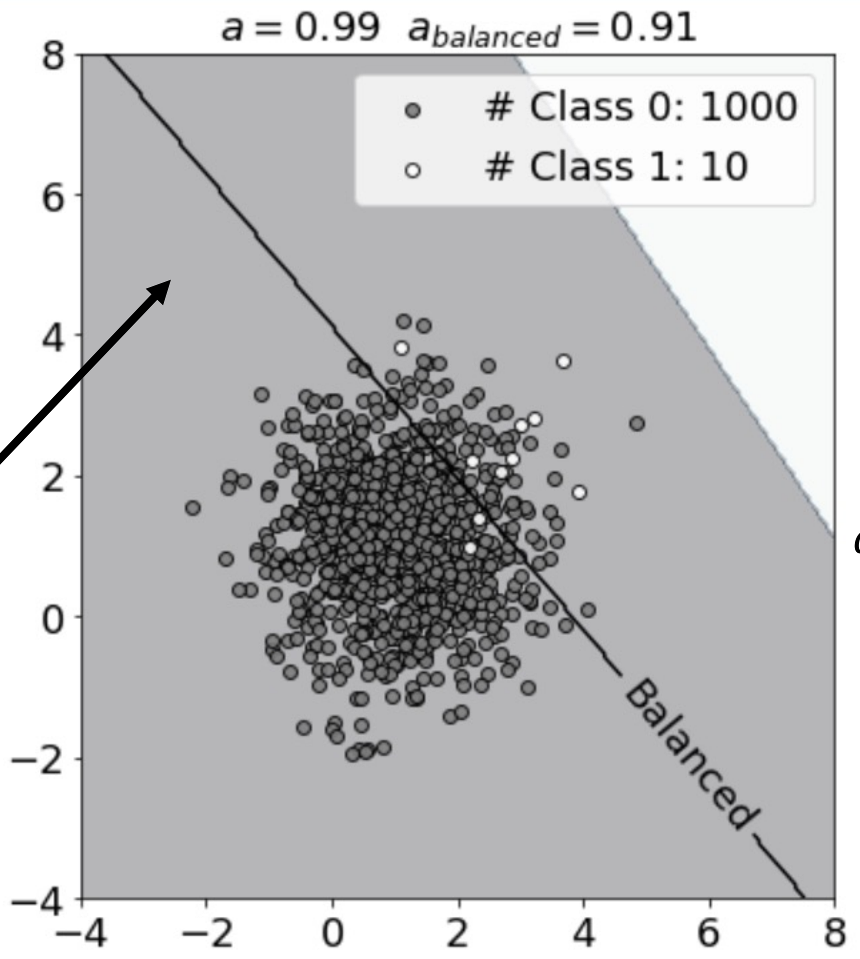
$$\rightarrow a = \frac{1}{1+0.01} \approx 1 - 0.01 = 0.99$$

Don't forget:

$\frac{1}{1+x} \approx 1 - x \quad$ for $x \ll 1$

Approx. 99% accuracy is achieved (with linear support vector classifier)

Unbalanced Dataset - Consequences



all points in this region are classified as Class 0

○ Number of points Class 1 - N_1

● Number of points Class 0 - N_0

In case of $N_1 \ll N_0$ (UNBALANCED DATASET)

$$a = \frac{number\ of\ correctly\ classified\ points}{total\ number\ of\ points}$$

$$a = \frac{N_0}{N_0 + N_1} \quad \rightarrow \quad a = \frac{1}{1 + N_1/N_0}$$

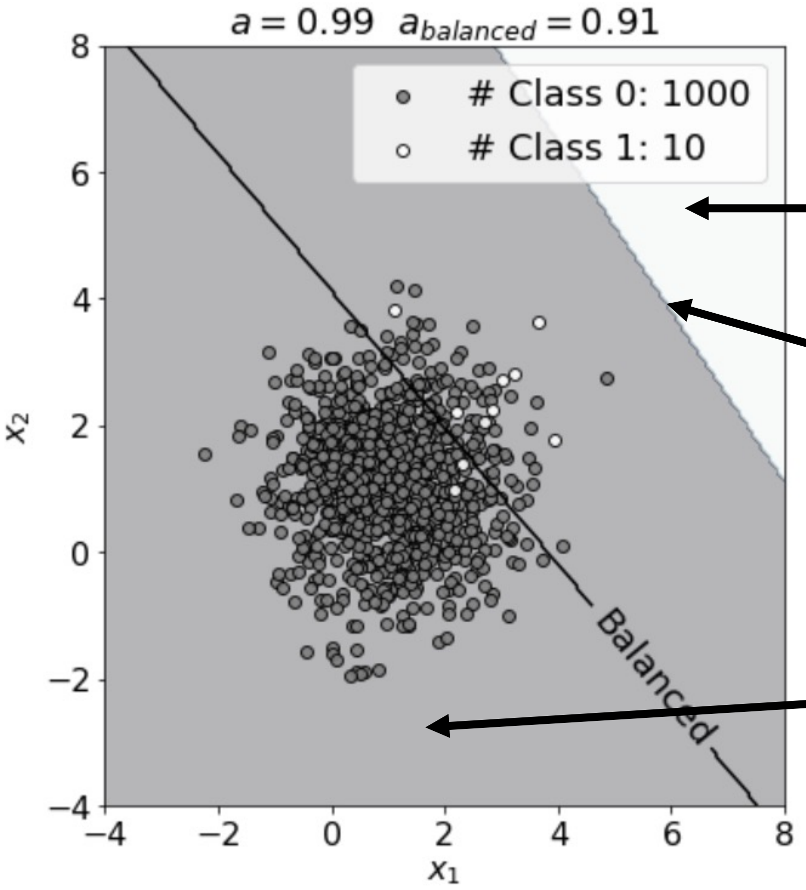
$$\rightarrow a = \frac{1}{1 + N_1/N_0} \approx 1 - N_1/N_0$$ Only valid for $N_1/N_0 \ll 1$

Don't forget:

$\frac{1}{1+x} \approx 1 - x \quad$ for $x \ll 1$

a	N_0	N_1
99%	1000	10
98%	1000	20
95%	1000	50

Class 0: 1000 – Class 1:10 - $a = 99\%$

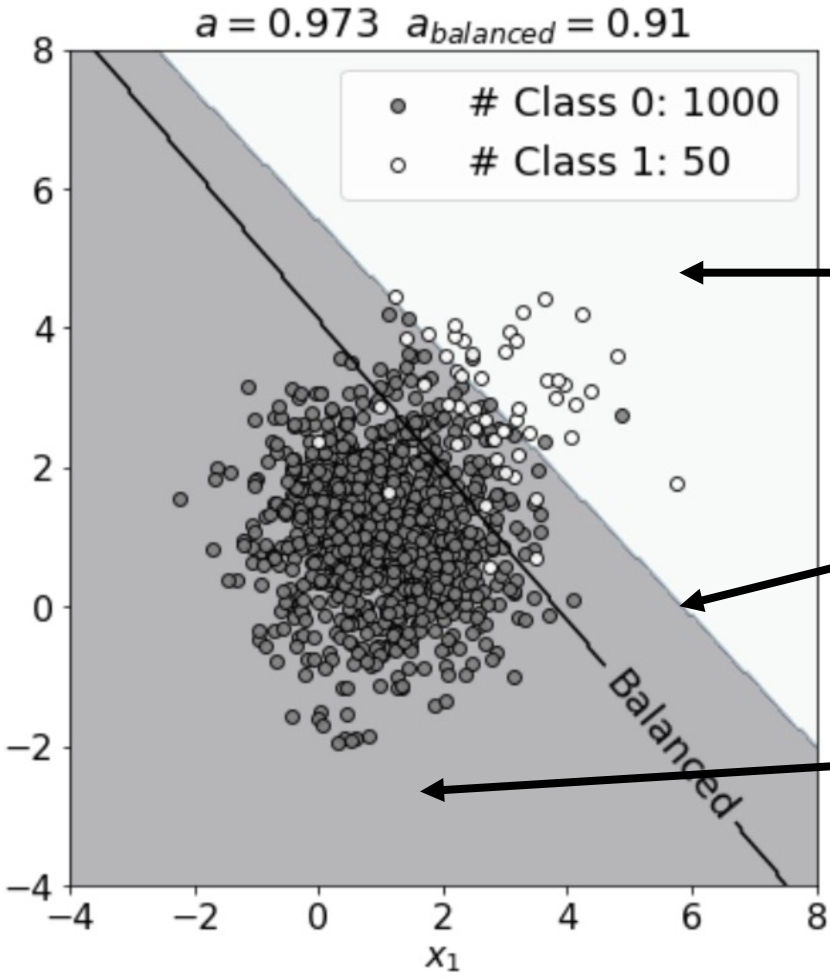


all points in this region are classified as Class 1

Decision boundary

all points in this region are classified as Class 0

Class 0: 1000 – Class 1:50 - $a = 97.3\%$

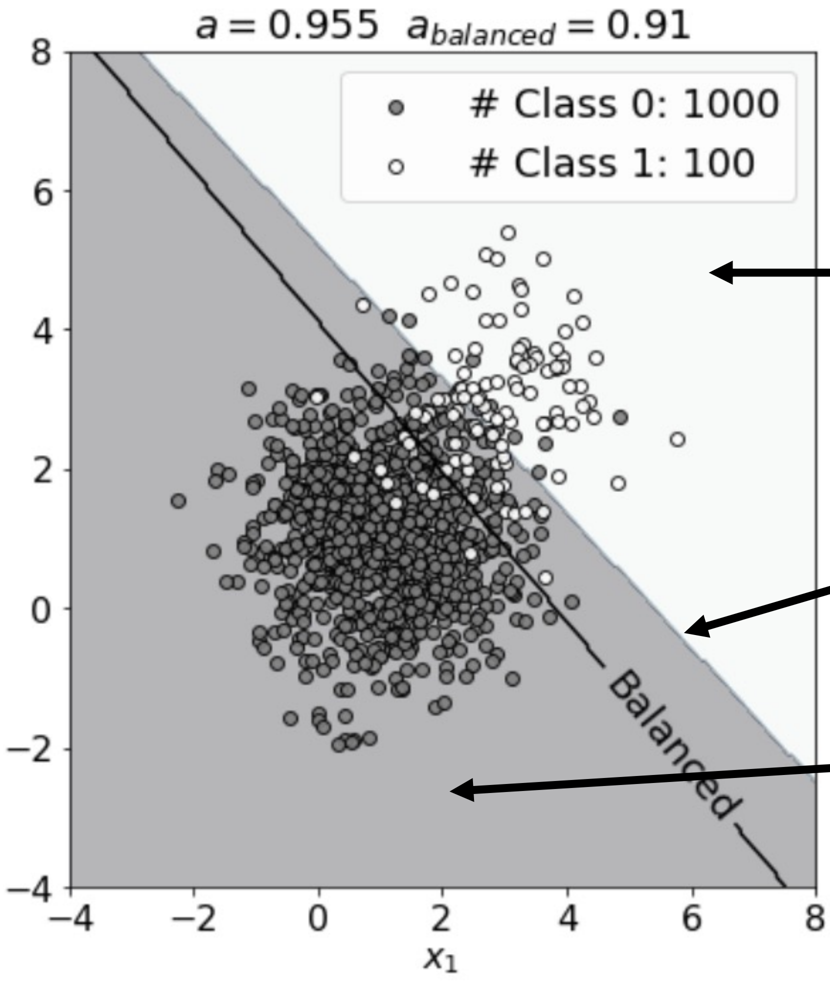


all points in this region are classified as Class 1

Decision boundary

all points in this region are classified as Class 0

Class 0: 1000 – Class 1:100 - $a = 95.5\%$

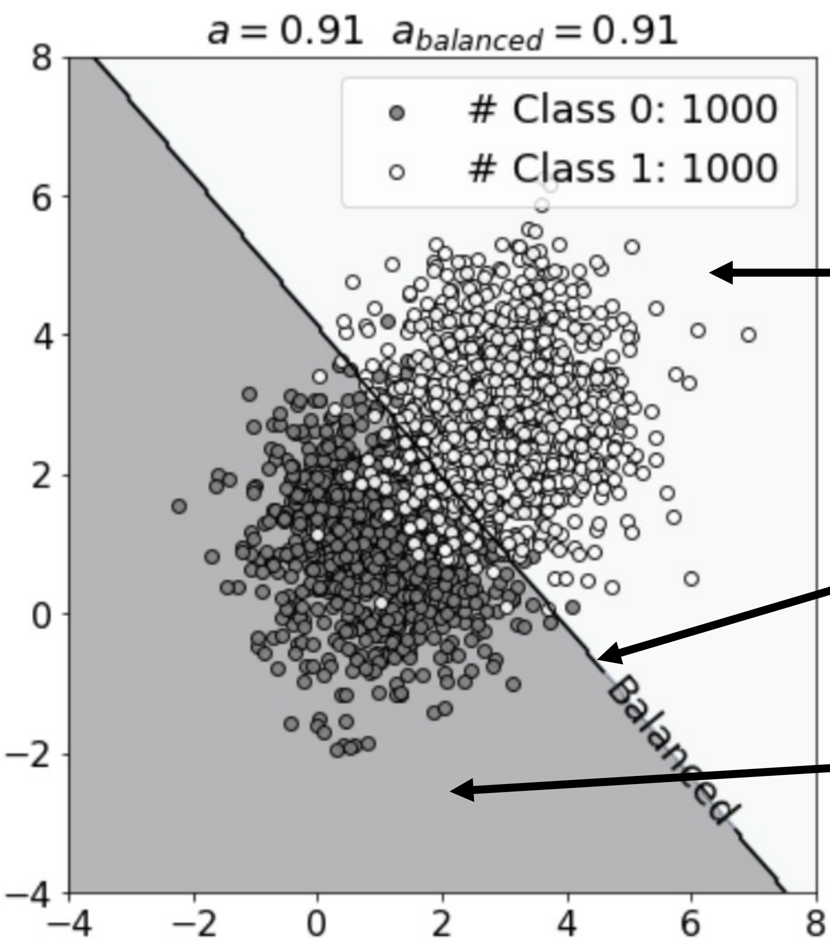


all points in this region are classified as Class 1

Decision boundary

all points in this region are classified as Class 0

Class 0: 1000 – Class 1:1000 - $a = 91\%$



all points in this region are classified as Class 1

Decision boundary

all points in this region are classified as Class 0

Real-life Scenario

- Study on hospital-acquired infections: "**Out of 683 patients, only 75 (11% of the total) were infected and 608 were not**" (Cohen, Gilles, et al. "Learning from imbalanced data in surveillance of nosocomial infection." *Artificial intelligence in medicine* 37.1 (2006): 7-18)

Table 1 Baseline performance (original class distribution: 0.11 pos, 0.89 neg)

Classifier	Sensitivity	Specificity	CWA	Accuracy
IB1 (kNN)	0.19	0.96	0.38	0.88
Nave Bayes	0.57	0.88	0.65	0.85
C4.5 (Decision Trees)	0.28	0.95	0.45	0.88
AdaBoost	0.45	0.95	0.58	0.90
SVM	0.43	0.92	0.55	0.86

Sensitivity: **the ability of a test to correctly identify patients with a disease**

$$Sensitivity = \frac{TP}{P} = \frac{\text{"Number of inputs classified as "sick" and have a true class of "sick""}}{\text{Number of sick patients}}$$

Strategies for dealing with unbalanced data sets

Type 1 consists in the pre-processing of the data in order to restore the class balance.

Type 2 consists of modifying the algorithms themselves so that they can handle unbalanced data.

Today we will deal with type 1.

Unbalanced Dataset - Solutions

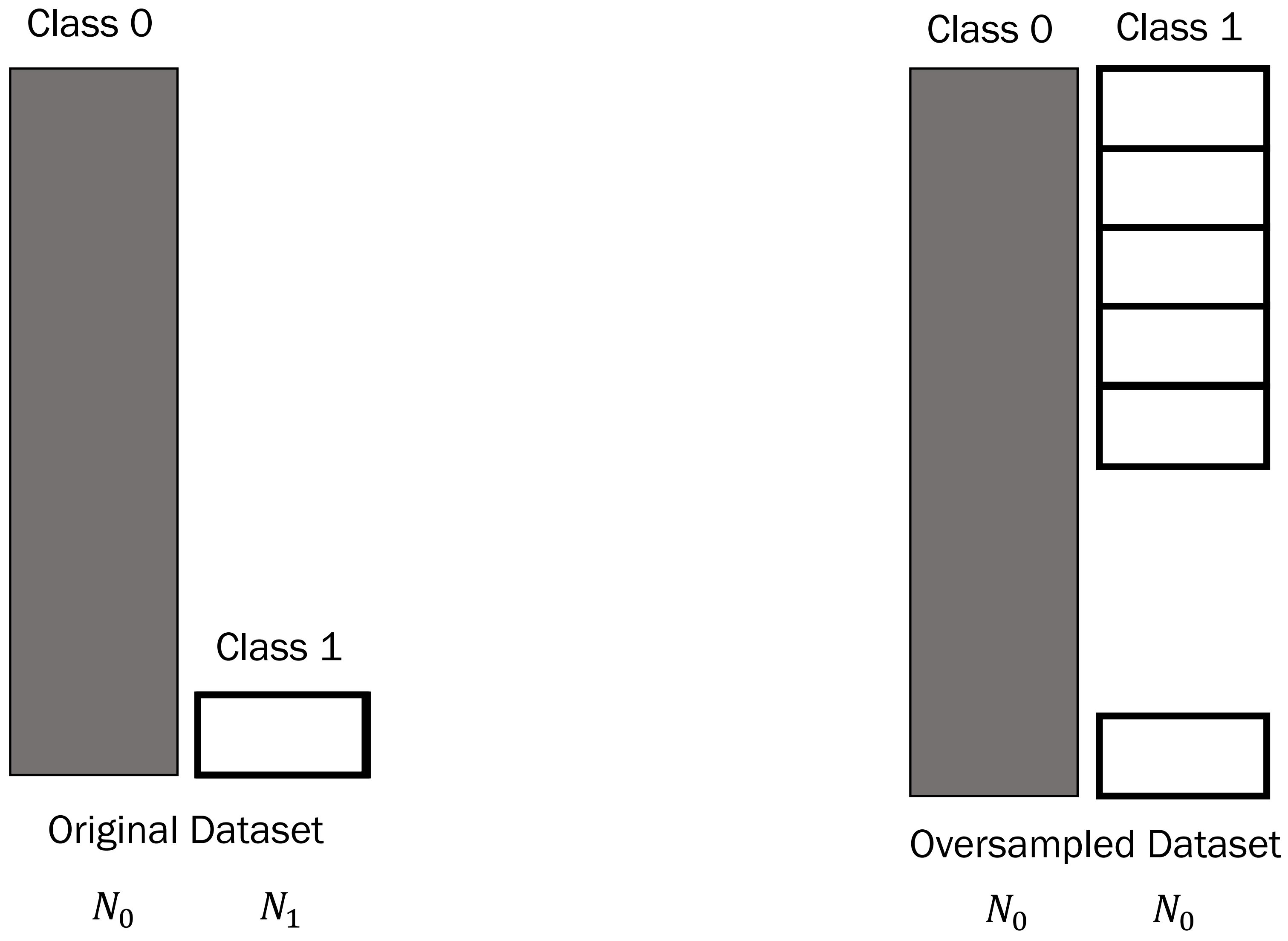
1) Collect more data on the less represented class

2) Undersampling (subsampling) and oversampling

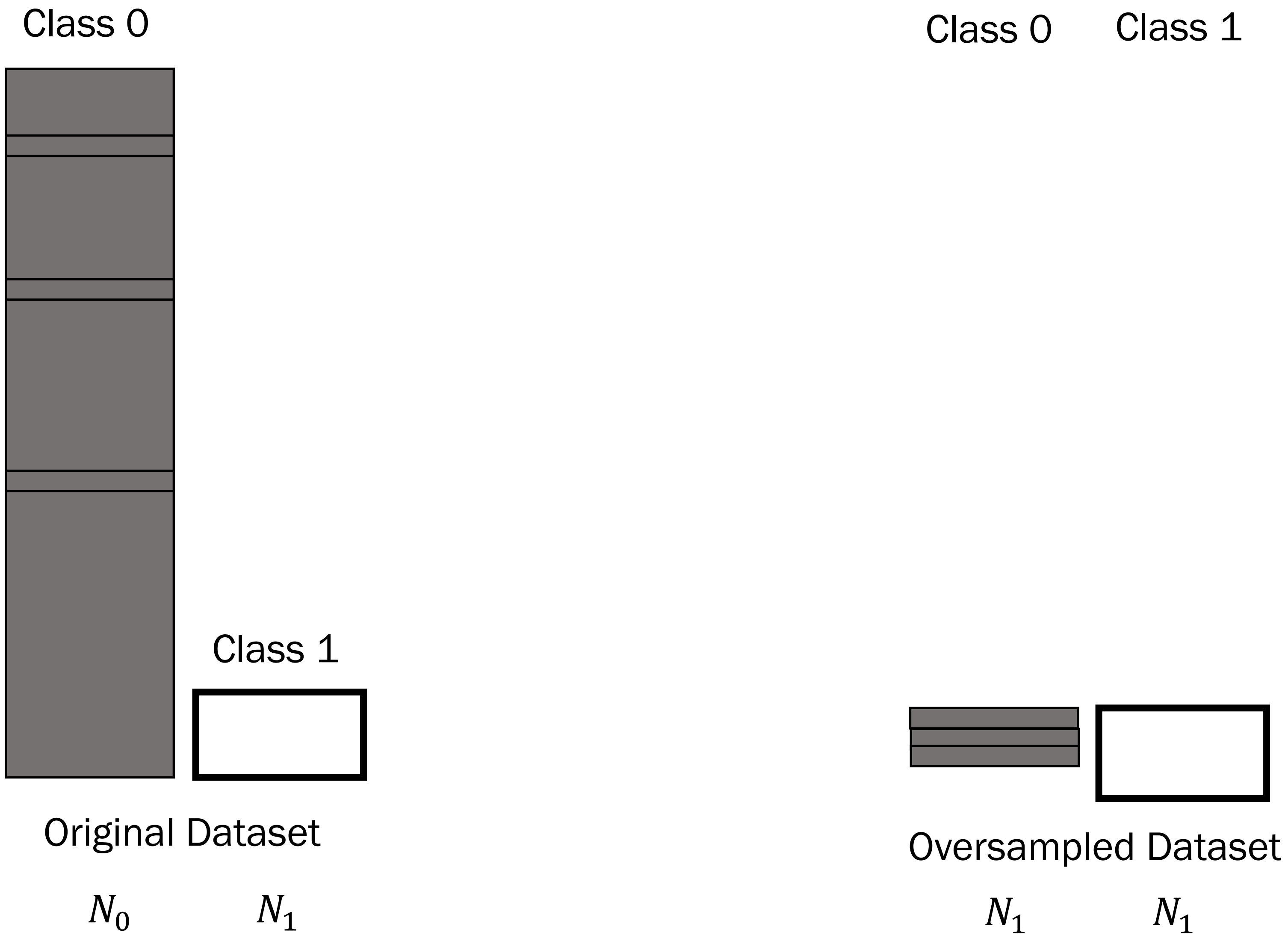
3) Use different metrics

Under- und Oversampling

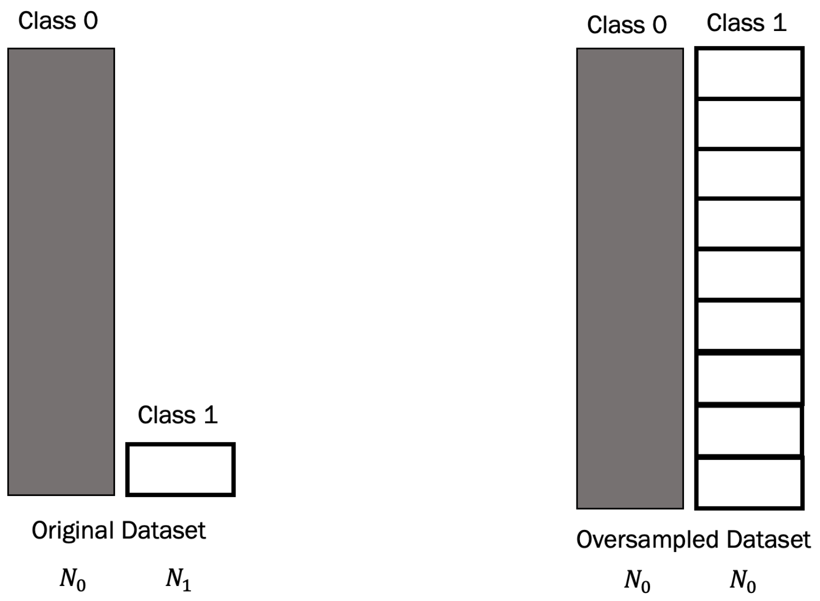
Oversampling



Undersampling (random undersampling)



Advantages and disadvantages



Oversampling

Advantages:

- You have a larger data set for training

- Disadvantages:

- The model generalizes worse because it has seen very few Class 1 cases

Advantages and disadvantages

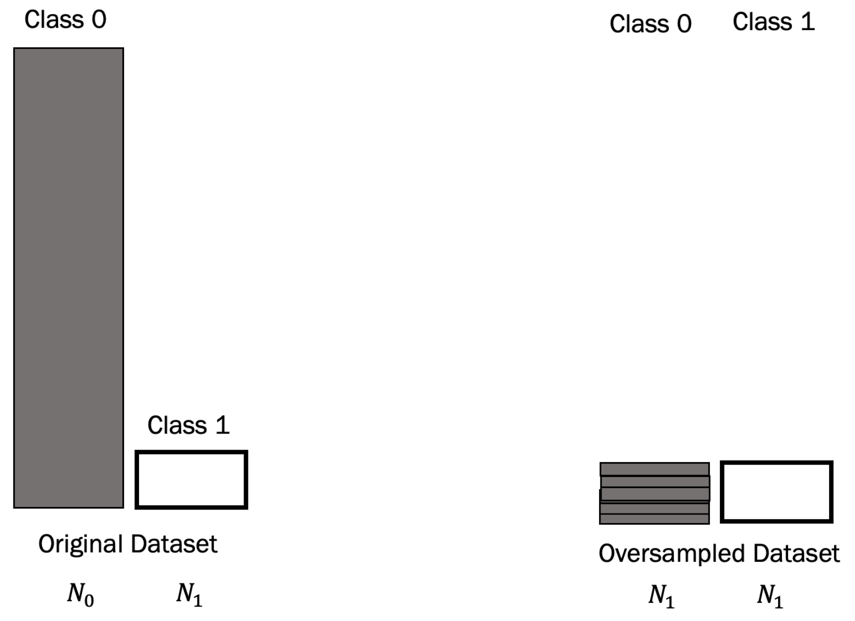
Downsampling

Advantages:

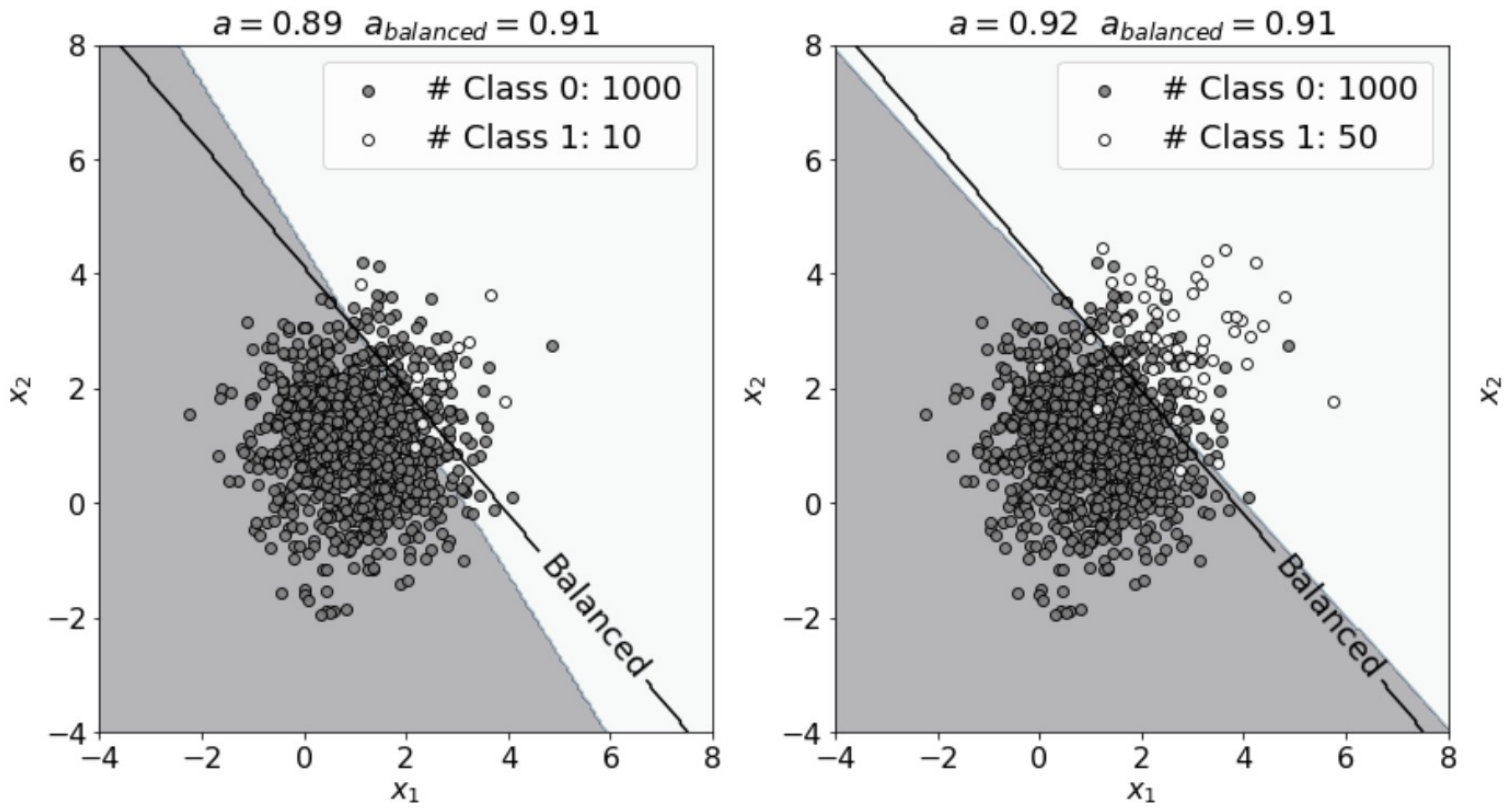
- You don't duplicate data

Disadvantages:

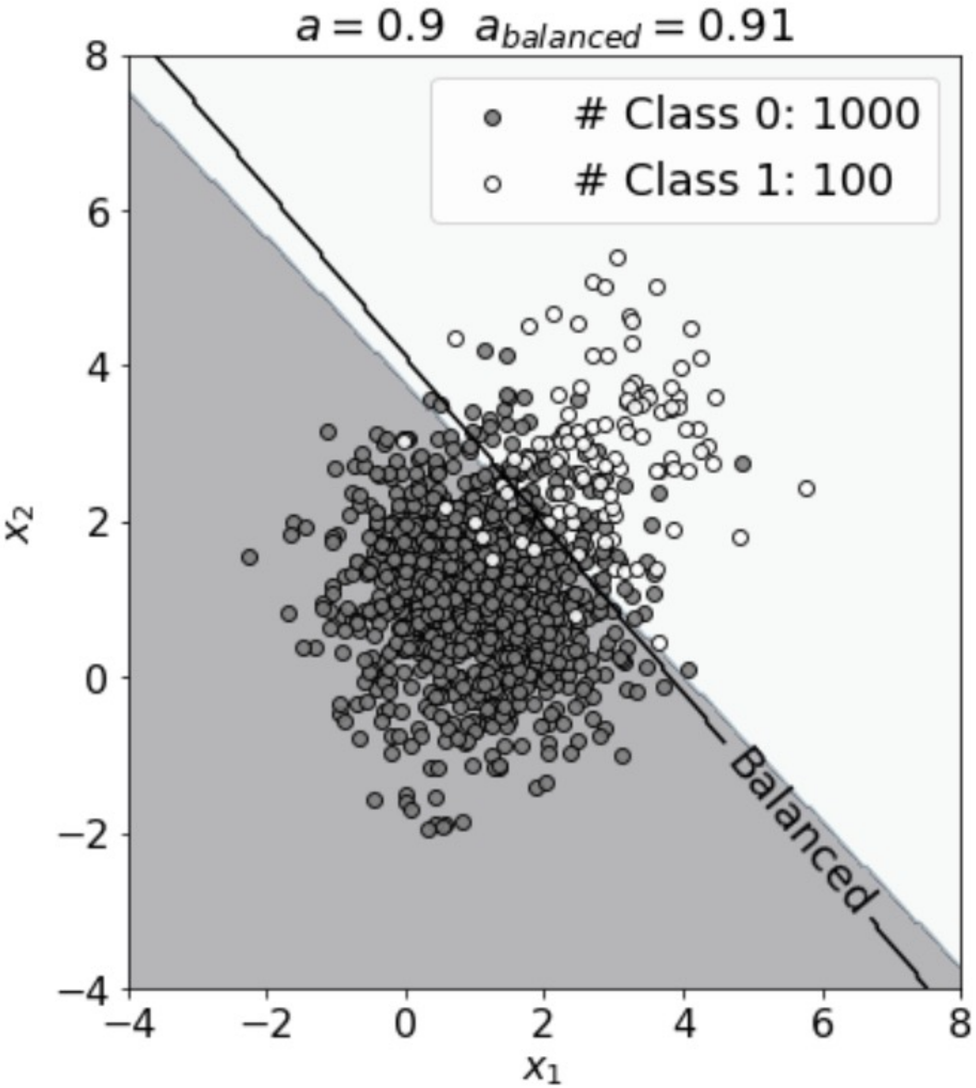
- The training record is much smaller, and you lose a lot of data in class 0.
- It can happen that the performance of the models quickly deteriorates with small data sets.

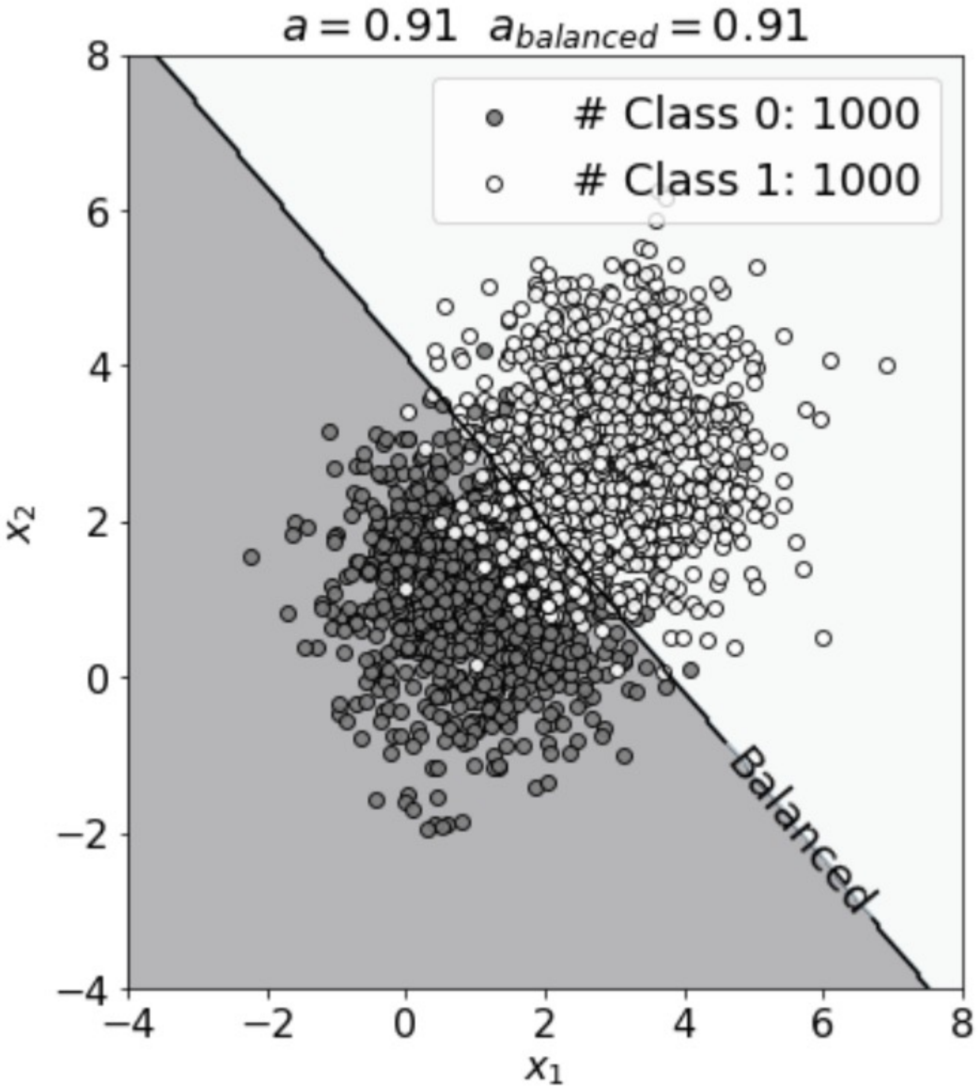


Oversampling

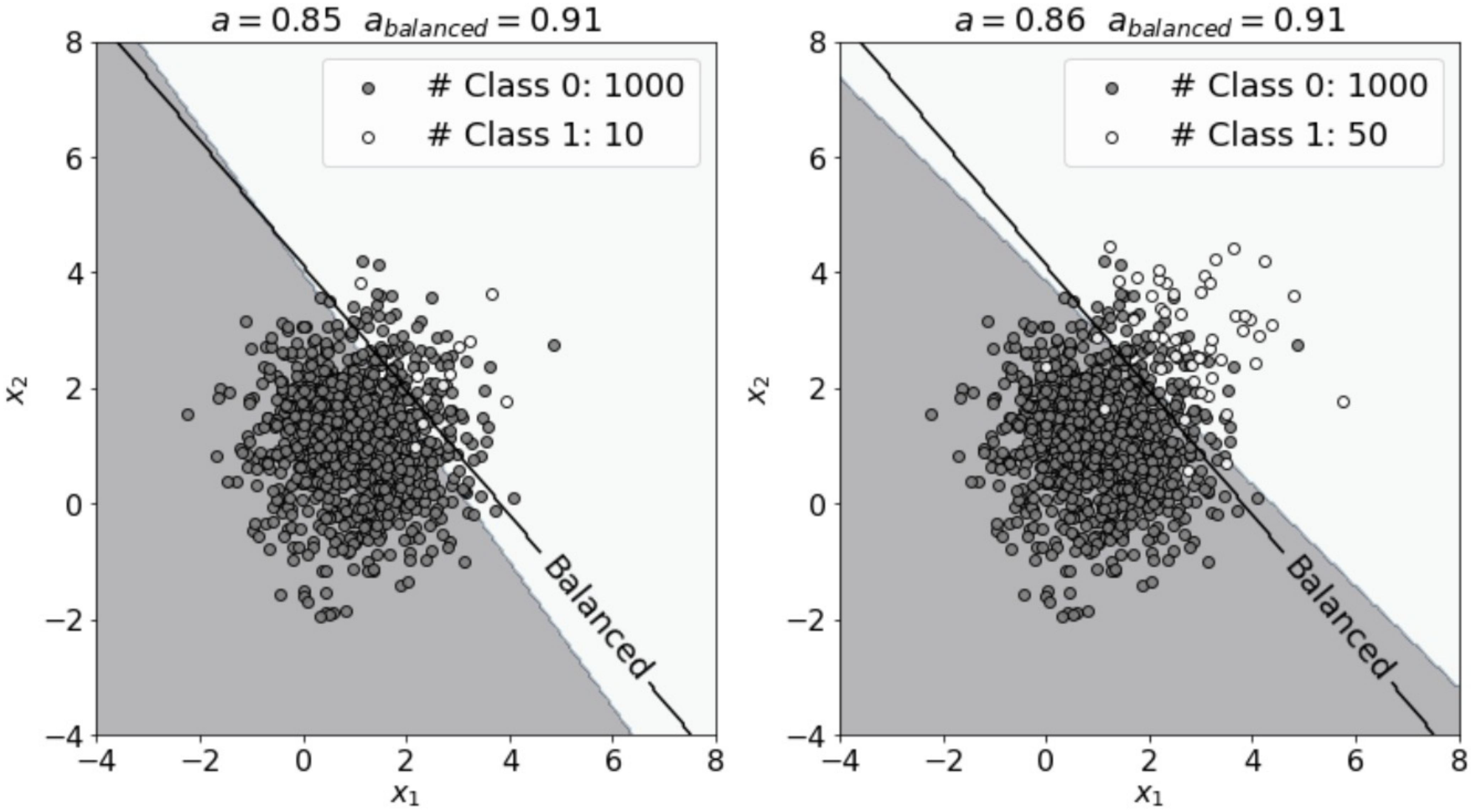


Oversampling

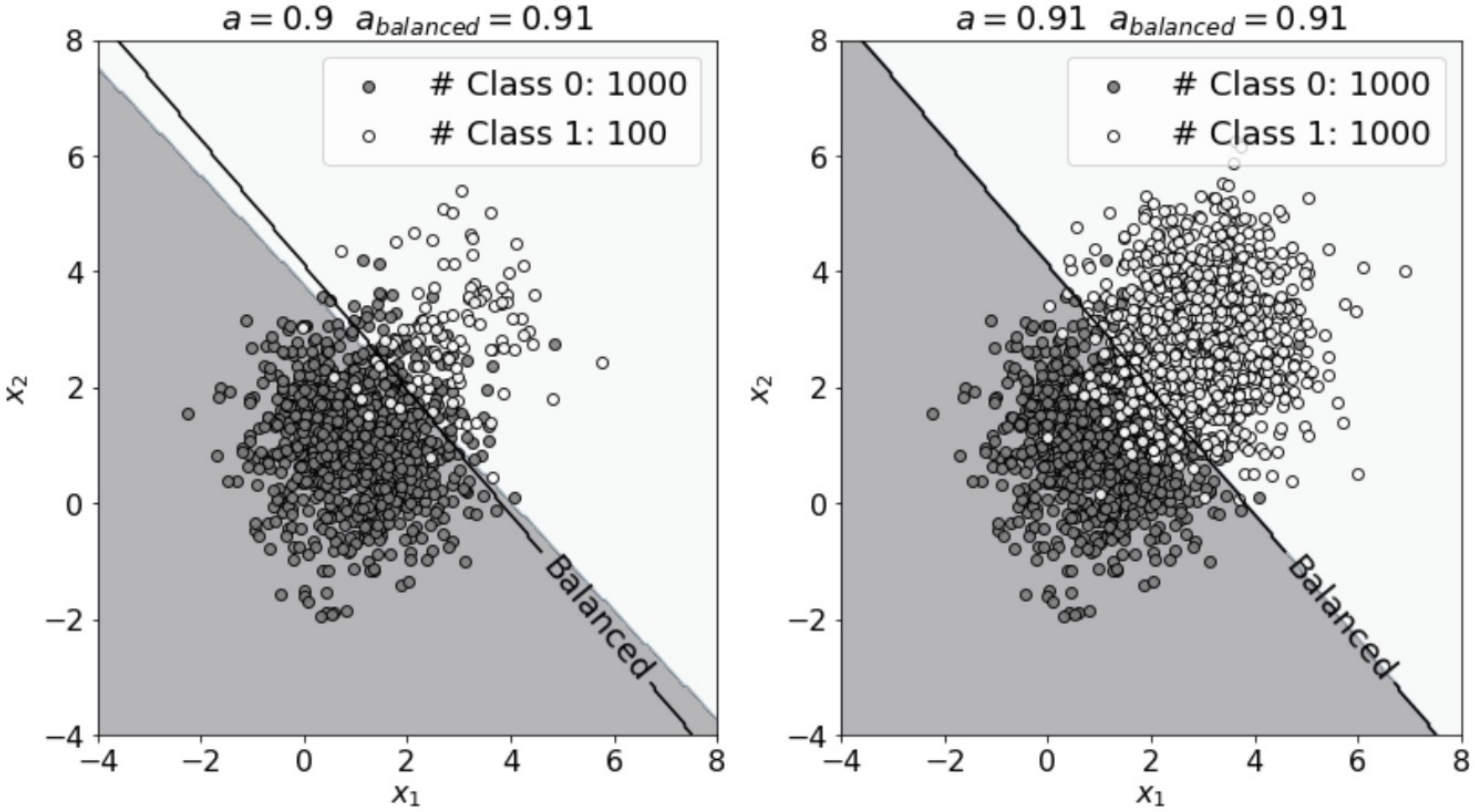




Undersampling



Undersampling



Metrics for Unbalanced Dataset Scenarios

Metrics - Overview

Binary Classification		Multi-class Classification	
Metric	**Discussed today**	**Metric**	**Discussed today**
Accuracy	•	Accuracy	•
Specificity	•	Accuracy pro class	
Sensitivity	•		
Balanced Accuracy	•		
F1 Score	•		
Area Under The Curve (AUC / Receiving Operating Curve)	•		

- Confusion Matrix → Not really a metric (it contains multiple numbers)

Confusion Matrix (binary classification)

	Predicted Label 0	Predicted Label 1
True Label 0	Number of inputs classified as 0 and that have a true class of 0 (TRUE POSITIVES, TP)	Number of inputs classified as 1 and that have a true class of 0 (FALSE NEGATIVES, FN)
True Label 1	Number of inputs classified as 0 and that have a true class of 1 (FALSE POSITIVES, FP)	Number of inputs classified as 1 and that have a true class of 1 (TRUE NEGATIVES, TN)

<u>In the case of the perfect classifier:</u>

FN = 0 and FP = 0

Sensitivity und Specificity

Sensitivity / Recall
True Positive Rate

$$\frac{TP}{P} = \frac{TP}{TP + FN}$$

How many of the cases in class 0 were correctly identified?
"How many sick among all the sick have I identified?"

Specificity / selectivity
True Negative Rate

$$\frac{TN}{N} = \frac{TN}{TN + FP}$$

How many of the cases in Class 1 were correctly identified?
"How many healthy people among all healthy people have I identified?"

Confusion Matrix

True Label		
	TP	FN
	FP	TN

Predicted Label

$$P = TP + FN$$

$$N = TN + FP$$

<u>PAY ATTENTION</u>: Sensitivity + Specificity ≠ 1

<u>In the case of the perfect classifier:</u>
FN = 0 und FP = 0 → P = TP und TN = N

Sensitivity = 1 AND **Specificity = 1**

Balanced Accuracy

Sensitivity / Recall
True Positive Rate

$$\frac{TP}{P} = \frac{TP}{TP + FN}$$

How many of the cases in class 0 were correctly identified?
"How many sick among all the sick have I identified?"

Specificity / selectivity
True Negative Rate

$$\frac{TN}{N} = \frac{TN}{TN + FP}$$

How many of the cases in Class 1 were correctly identified?
"How many healthy people among all healthy people have I identified?"

Confusion Matrix

True Label \ Predicted Label		
	TP	FN
	FP	TN

Predicted Label

$$P = TP + FN$$

$$N = TN + FP$$

Balanced Accuracy = a_B **=** $\frac{Sensitivity + Specificity}{2}$

F1 Score

Sensitivity / Recall
True Positive Rate

$$\frac{TP}{P} = \frac{TP}{TP + FN}$$

How many of the cases in class 0 were correctly identified?
"How many sick among all the sick have I identified?"

Specificity / selectivity
True Negative Rate

$$\frac{TN}{N} = \frac{TN}{TN + FP}$$

How many of the cases in Class 1 were correctly identified?
"How many healthy people among all healthy people have I identified?"

Confusion Matrix

True Label		
	TP	FN
	FP	TN

Predicted Label

$$P = TP + FN$$

$$N = TN + FP$$

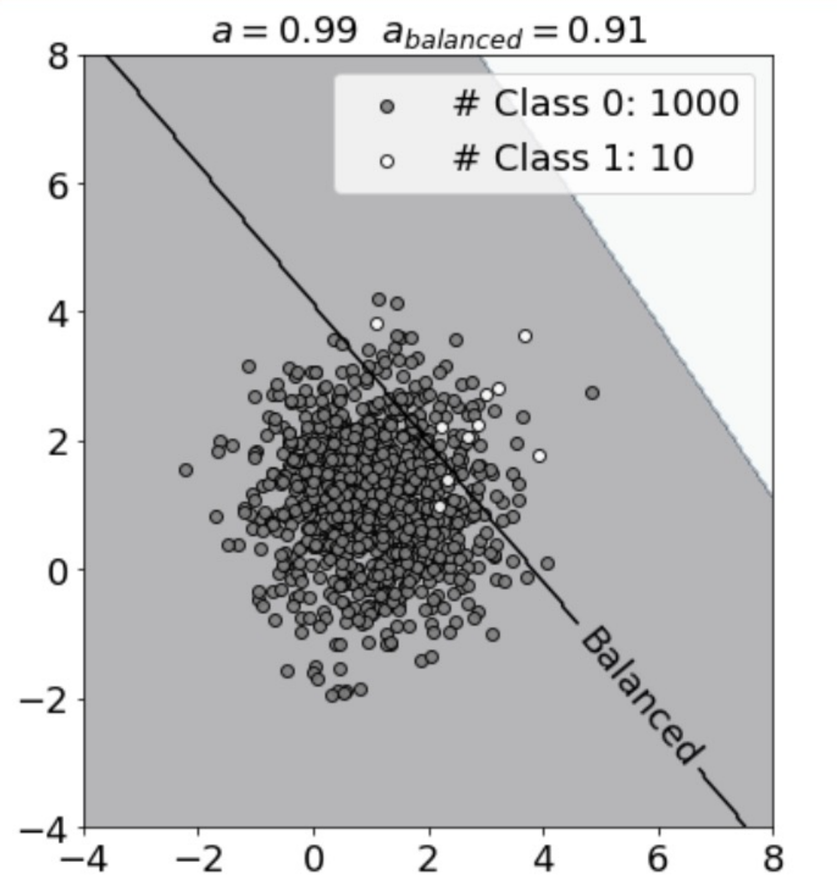
$$\mathbf{F1} = \frac{2}{\frac{1}{Sensitivity} + \frac{1}{Specificity}}$$

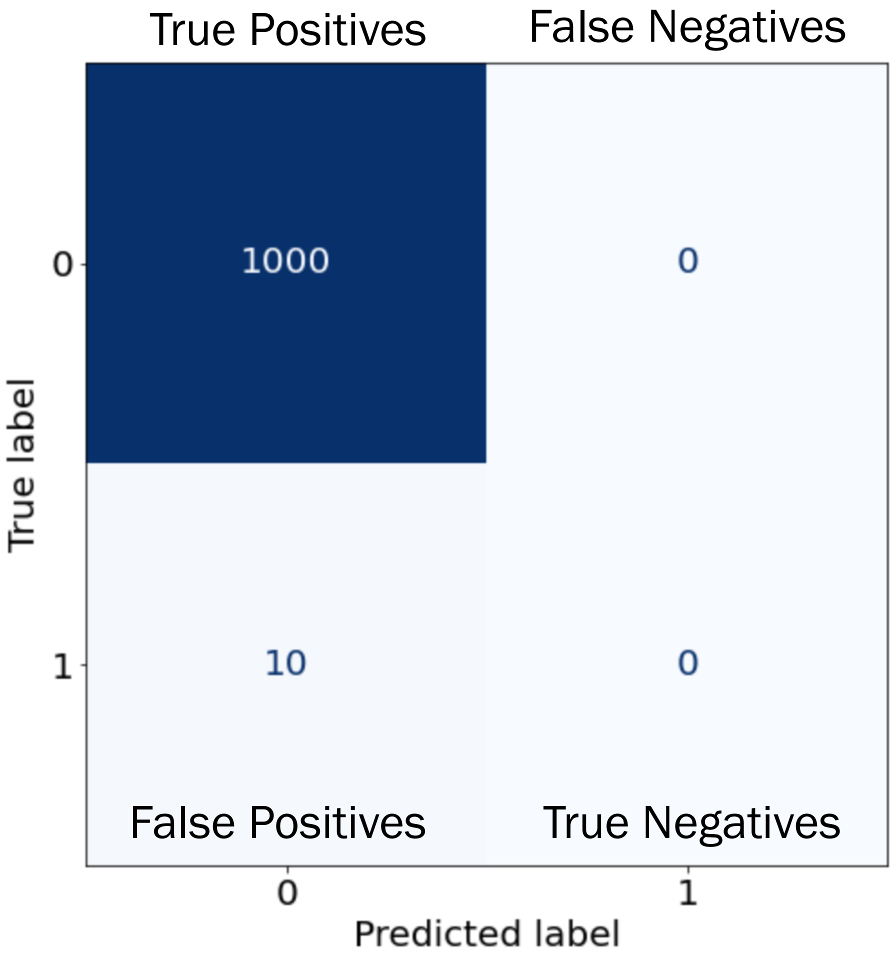
* harmonic mean

Summary – confusion matrix

		PREDICTED CLASS			
		POSITIVE PP	**NEGATIVE** PN		
ACTUAL (TRUE) CLASS	**POSITIVE** P	TRUE POSITIVES (TP)	FALSE NEGATIVES (FN)	**SENSITIVITY** $\frac{TP}{TP+FN}=\frac{TP}{P}$	**BALANCED ACCURACY** $a_B = \frac{\text{Sensitivity} + \text{Specificity}}{2}$ **F1 SCORE** $F1 = \frac{2}{\text{Sensitivity}^{-1} + \text{Specificity}^{-1}}$
	NEGATIVE N	FALSE POSITIVES (FP)	TRUE NEGATIVES (TN)	**SPECIFICITY** $\frac{TN}{TN+FP}=\frac{TN}{N}$	
		PRECISION $\frac{TP}{TP+FP}=\frac{TP}{PP}$	**NEGATIVE PREDICTIVE VALUE** $\frac{TN}{TN+FN}=\frac{TN}{PN}$	**ACCURACY** $\frac{TP+TN}{TP+TN+FP+FN}$	

Class 0:1000 – Class 1:10 – Confusion Matrix





SENSITIVITY

$$\frac{TP}{TP+FN} = \frac{TP}{P}$$

= 100 %

SPECIFICITY

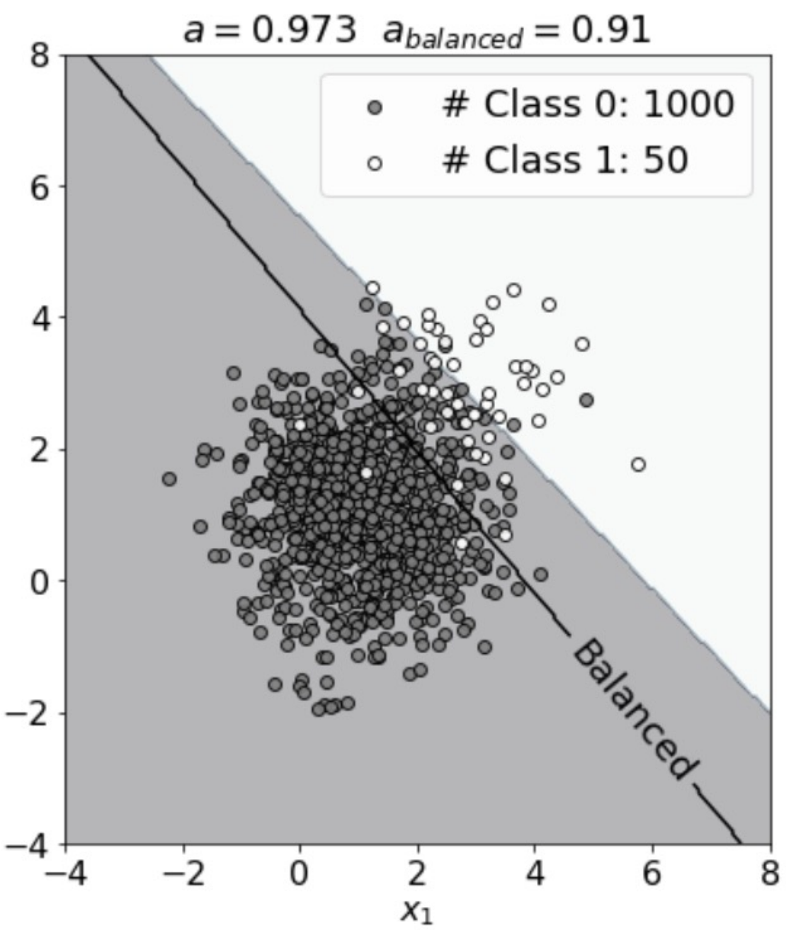
$$\frac{TN}{TN+FP} = \frac{TN}{N}$$

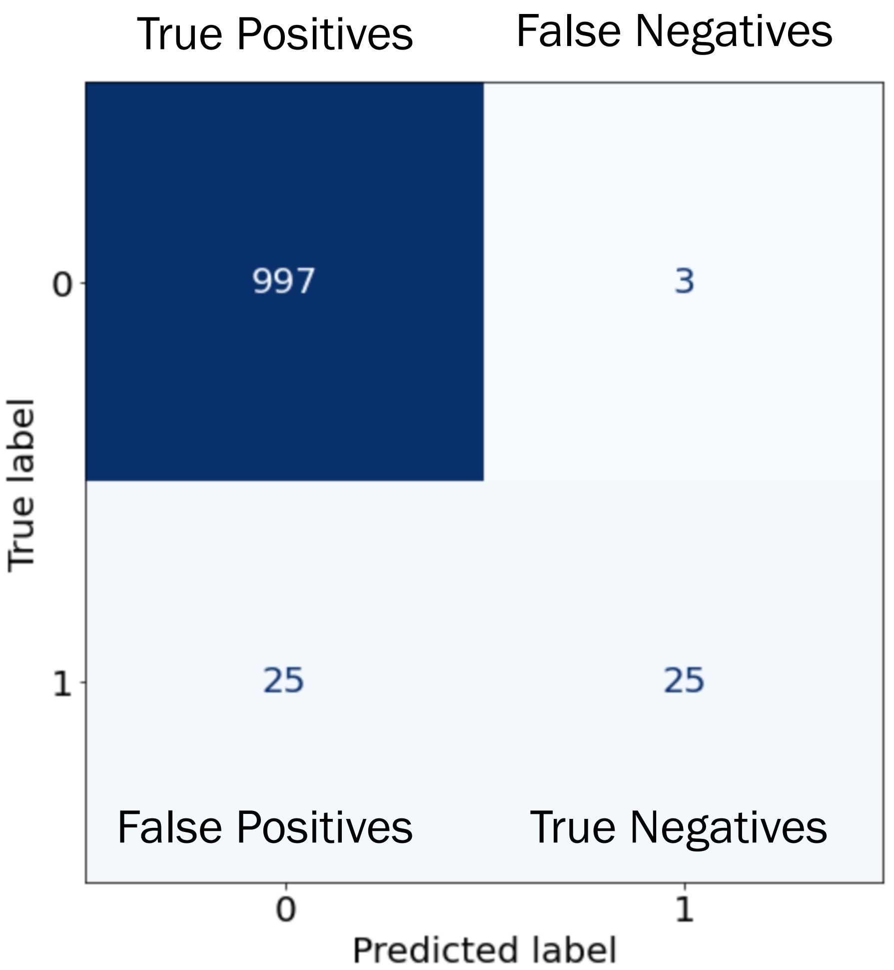
= 0 %

Balanced Accuracy

$$a_B = 50\%$$

Class 0:1000 – Class 1:50 – Confusion Matrix





SENSITIVITY

$$\frac{TP}{TP+FN} = \frac{TP}{P}$$

= 99.7 %

SPECIFICITY

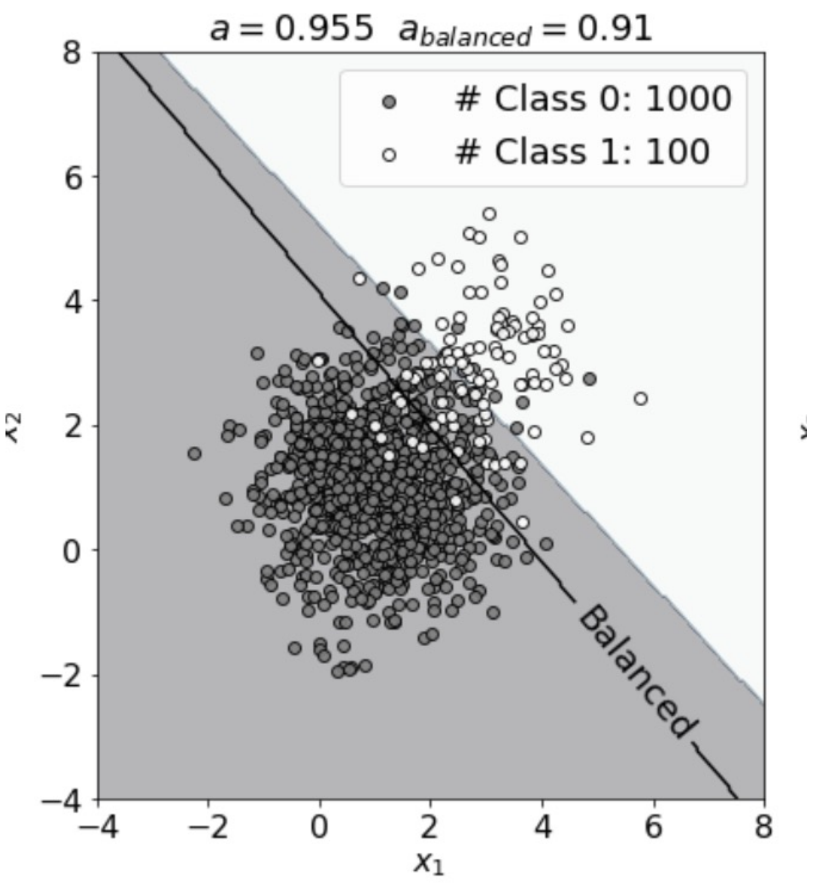
$$\frac{TN}{TN+FP} = \frac{TN}{N}$$

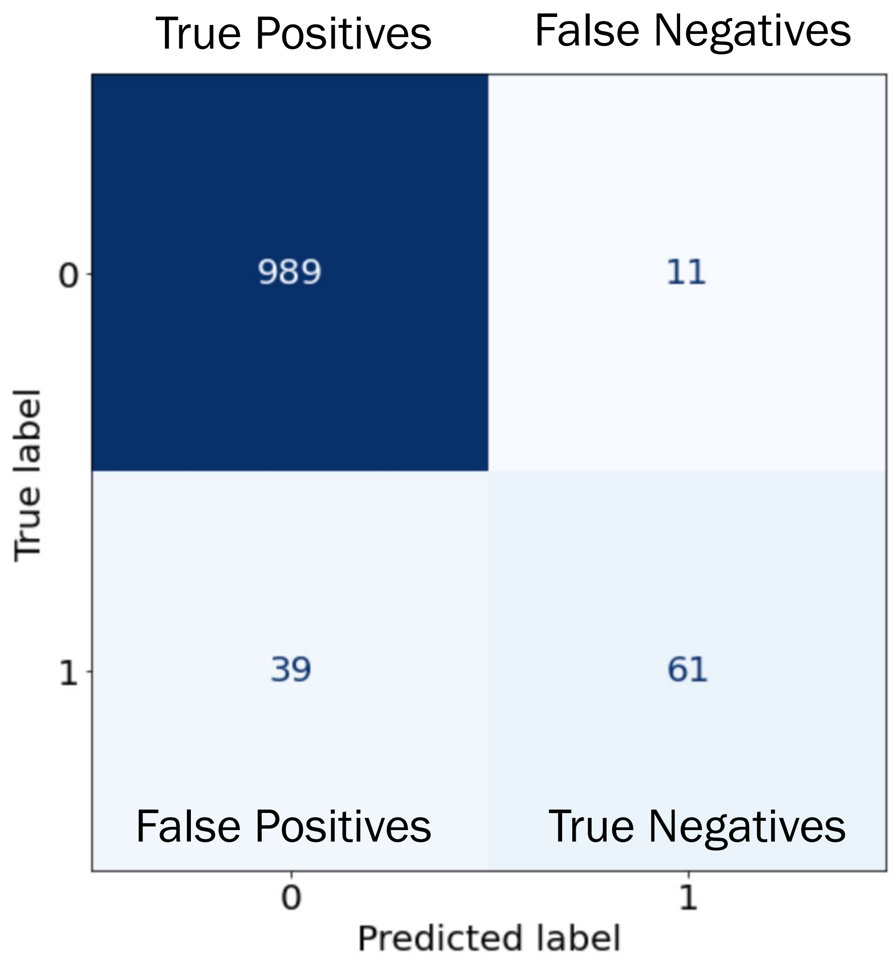
= 50 %

Balanced Accuracy

$$a_B = 74.9\%$$

Class 0:1000 – Class 1:100 – Confusion Matrix





SENSITIVITY

$$\frac{TP}{TP+FN} = \frac{TP}{P}$$

= 98.9 %

SPECIFICITY

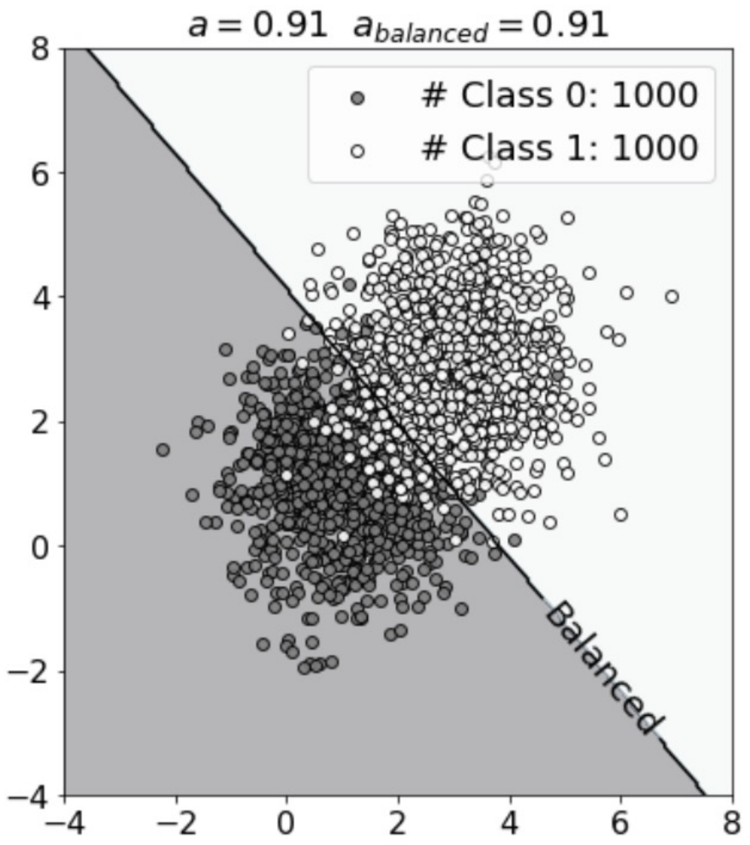
$$\frac{TN}{TN+FP} = \frac{TN}{N}$$

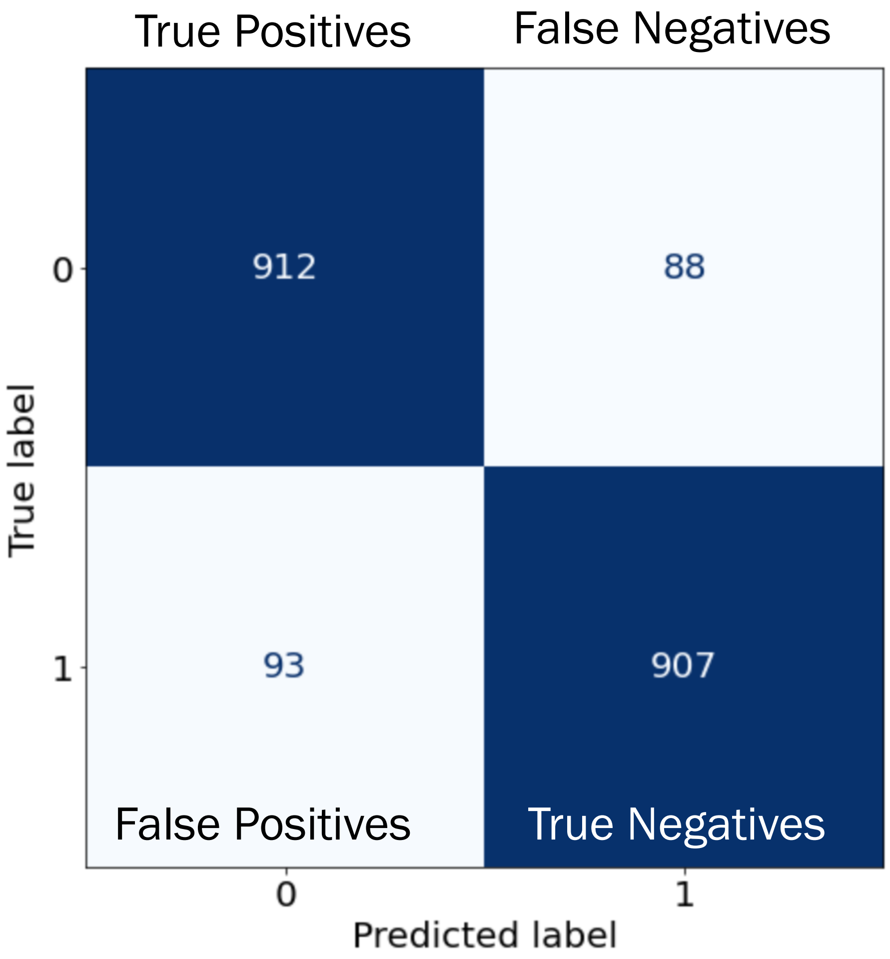
= 61 %

Balanced Accuracy

$$a_B = 80\%$$

Class 0:1000 – Class 1:1000 – Confusion Matrix





SENSITIVITY

$$\frac{TP}{TP+FN} = \frac{TP}{P}$$

= 91.2 %

SPECIFICITY

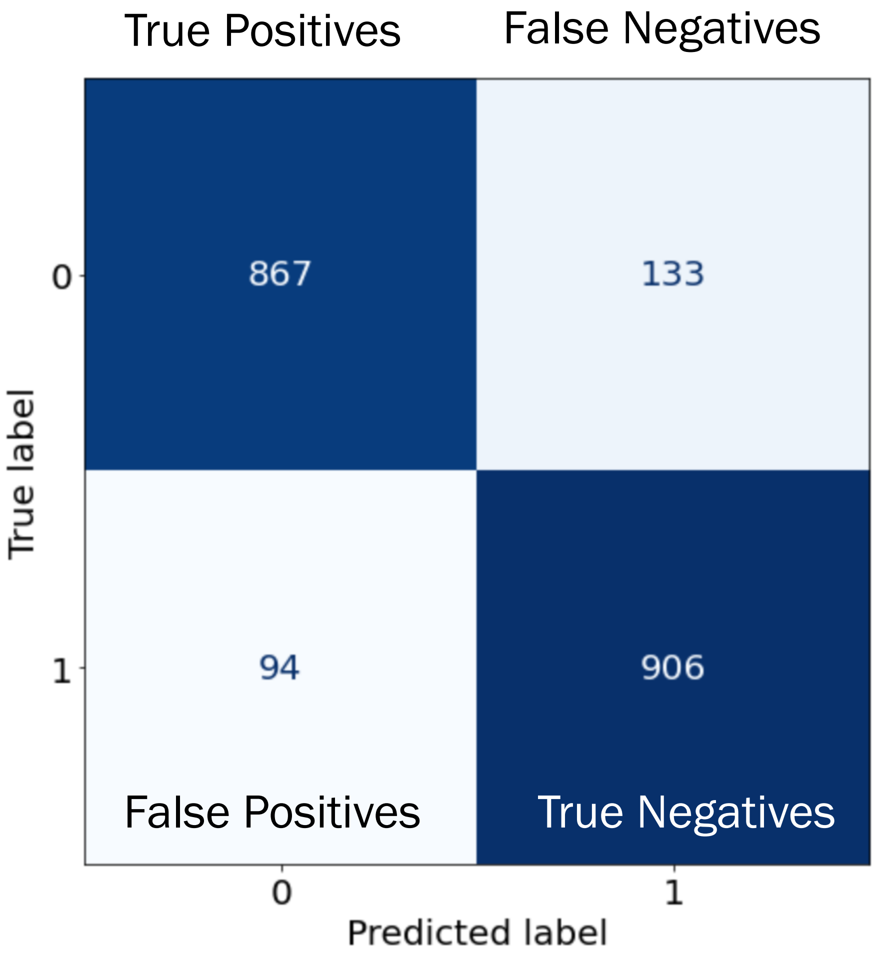
$$\frac{TN}{TN+FP} = \frac{TN}{N}$$

= 90.7 %

Balanced Accuracy

$$a_B = 91\%$$

Class 0:1000 – Class 1:10 – Confusion Matrix (mit vs. ohne Oversampling)



WITH OVERSAMPLING

SENSITIVITY

$$\frac{TP}{TP+FN} = \frac{TP}{P}$$

= 86.7 %

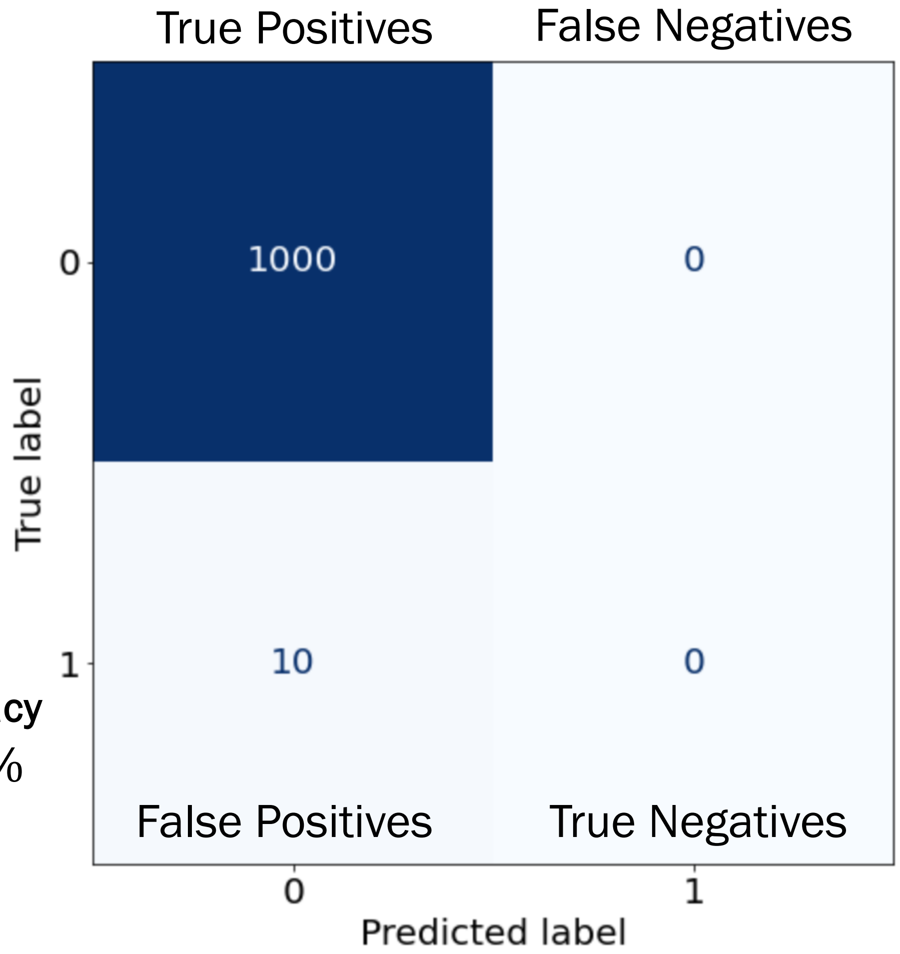
SPECIFICITY

$$\frac{TN}{TN+FP} = \frac{TN}{N}$$

= 90.6 %

Balanced Accuracy

$a_B = 88.7\%$



WITHOUT OVERSAMPLING

SENSITIVITY

$$\frac{TP}{TP+FN} = \frac{TP}{P}$$

= 100 %

SPECIFICITY

$$\frac{TN}{TN+FP} = \frac{TN}{N}$$

= 0 %

Balanced Accuracy

$a_B = 50\%$

Real-life Scenarios

Study on hospital-acquired infections: "**Out of 683 patients, only 75 (11% of the total) were infected and 608 were not**" (Cohen, Gilles, et al. "Learning from imbalanced data in surveillance of nosocomial infection." *Artificial intelligence in medicine* 37.1 (2006): 7-18)

Table 1 Baseline performance (original class distribution: 0.11 pos, 0.89 neg)

Classifier	Sensitivity	Specificity	a_B	CWA	Accuracy
IB1 (kNN)	0.19	0.96	0.58	0.38	0.88
Nave Bayes	0.57	0.88	0.73	0.65	0.85
C4.5 (Decision Trees)	0.28	0.95	0.62	0.45	0.88
AdaBoost	0.45	0.95	0.70	0.58	0.90
SVM	0.43	0.92	0.68	0.55	0.86

* CWA – Class Weighted Accuracy (proposed in the paper)

Real-life Scenarios

- Study on hospital-acquired infections: “**Out of 683 patients, only 75 (11% of the total) were infected and 608 were not**” (Cohen, Gilles, et al. "Learning from imbalanced data in surveillance of nosocomial infection." *Artificial intelligence in medicine* 37.1 (2006): 7-18)

Table 2 Random subsampling and oversampling (0.5 pos, 0.5 neg)

Classifier	(a) Random subsampling				(b) Random oversampling			
	Sens	Spec	CWA	Accu	Sens	Spec	CWA	Accu
IB1 (kNN)	0.01	0.99	0.26	0.88	0.19	0.96	0.38	0.88
Nave Bayes	0.21	0.96	0.40	0.88	0.68	0.83	0.72	0.81
C4.5 (Decision Trees)	0.00	1.00	0.25	0.89	0.49	0.87	0.59	0.83
AdaBoost	0.04	1.00	0.28	0.89	0.73	0.87	0.77	0.85
SVM	0.05	0.99	0.29	0.88	0.60	0.89	0.67	0.86

Example – Images and Neural Networks

Question

Are neural networks smarter? Can they handle unbalanced records?

The goal is to convince you that what we've seen can happen with almost any algorithm.

Problem Description

I want to develop a classifier that distinguishes the 1 from all the others.

 # Number 1s: 6742 (11.2%)
 # Number 0,2,3,4,5,6,7,8,9: ca. 53258 (88.8%)

Handwritten digits – 28x28 pixel images

Confusion Matrix 1 vs. all (MNIST)

	Predicted Label 1	Predicted Label 0,2,3,4,5,6,7,8,9
True Label 1	0	6742
True Label 0,2,3,4,5,6,7,8,9	0	53258

* Not relevant: Results of a neural network with 1 neuron with sigmoid activation function.

Receiving Operating Curve (ROC)

- The ROC curve is a very important method for studying binary classification metrics.

- It is used to derive the Area Under the Cruve (AUC) metric.

ROC Curve (I)

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate (TPR)
- False Positive Rate (FPR)

$$\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$\mathrm{FPR} = \frac{FP}{N} = \frac{FP}{FP + TN}$$

How can we draw the curve? A model has only one value for TPR and FPR!

https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

ROC Curve (II) - Derivation

Let us suppose we have a model that has, as output, the probability $\hat{y}_i$ of an observation x_i of being in class 1.

The input observation is classified according to the following rule[1)]

$$\begin{cases} Class\ 1\ if\ \hat{y}_i > \alpha \\ Class\ 0\ if\ \hat{y}_i \leq \alpha \end{cases}$$

With $\alpha \in [0,1]$ is a real number. Normally one chooses $\alpha = 0.5$.

[1)] For the more mathematical savy of you, this is the translated Heaviside step function.

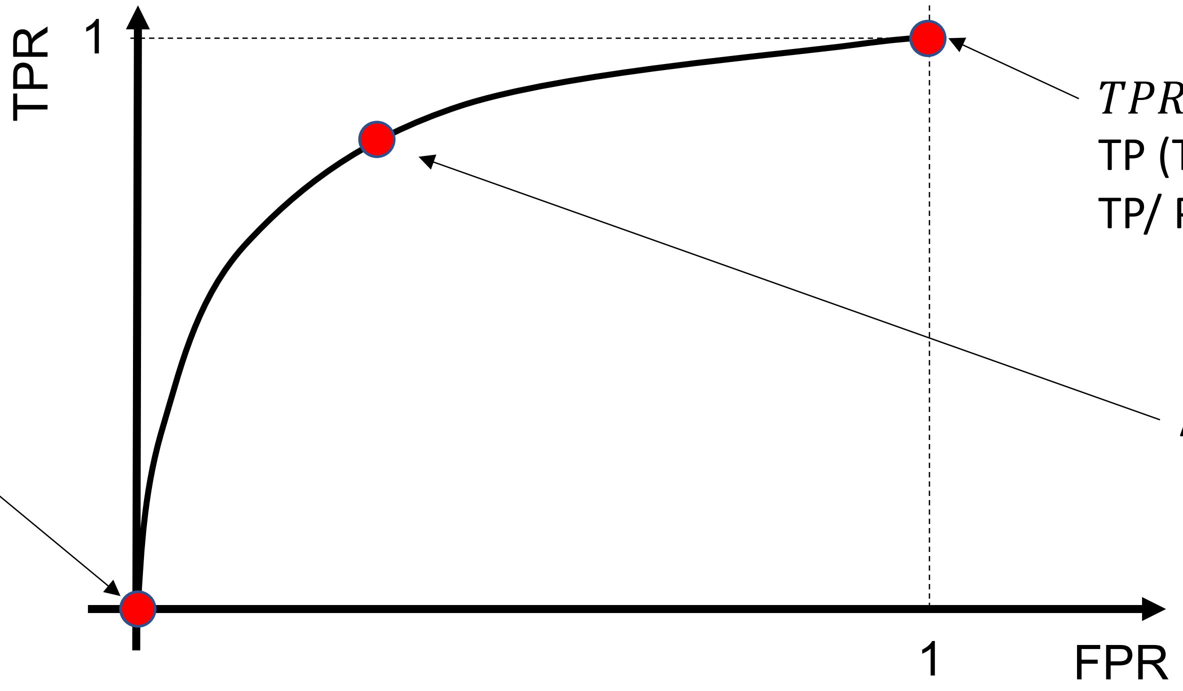
ROC Curve (III) - Derivation

For each value of α one get a value of TPR(α) and FPR(α).

The ROC curve is obtained by plotting TPR(α) and FPR(α) by varying α from 0 to 1.

$\alpha = 0$ – All observations Classified as in class 1

$\alpha = 1$ – All observations Classified as in class 0



$TPR = TP/P$
TP (True Positive) = P (Positive) → TP/ P = 1

$TPR = TP/P$
TP (True Positive) = 0 (Positive) → TP/ P = 0

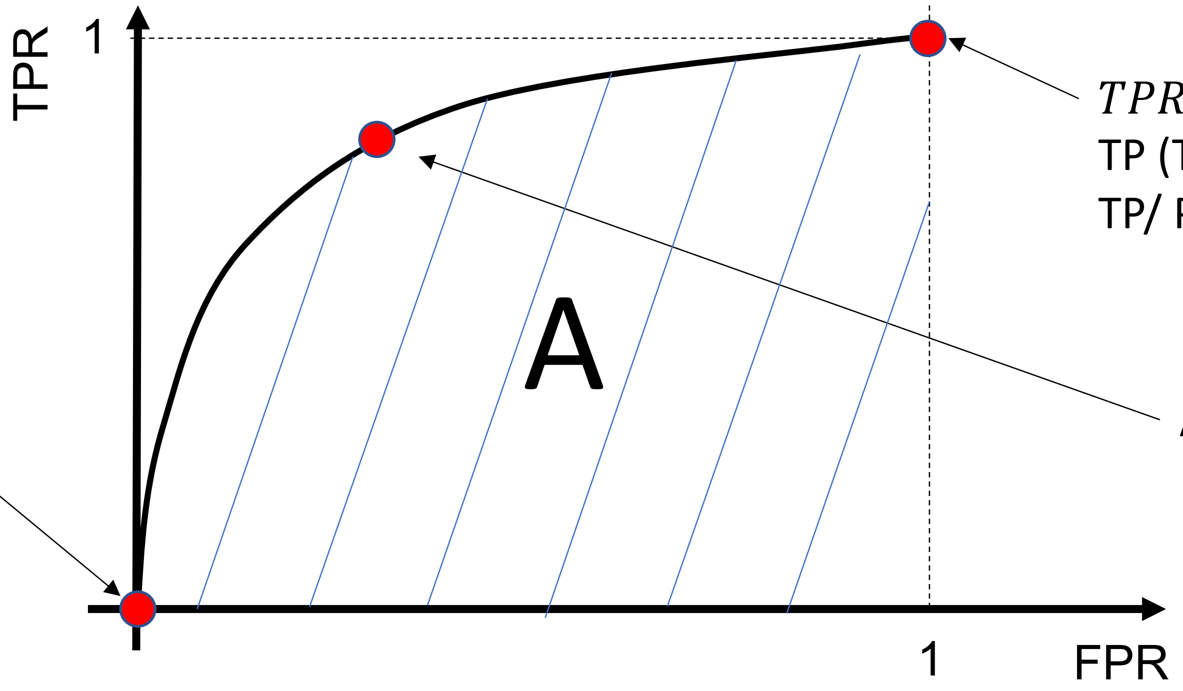
A value of α between 0 and 1

ROC Curve (IV) – AUC (Area Under the Curve)

To get a general metric on all possible cases (or possible α), the area under the curve (indicated with A) is often given as a metric.

$\alpha = 0$ – All observations Classified as in class 1

$\alpha = 1$ – All observations Classified as in class 0



$TPR = TP/P$
TP (True Positive) = P (Positive) → TP/ P = 1

$TPR = TP/P$
TP (True Positive) = 0 (Positive) → TP/ P = 0

A value of α between 0 and 1

Metrics for Multi-Class Problems

Multiclass classification is the problem of classifying instances into one of three or more classes.

e.g. diabetes type I, II or "gestational".

Metrics - Overview

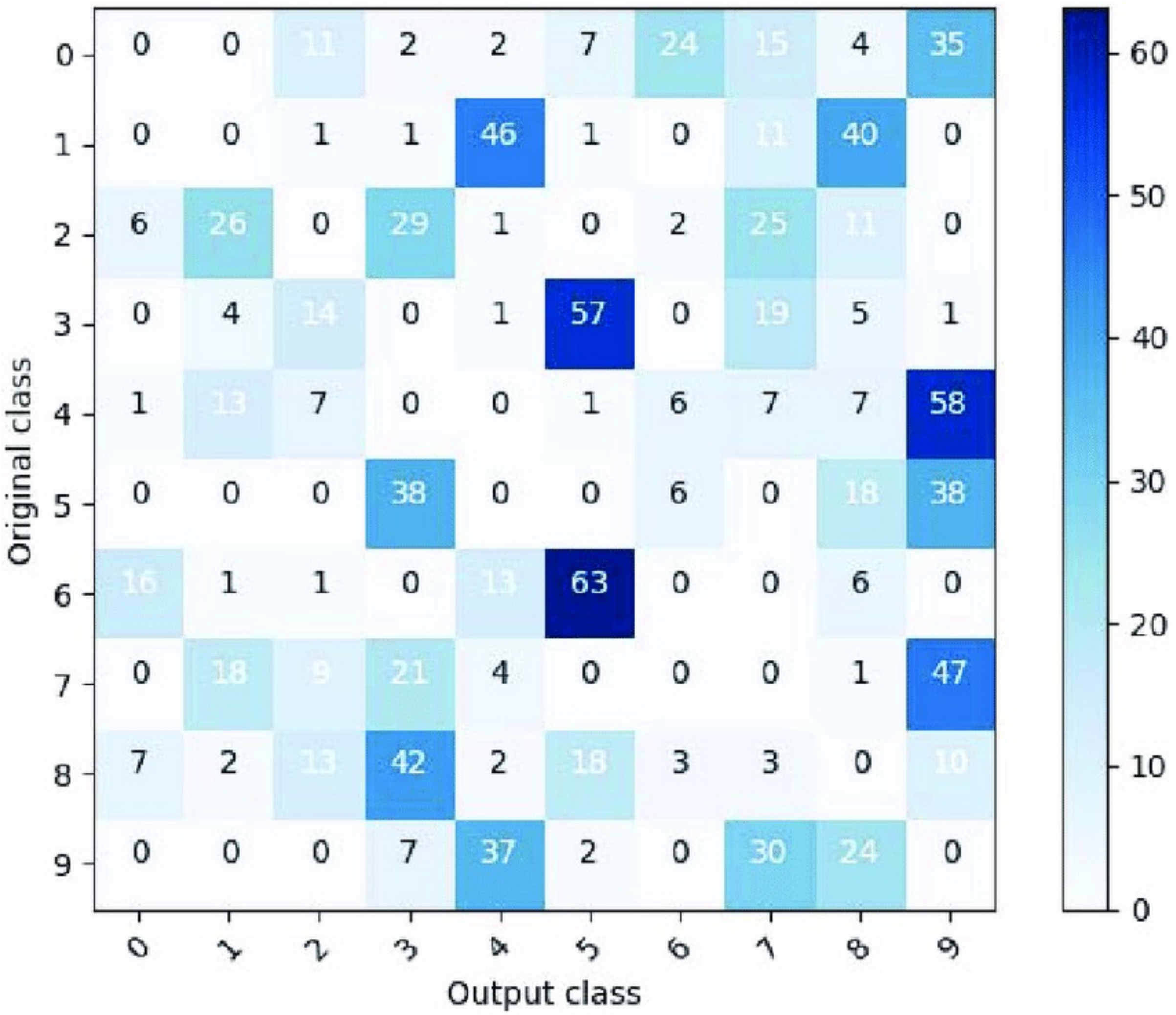
Binary Classification		Multi-class Classification	
Metric	**Discussed today**	**Metric**	
Accuracy	•	Accuracy	•
Specificity	•	Accuracy pro class	
Sensitivity	•		
Balanced Accuracy	•		
F1 Score	•		
Area Under The Curve (AUC / Receiving Operating Curve)	•		

- Confusion Matrix → Not really a metric (it contains multiple numbers)

Confusion Matrix (Multi-class classification)

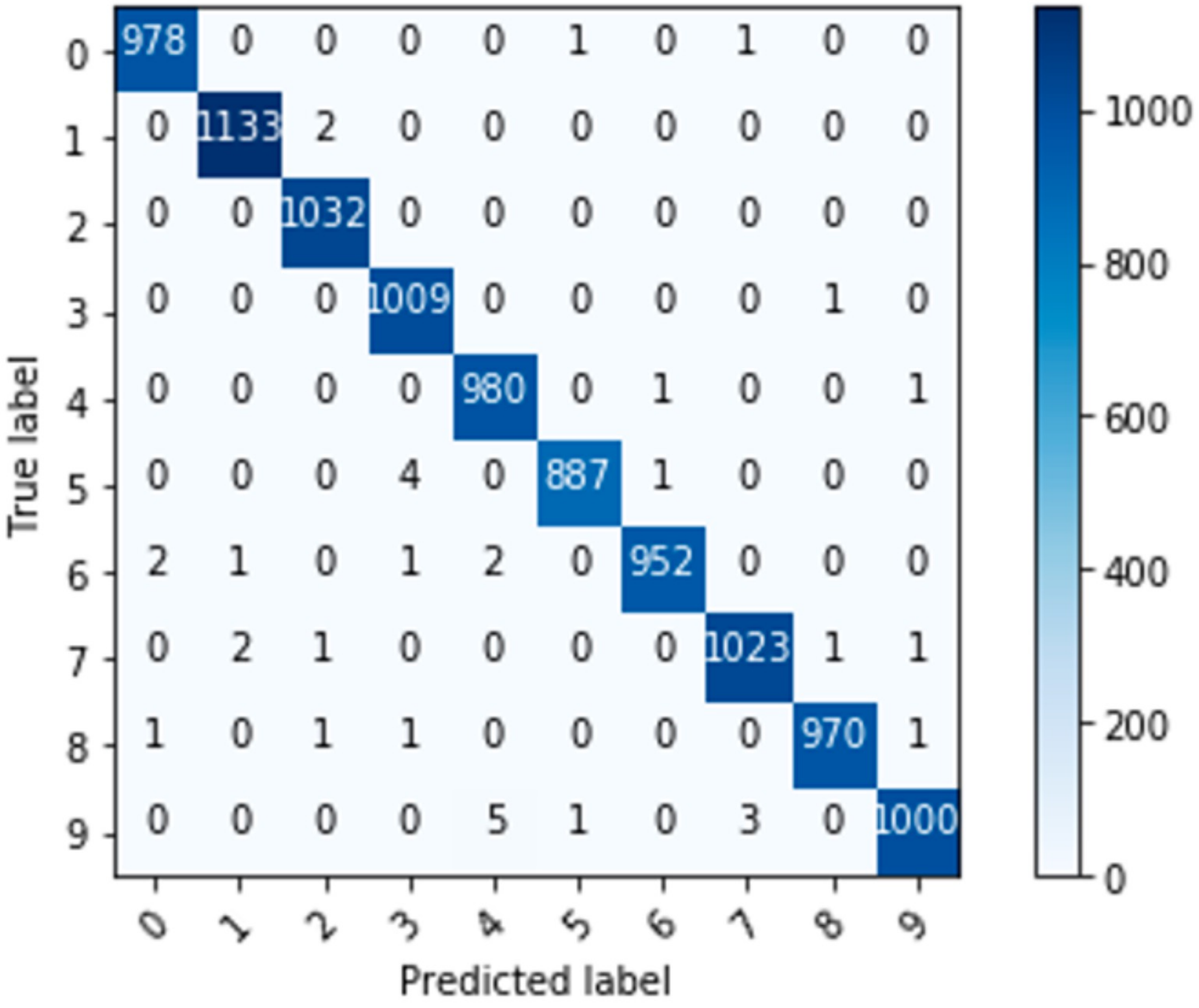
	Predicted Label 0	Predicted Label 1	Predicted Label 2	
True Label 0	Number of inputs classified as 0 and that have a true class of 0	Number of inputs classified as 1 and that have a true class of 0	Number of inputs classified as 2 and that have a true class of 0	…
True Label 1	Number of inputs classified as 0 and that have a true class of 1	Number of inputs classified as 1 and that have a true class of 1	⋱	
…	⋮	⋮	⋮	

Quiz: Classifier is good or bad?



Kwon, Hyun, et al. "Random untargeted adversarial example on deep neural network." *Symmetry* 10.12 (2018): 738.

Quiz: Classifier is good or bad?



Oyedotun, Oyebade K., Djamila Aouada, and Bjorn Ottersten. "Improving the capacity of very deep networks with maxout units." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Honorable Mention for Computer Vision Problems

This really works

Data Augmentation

Especially in computer vision, **data augmentation** (*a.k.a. generating new images from existing one, but slightly different*) is a really powerful techniques that will make your model performance better! We will see it when we will discuss computer vision techniques.