About me

prof. dr. Sven Degroeve

Master in Informatics, PhD in Computer Sciences

Computational Omics and Systems Biology Group (CompOmics, prof. dr. Lennart Martens)

Vakgroep Biomoleculaire Geneeskunde
VIB-UGent Center for Medical Biotechnology

Assistants

dr. Robbin Bouwmeester

dr. Ralf Gabriels

Toon Callens

Tine Clays

Alireza Nameni

Arthur Declercq

Course specifications

5 credits

contents:

linear regression, logistic regression
support vector machine, random forest, gradient boosting, ensemble learning
deep neural networks, AI
programming with scikit-learn
build and evaluate your own models

final competences:

The student is capable of interacting at high level with data analysis specialists.
The student is capable of understanding the specific literature on machine learning based data analysis.

Course specifications

5 credits

contents:

linear regression, logistic regression
~~support vector machine~~, random forest, gradient boosting, ensemble learning
deep neural networks, AI
programming with scikit-learn
build and evaluate your own models

final competences:

The student is capable of interacting at high level with data analysis specialists.
The student is capable of understanding the specific literature on machine learning based data analysis.

Course specifications

calculation of the examination mark:

(continuous evaluation)
Jupyter notebook project (15%)
microteaching: teach other students about a machine learning method not discussed during the lectures (15%)
2 page report about the obtained Kaggle contest results (20%)

(period-specific evaluation)
written exam on 50% of the total score

Course specifications

Jupyter notebook project (15%)

Intro to Machine Learning
Intermediate Machine Learning

certificate send to me by email
sven.degroeve@ugent.be
(subject: kaggle certificate)

Intro to Machine Learning

Learn the core ideas in machine learning, and build your first models.

Begin Course

3 hours to go

Course Discussion

Lessons

		Tutorial	Exercise
1	**How Models Work** The first step if you're new to machine learning.		
2	**Basic Data Exploration** Load and understand your data.		
3	**Your First Machine Learning Model** Building your first model. Hurray!		
4	**Model Validation** Measure the performance of your model, so you can test and compare alternatives.		
5	**Underfitting and Overfitting** Fine-tune your model for better performance.		
6	**Random Forests** Using a more sophisticated machine learning algorithm.		
7	**Machine Learning Competitions** Enter the world of machine learning competitions to keep improving and see your progress.		

Course specifications

Jupyter notebook project (15%)

Intro to Machine Learning
Intermediate Machine Learning

certificate send to me by email
sven.degroeve@ugent.be
(subject: kaggle certificate)

Intermediate Machine Learning

Handle missing values, non-numeric values, data leakage, and more.

Begin Course

4 hours to go

Course Discussion

Lessons

		Tutorial	Exercise
1	**Introduction** Review what you need for this course.		
2	**Missing Values** Missing values happen. Be prepared for this common challenge in real datasets.		
3	**Categorical Variables** There's a lot of non-numeric data out there. Here's how to use it for machine learning.		
4	**Pipelines** A critical skill for deploying (and even testing) complex models with pre-processing.		
5	**Cross-Validation** A better way to test your models.		
6	**XGBoost** The most accurate modeling technique for structured data.		
7	**Data Leakage** Find and fix this problem that ruins your model in subtle ways.		

Course specifications

Jupyter notebook project (15%)

Intro to Machine Learning
Intermediate Machine Learning

certificate send to me by email
sven.degroeve@ugent.be
(subject: kaggle certificate)

optional:
Data Visualization

Data Visualization

Make great data visualizations. A great way to see the power of coding!

Begin Course

4 hours to go

Course Discussion

Lessons

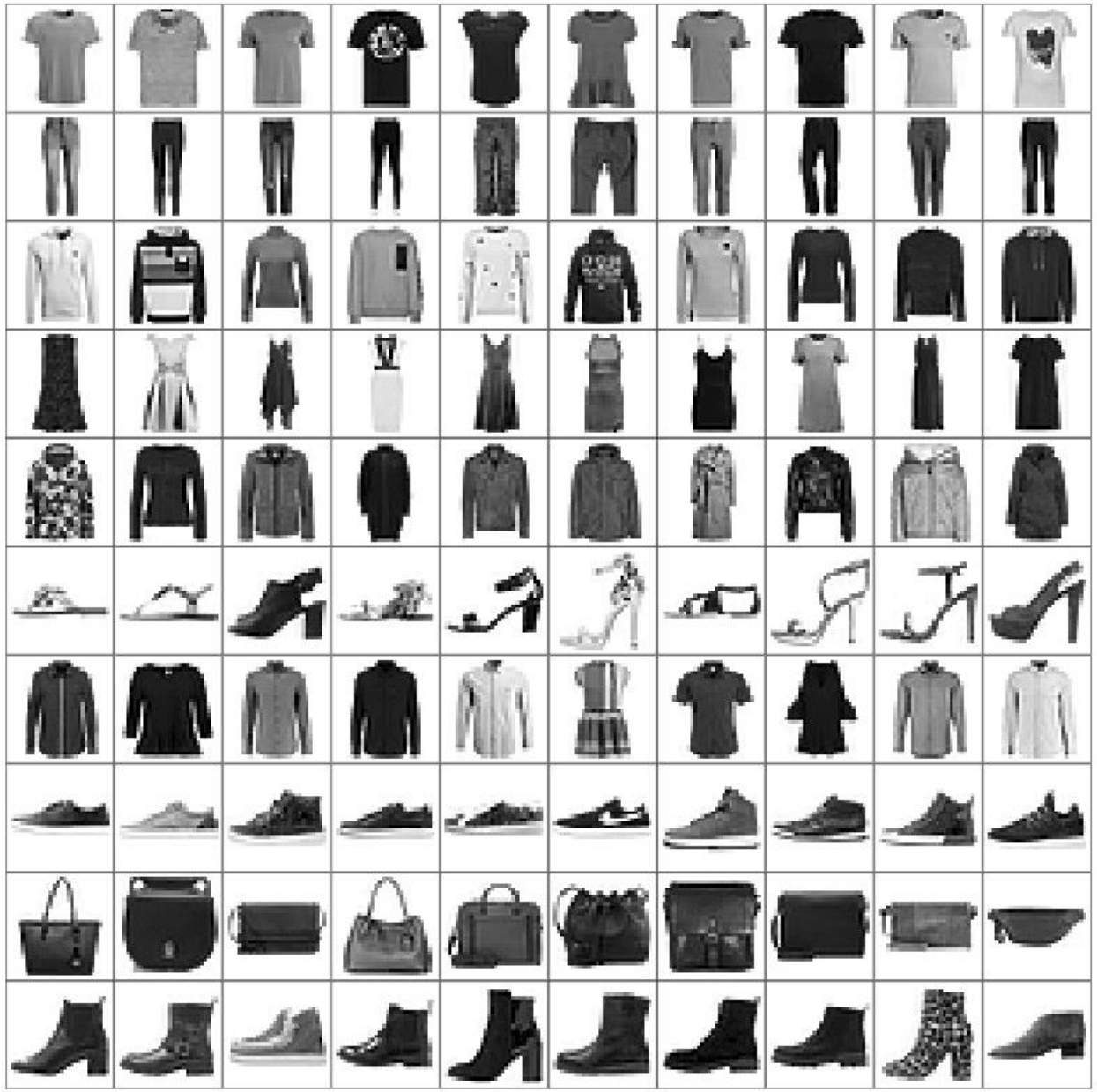
		Tutorial	Exercise
1	**Hello, Seaborn** Your first introduction to coding for data visualization		
2	**Line Charts** Visualize trends over time		
3	**Bar Charts and Heatmaps** Use color or length to compare categories in a dataset		
4	**Scatter Plots** Leverage the coordinate plane to explore relationships between variables		
5	**Distributions** Create histograms and density plots		
6	**Choosing Plot Types and Custom Styles** Customize your charts and make them look snazzy		
7	**Final Project** Practice for real-world application		
8	**Creating Your Own Notebooks** How to put your new skills to use for your next personal or work project		

Course specifications

2 page (Calibri (body) font, size 11, A4) report about the obtained Kaggle contest results (20%)

Course specifications

2 page (Calibri (body) font, size 11, A4) report about the obtained Kaggle contest results (20%)

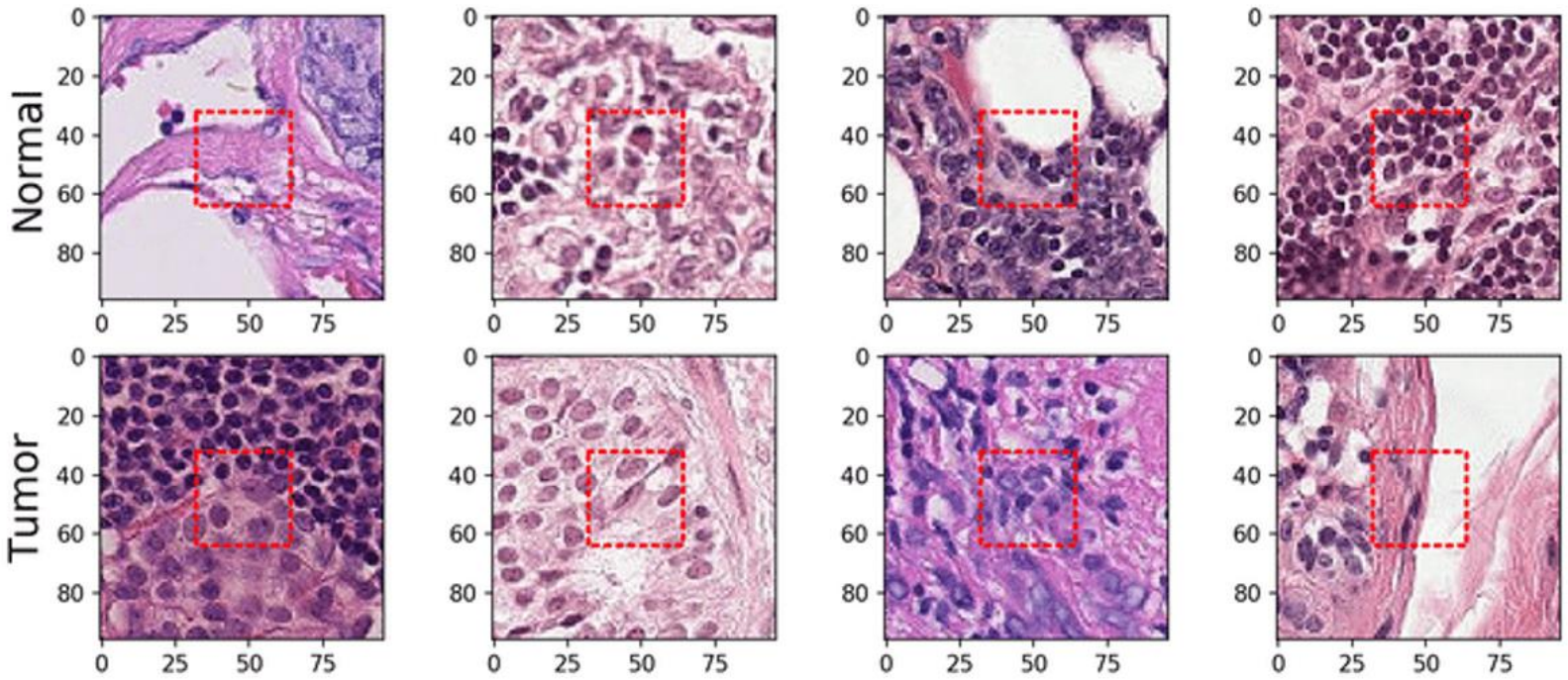


Course specifications

2 page (Calibri (body) font, size 11, A4) report about the obtained Kaggle contest results (20%)

histopathologic scans of lymph node sections

binary label indicating presence of metastatic tissue



Course specifications

2 page (Calibri (body) font, size 11, A4) report about the obtained Kaggle contest results (20%)

format: technical report

should contain:

- a clear and concise description of the method(s) used to fit your model(s)

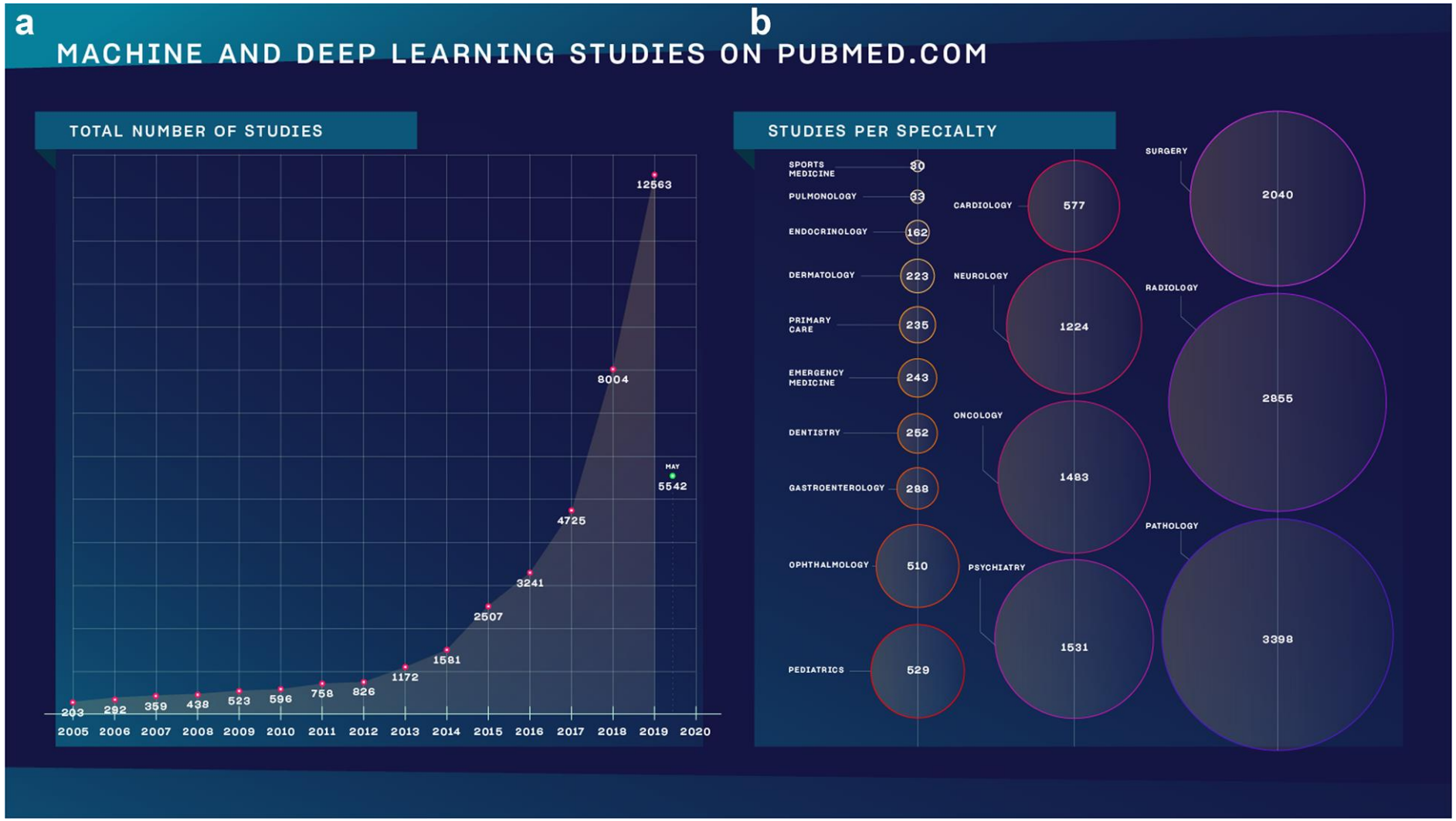
(show me you understand what you have done)

- a Table with optimal hyperparameter values
- a Table with evaluation results (both local CV and leaderboard test set)
- a section about what feature(s) are considered most important
- a short Discussion section
- clear plots are of course allowed as well!

Course specifications

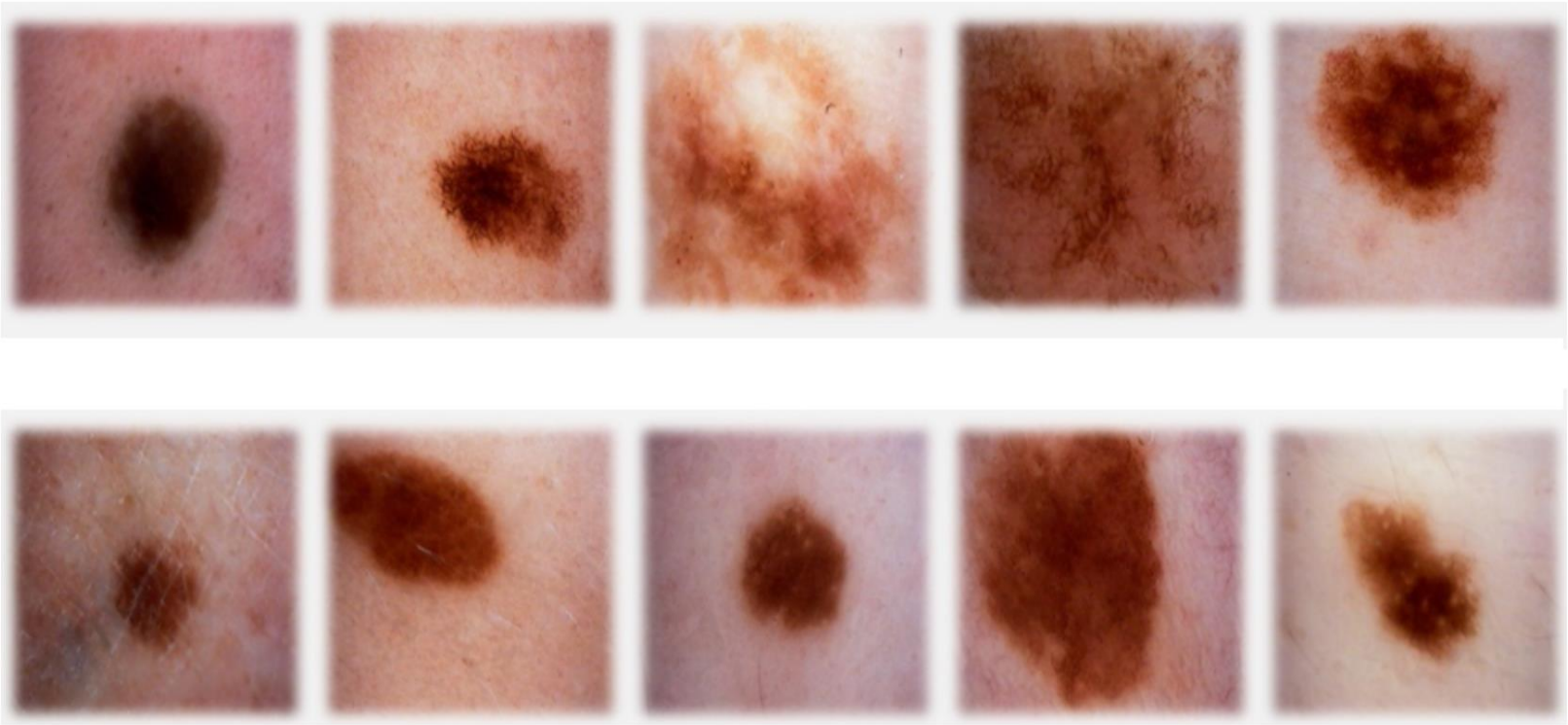
microteaching: teach other students about a machine learning method not discussed during the lectures (15%)

Why follow this course?



Meskó, B., Görög, M. A short guide for medical professionals in the era of artificial intelligence. npj Digit. Med. 3, 126 (2020)

classification

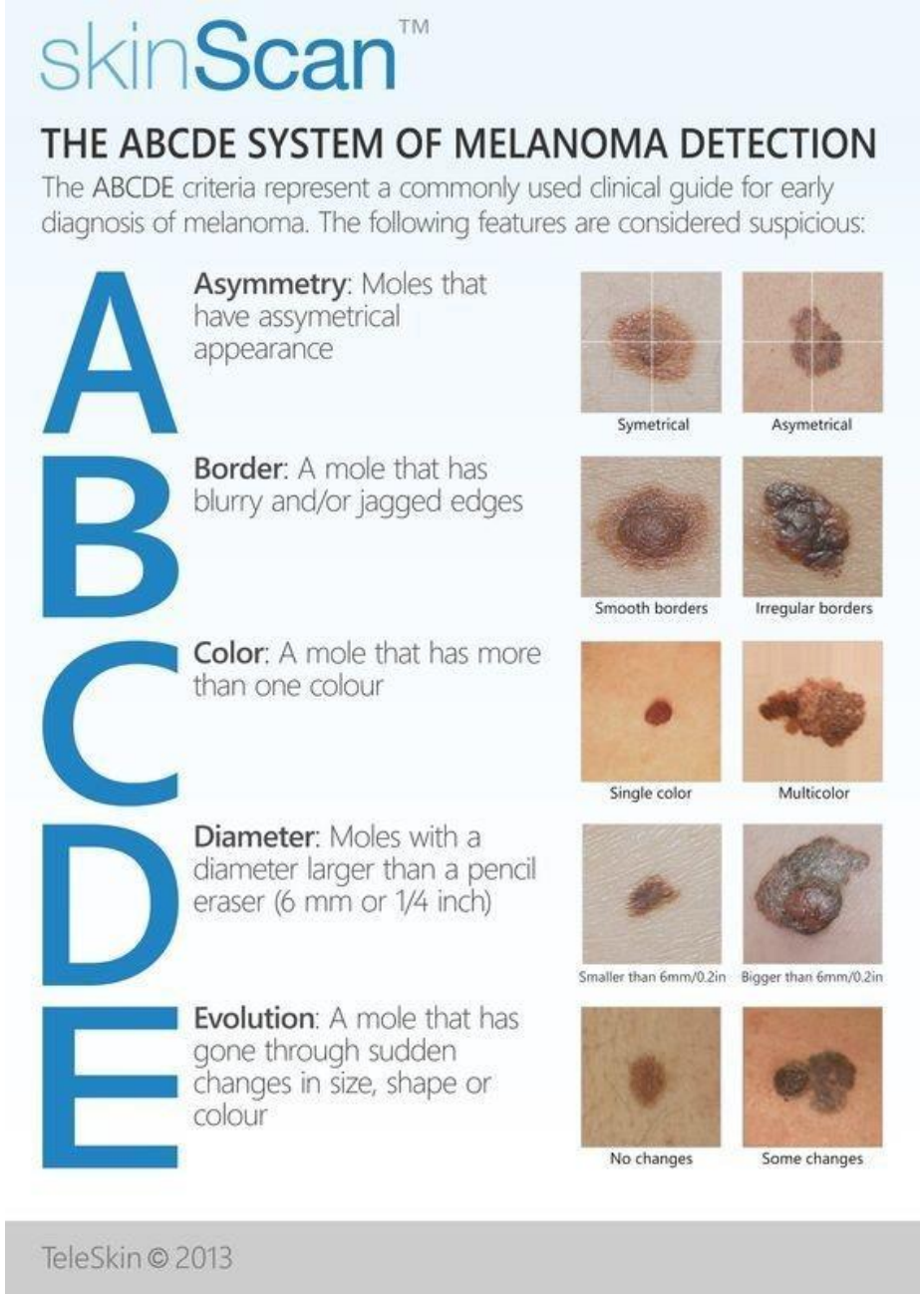


sign of cancer

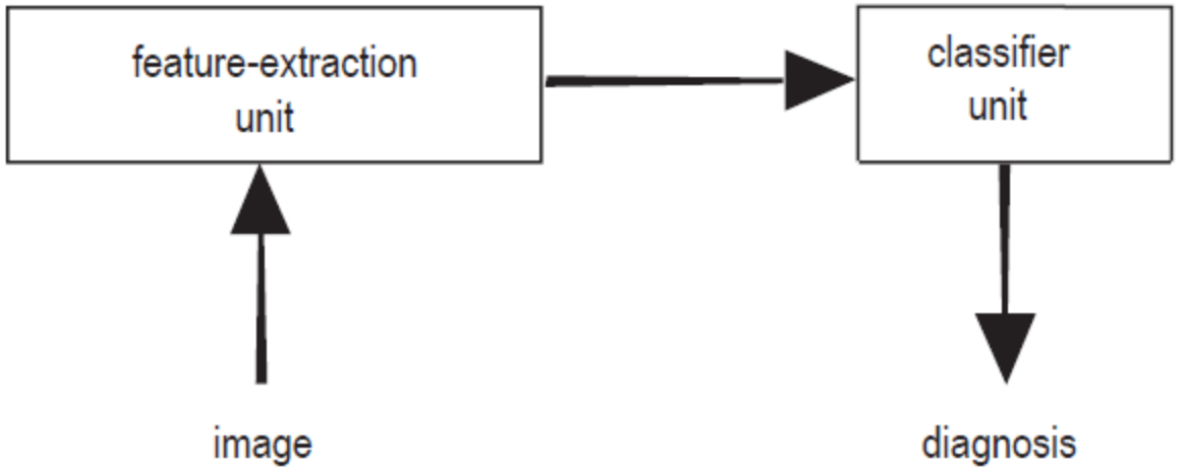
top row malignant

bottom row benign

classification



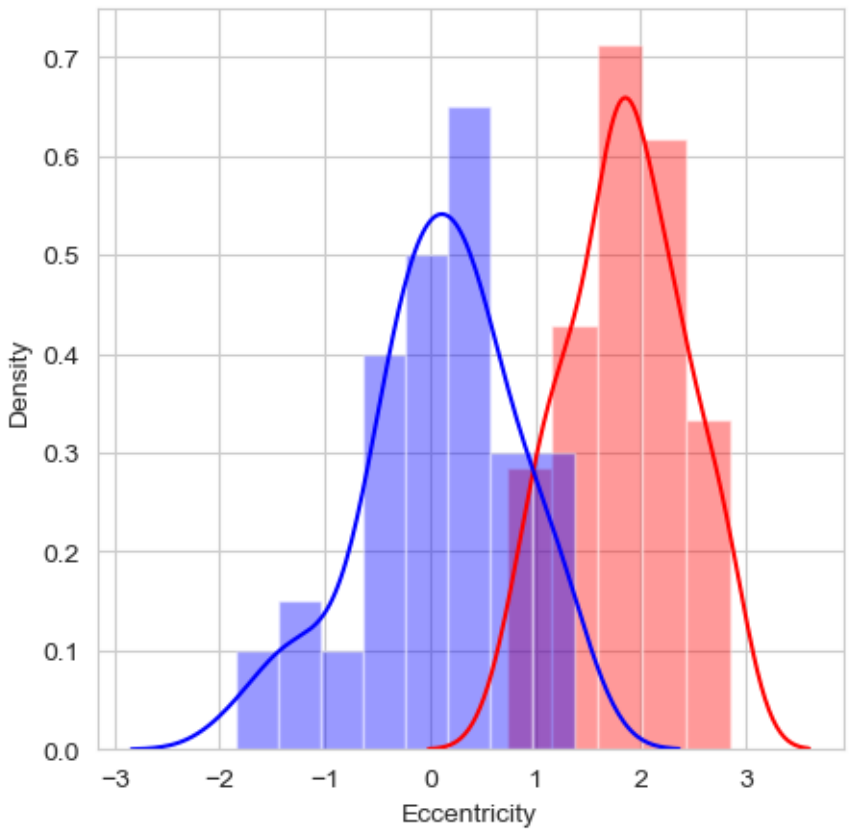
classification: terminology



feature extraction: features (a.k.a. properties or attributes)

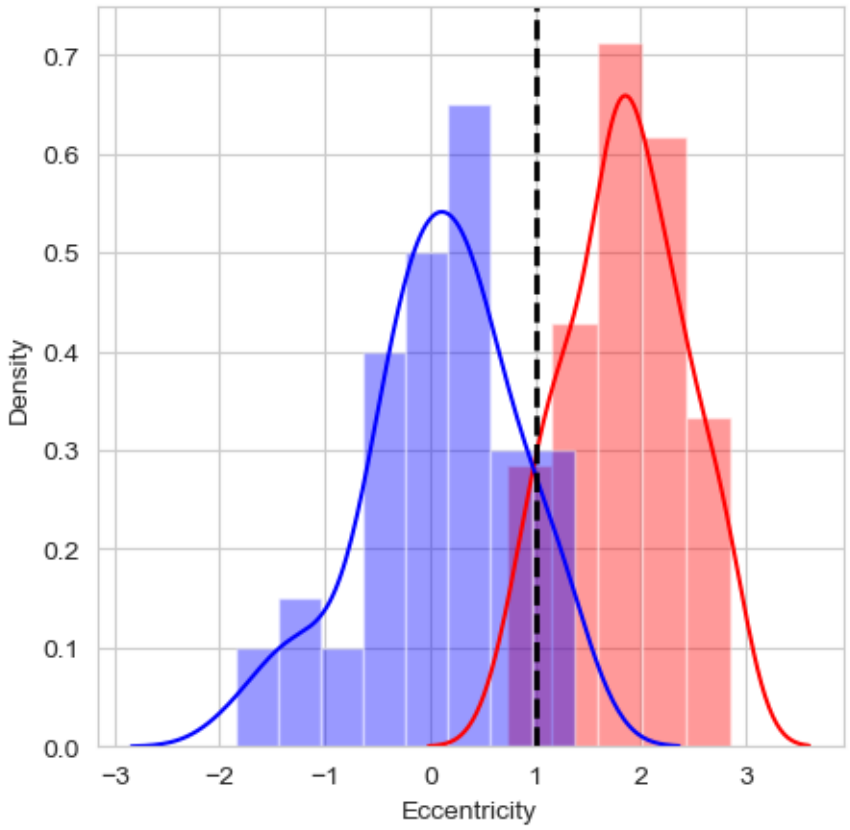
data set, sample (a.k.a. example, instance or data point), label (a.k.a. target)

classification: a feature



feature: eccentricity of lesion (how nearly circular the lesion is)

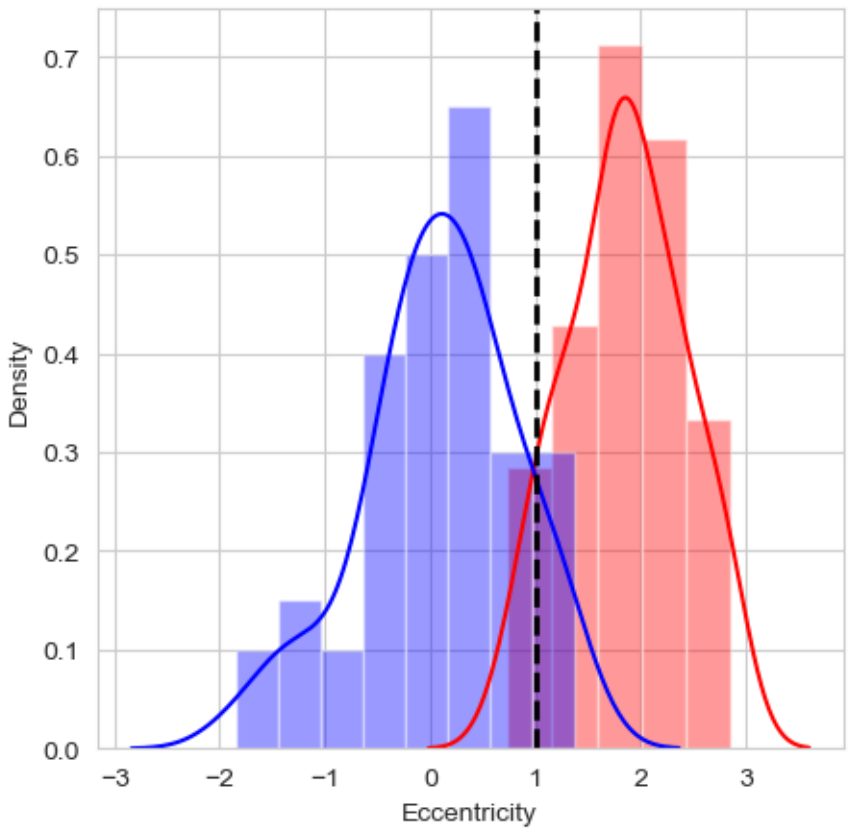
classification: the model

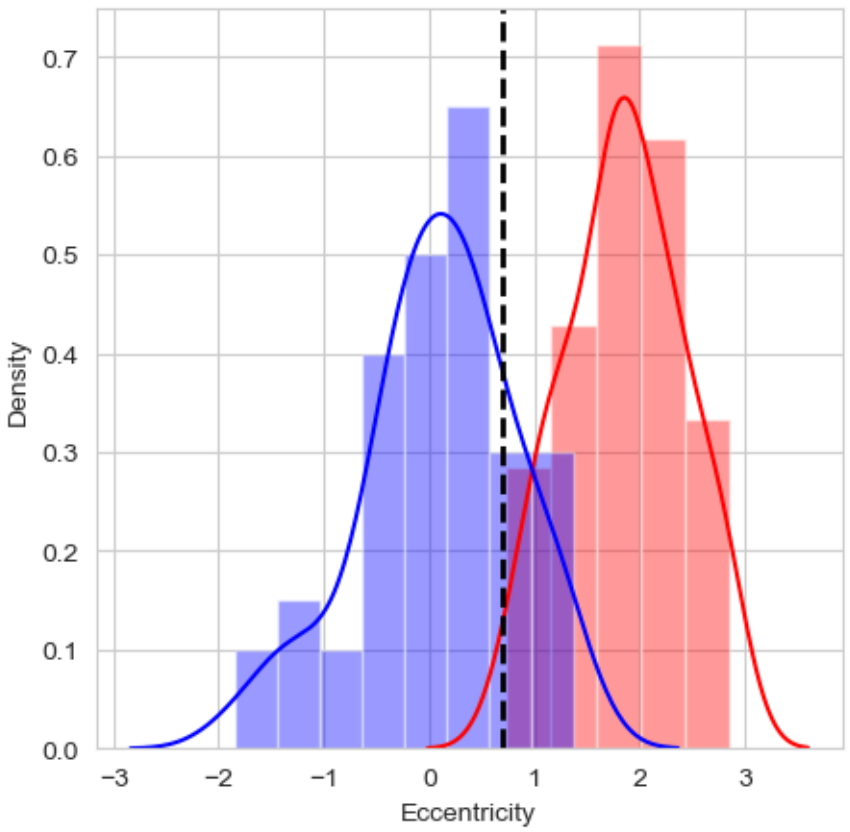


feature: eccentricity of lesion (how nearly circular the lesion is)

model: threshold ***t***

classification: the model





feature: eccentricity of lesion (how nearly circular the lesion is)

model: threshold ***t:*** consequence of the predictions

classification: prediction errors

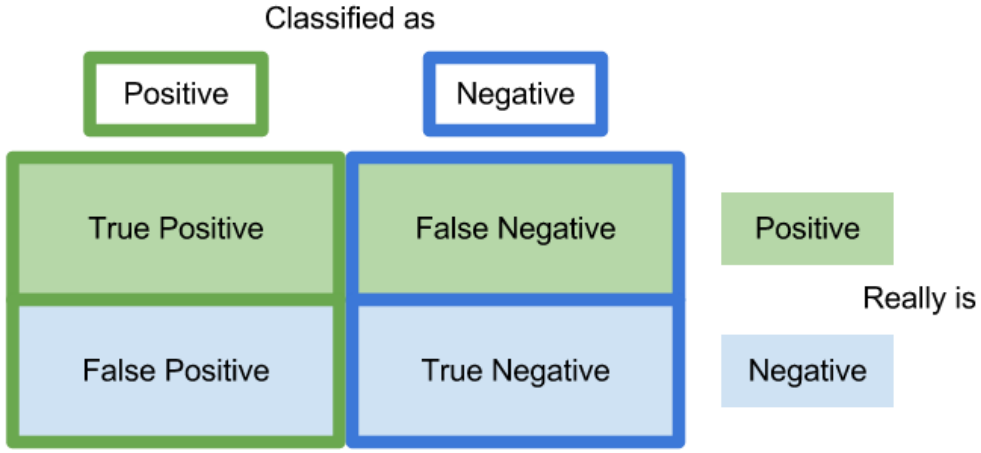
malignant: **positive** class
benign: **negative** class

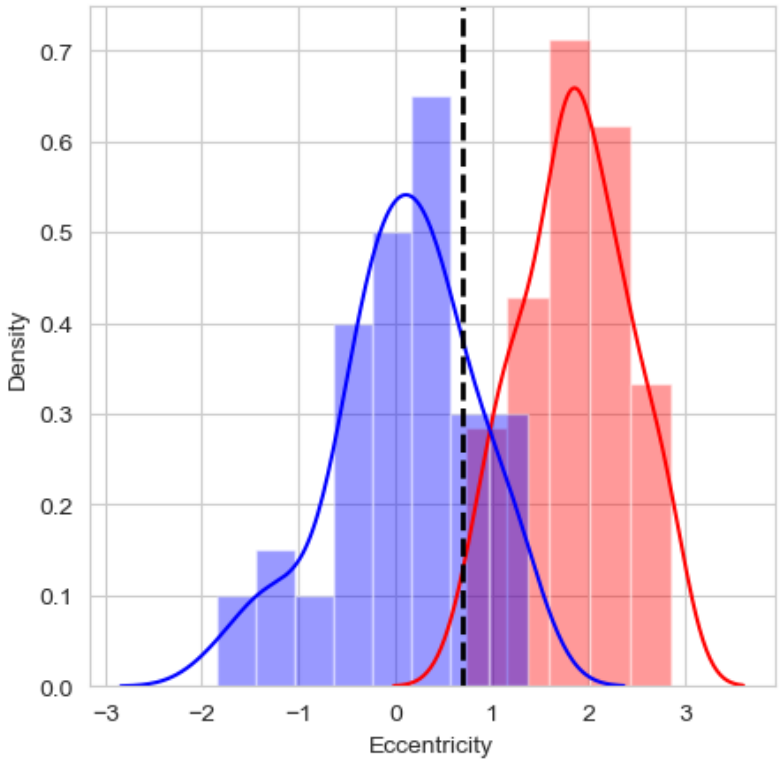
count the number of malignant images with eccentricity value $\geq t$: **true positive** predictions (TP)

count the number of malignant images with eccentricity value $< t$: **false negative** predictions (FN)

count the number of benign images with eccentricity value $\geq t$: **false positive** predictions (FP)

count the number of benign images with eccentricity value $< t$: **true negative** predictions (TN)





classification: prediction errors

	Classified as: Positive	Classified as: Negative
Really is: Positive	True Positive	False Negative
Really is: Negative	False Positive	True Negative

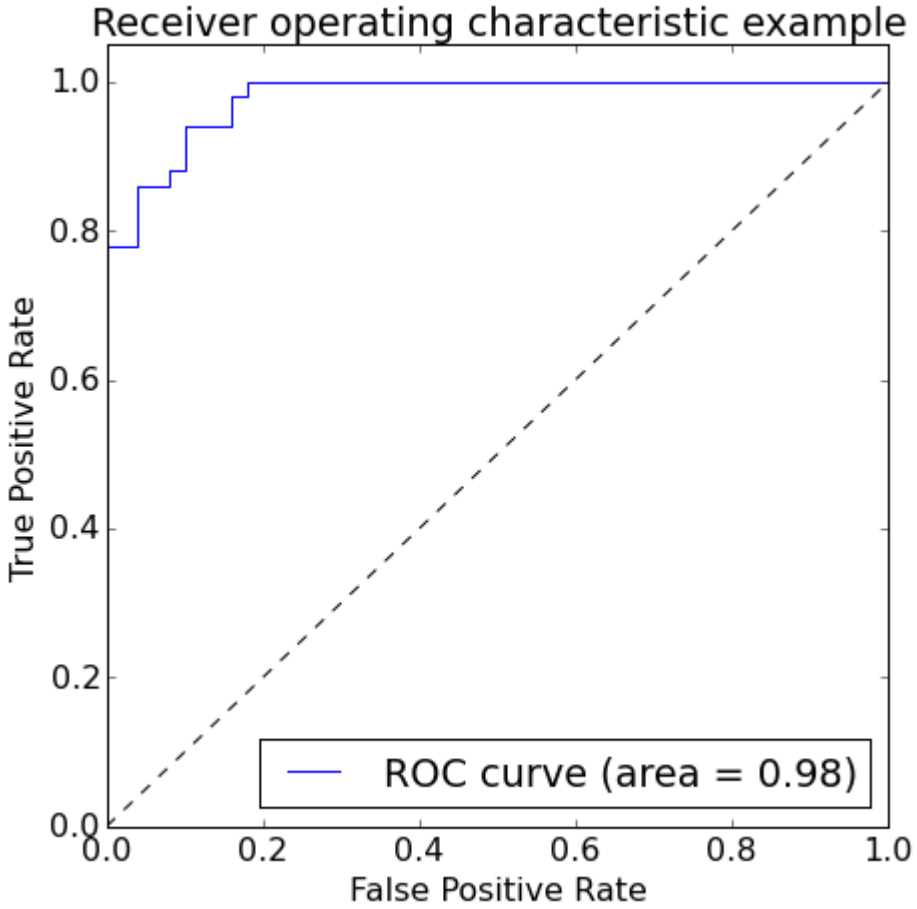
$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

classification: prediction errors

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



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

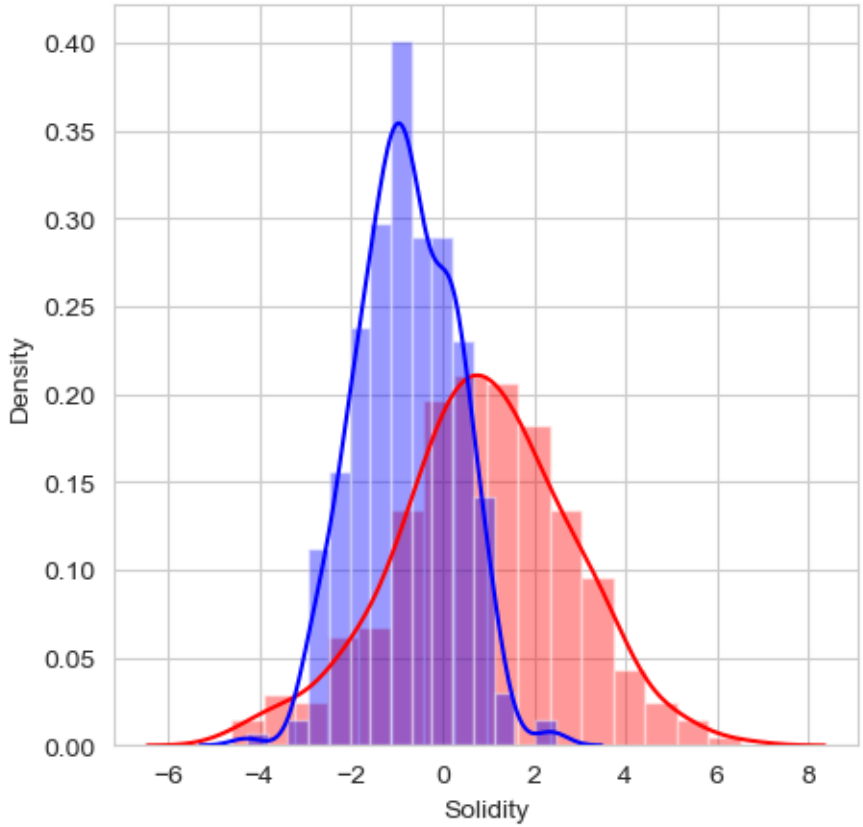
model that classifies all images as malignant:
TPR=1 and FPR=1

model that classifies all images a benign:
TPR=0 and FPR=0

vary threshold *t*

Area Under the Curve (AUC)

classification: multi-dimensional

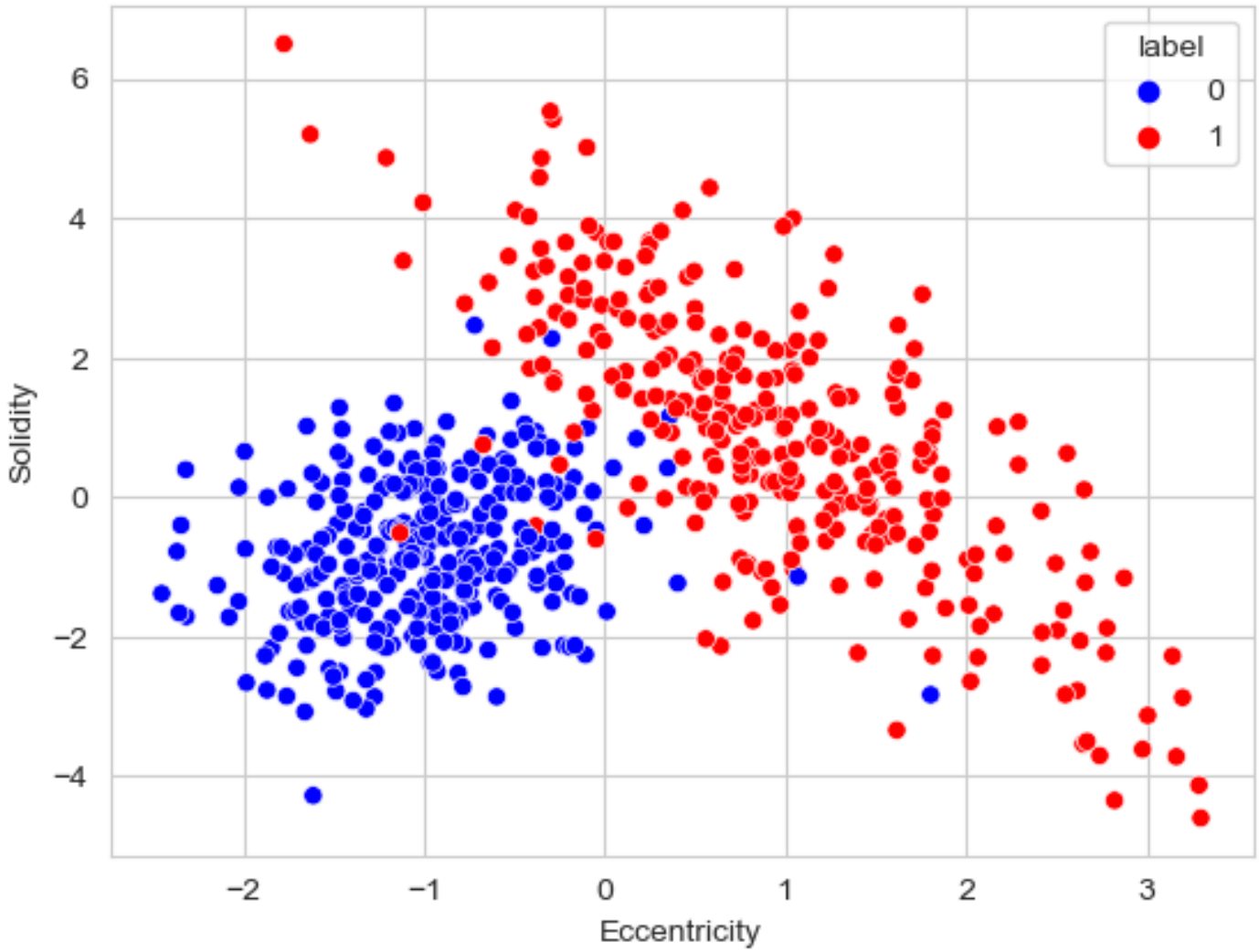


add another feature?

feature vector X

Euclidean vector space

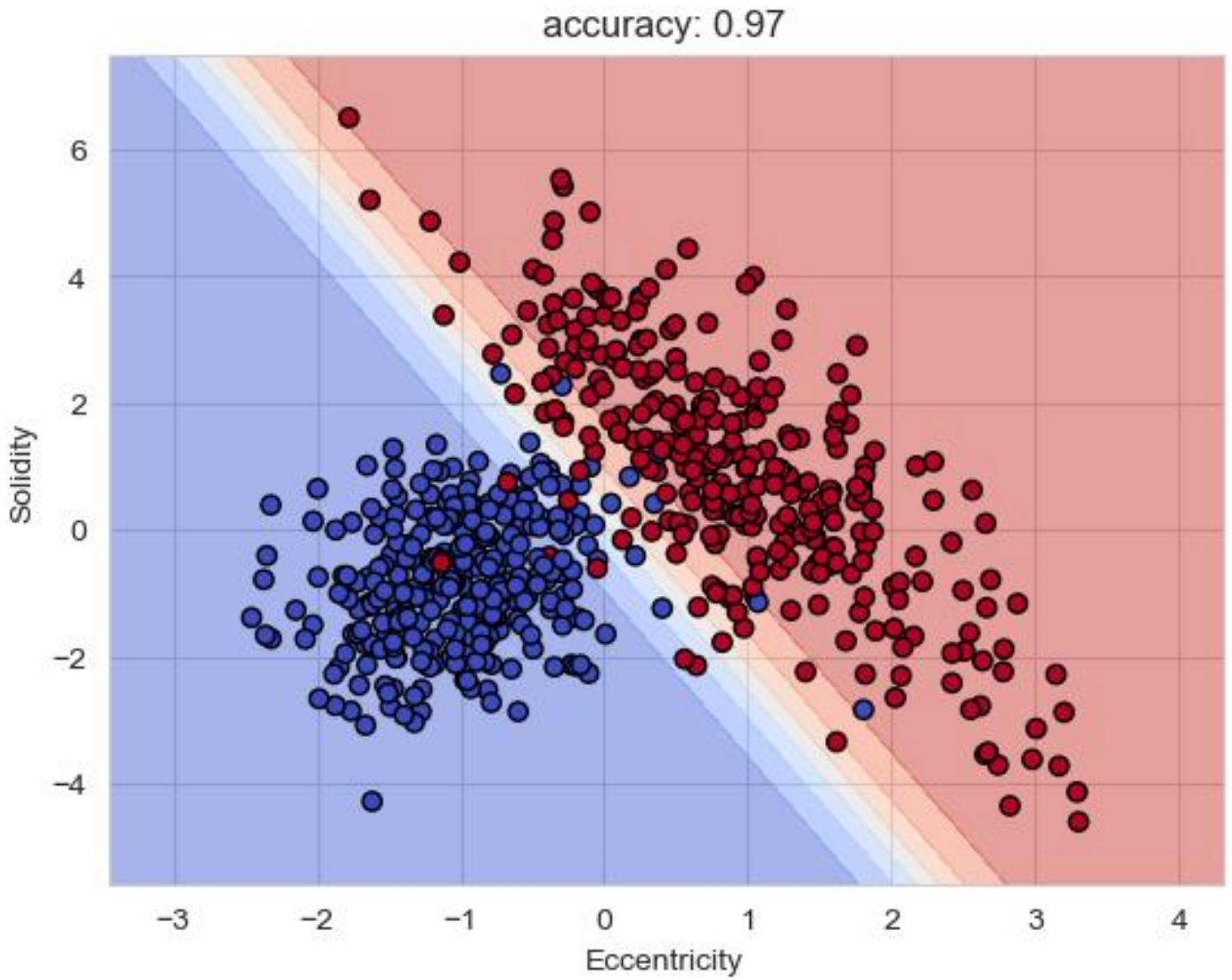
classification: multi-dimensional



feature vector X

Euclidean vector space

classification: multi-dimensional



linear decision boundary

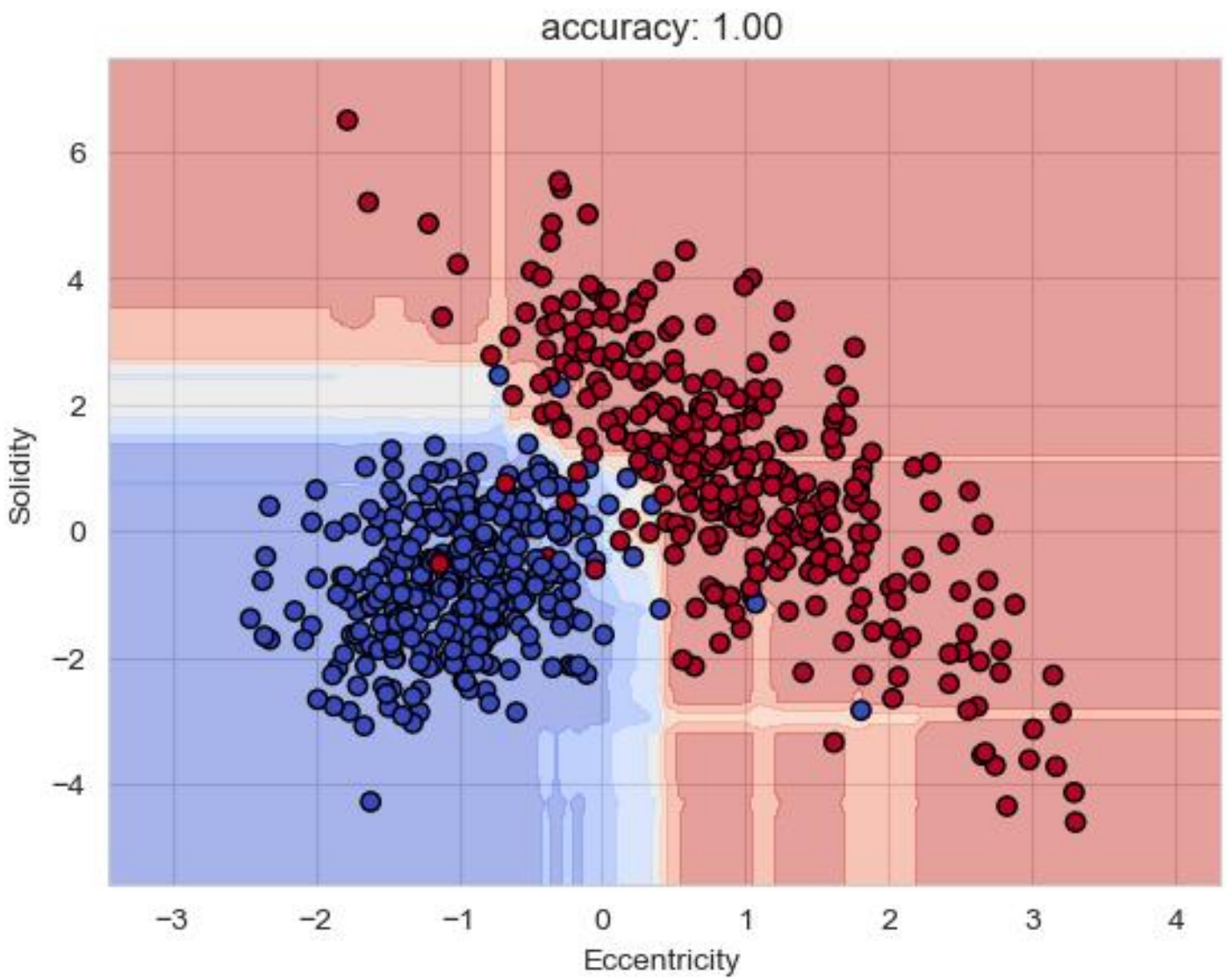
blue region malignant, red region benign

yet more features

can’t look at the decision boundary

more complex

classification: model complexity

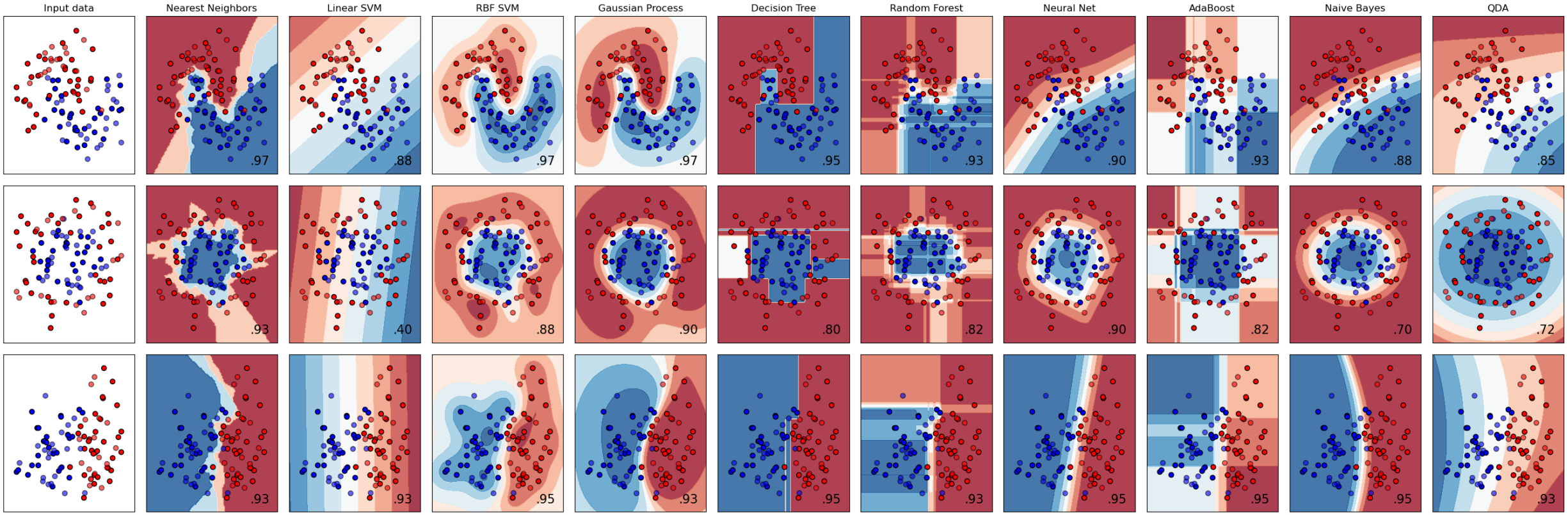


unseen external images

generalization

overfitting

scikit-learn



data normalization

make all features same scale

Eccentricity [0,100], Solidity [-5,7]

weights all features equally in their representation

standardization

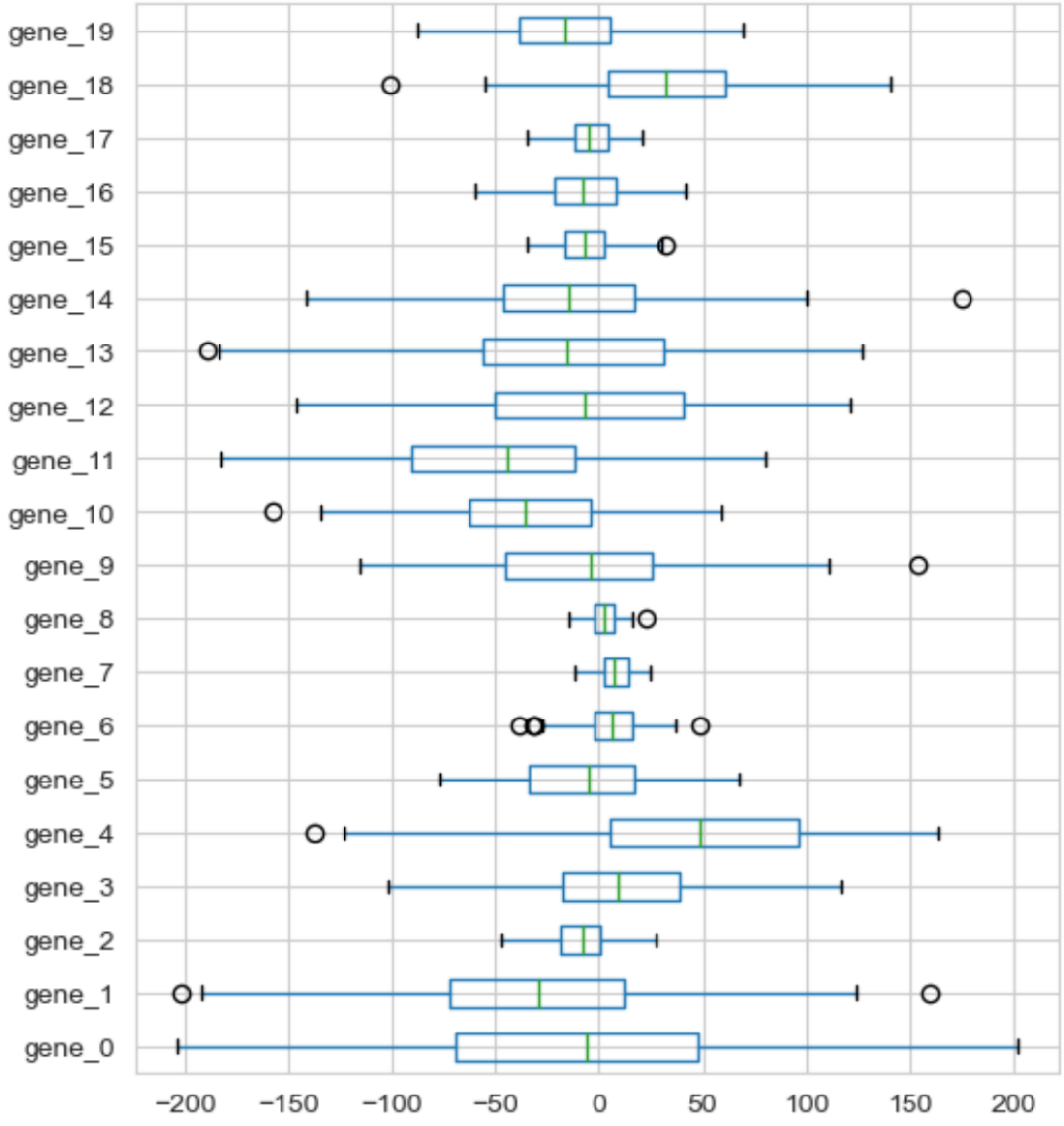
$$\mu = 0 \quad \sigma = 1$$

$$x_{norm} = \frac{x - \mu}{\sigma}$$

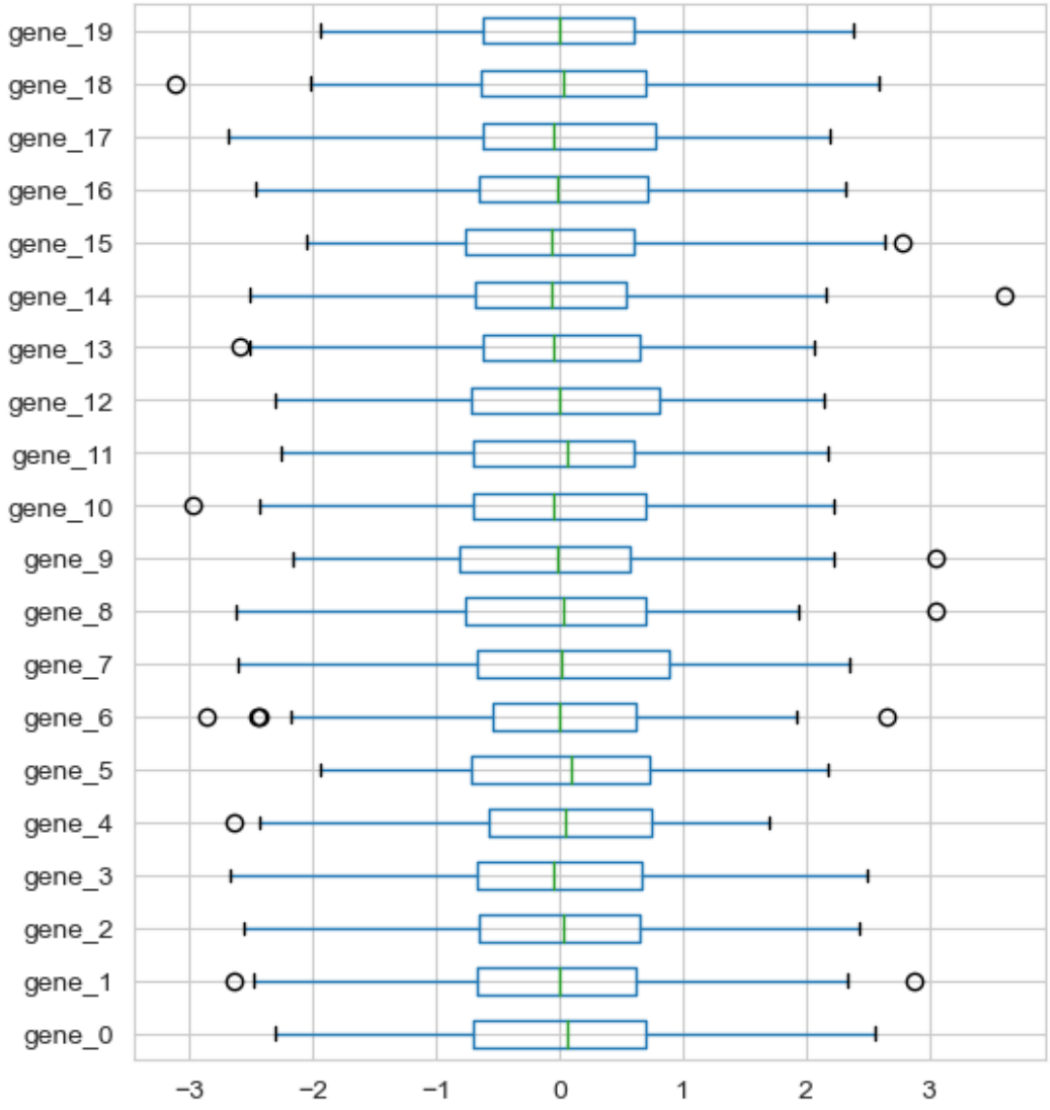
$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

min-max scaling: scale the features to a fixed range

data normalization

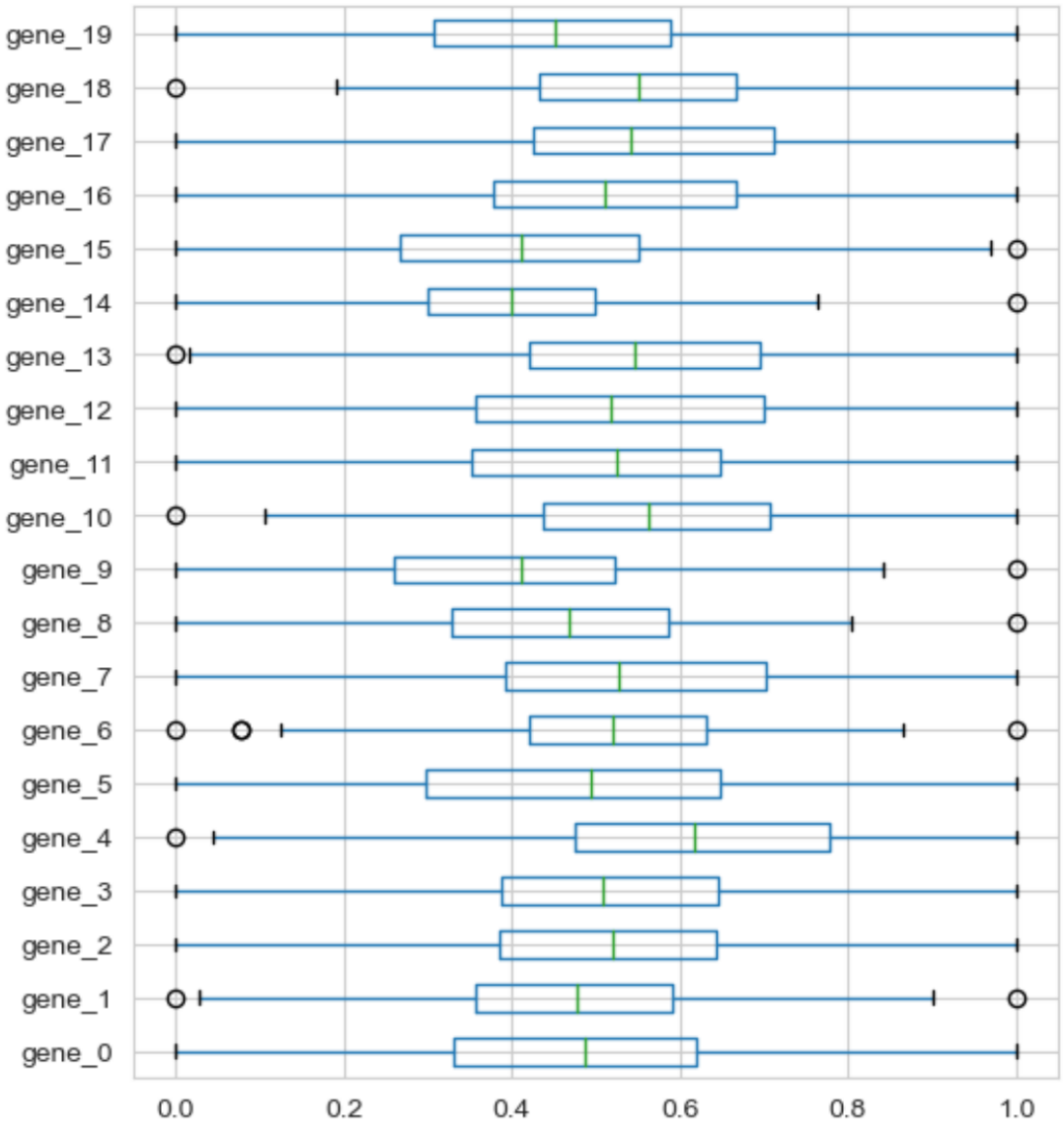


data normalization: standardization



$$x_{norm} = \frac{x - \mu}{\sigma}$$

data normalization: min-max scaling



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