



UNIVERSITÀ  
degli STUDI  
di CATANIA

DIPARTIMENTO DI  
MATEMATICA E INFORMATICA

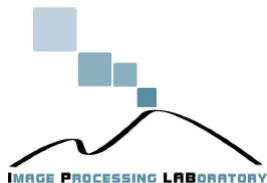
# Social Media Data Analysis 2023/2024

## *Nearest Neighbors*

Francesco Ragusa

[francesco.ragusa@unict.it](mailto:francesco.ragusa@unict.it)

<https://iplab.dmi.unict.it/ragusa/>



<https://iplab.dmi.unict.it/fpv/>



# Tasks: Classification



**Donald J. Trump**

Sponsored • Paid for by DONALD J. TRUMP FOR PRESIDENT, INC.

ID: 203601667365838

Everyone who enters before MIDNIGHT will have their name entered TWICE to win dinner with me. Please contribute NOW to get your name DOUBLE-ENTERED to win.



**VS.**



**Oppenheimer**

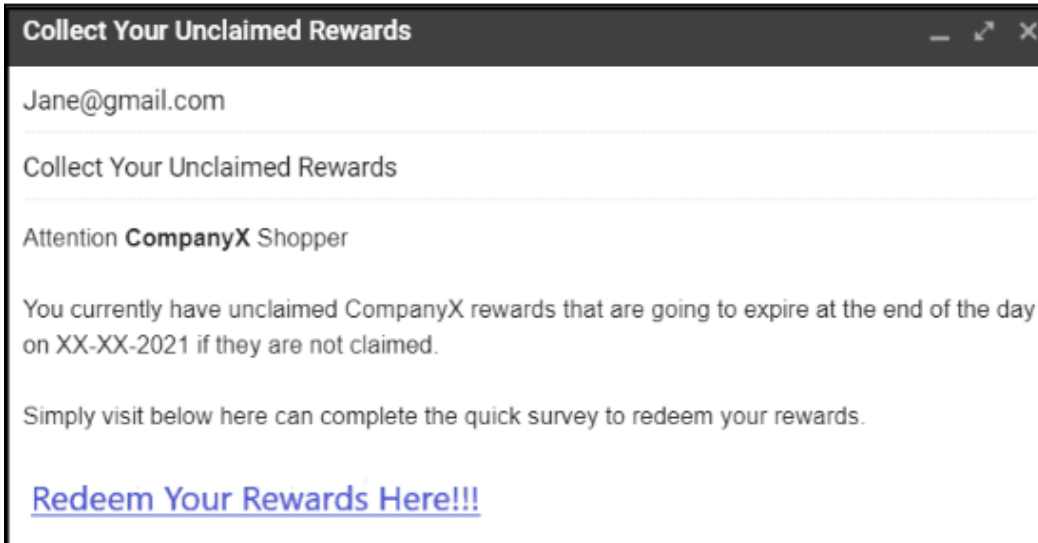
1 novembre alle ore 16:00 · 🌐

This Friday, [#Oppenheimer](#) is back in **IMAX** for one week. Experience it on the largest screen possible and get tickets now. [www.oppenheimertickets.com](http://www.oppenheimertickets.com)





# Tasks: Classification



VS.



**Dear John Smith:**

As part of our ongoing effort to provide better services and support, we would like to request your feedback via a short online survey. It should only take about 15 minutes to complete.

The survey is active for a limited time only, so please respond as soon as possible. This survey is hosted by an external company (VendorName), so the link below does not lead to our website. Your responses will be subject to [Amazon's Privacy Notice](#).

If you have any concerns about the authenticity of this email or to find out more about Amazon's survey program, please visit the [Amazon Customer Service help page](#).

In order to take the survey, please click here:  
<http://www1.vendordomain.com/study/rjyl/index.html?var=2739974>

Thank you very much for your time and effort!

Sincerely,

The Kindle Marketing Team

We hope you enjoyed receiving this message. However, if you'd rather not receive future e-mails of this sort from Amazon.com, please visit the opt-out link below.

<http://www.amazon.com/qp/qss/o/1pLvnyIpGCNHTOEZRxxRJ-2TzZBYS00ucDzsdMUeRtSI>

Please note that this message was sent to the following e-mail address: email-qa-blast+testing@amazon.com

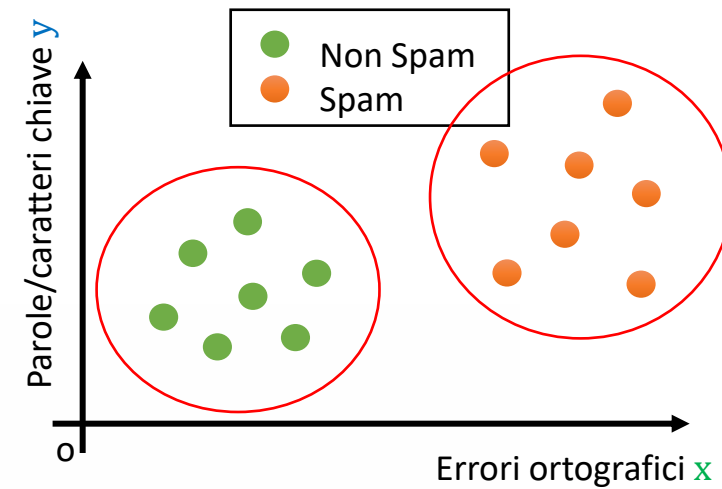
Copyright 2012 Amazon.com, Inc. or its affiliates. All Rights Reserved. Amazon, Amazon.com and the Amazon.com logo are registered trademarks of Amazon.com, Inc and its affiliates. Amazon.com, 410 Terry Avenue N, Seattle, WA 98109.

# Tasks: Classification



# Tasks: Classification

$$y^{(j)} = f(x^{(j)})$$



From: cheapsales@buystufffromme.com  
To: ang@cs.stanford.edu  
Subject: Buy now!

Deal of the week! Buy now!  
Rolex w4tchs - \$100  
Medicine (any kind) - \$50  
Also low cost M0rgages  
available.

Spam

$$y = 1$$

From: Alfred Ng  
To: ang@cs.stanford.edu  
Subject: Christmas dates?

Hey Andrew,  
Was talking to Mom about plans  
for Xmas. When do you get off  
work. Meet Dec 22?  
Alf

Non-spam

$$y = 0$$

# Tasks: Classification

## **Parametric and Non-parametric models**

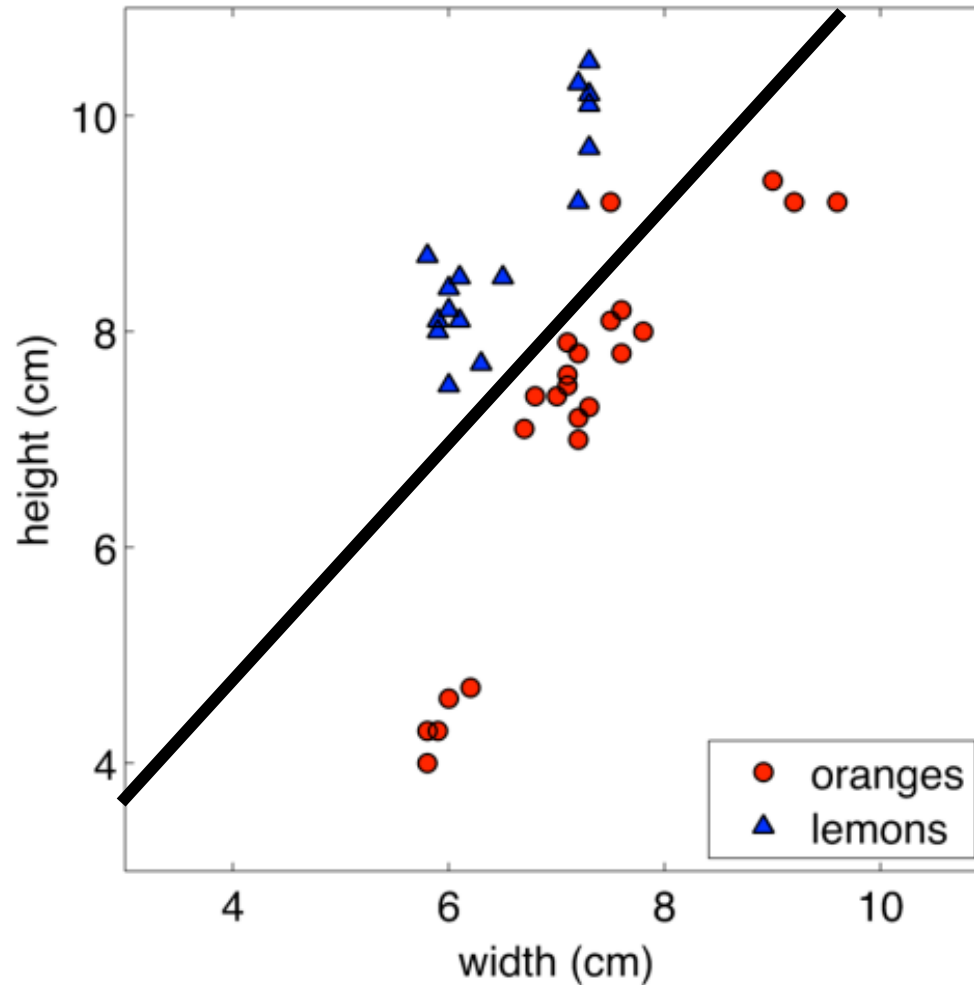
# Tasks: Classification

## Parametric and Non-parametric models

- Distance
- Non-linear decision boundaries

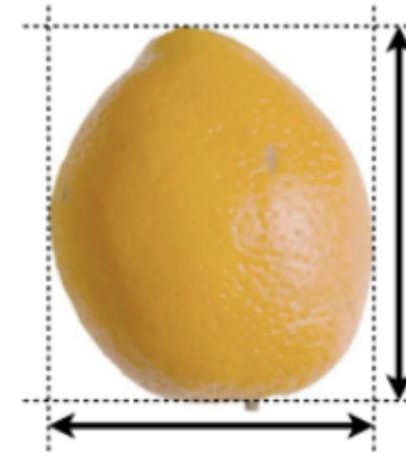


# Classification: Oranges and Lemons



Linear decision boundary:

$$w_0 + w_1x_1 + w_2x_2$$





# Classification: non-parametric models

**They work for classification or regression problems**

**Learning amounts to simply storing training data**

**Test instances classified using similar training instances**

# Data representation: notation

Examples:  $(\boldsymbol{x}^{(j)}, y^{(j)})_j$

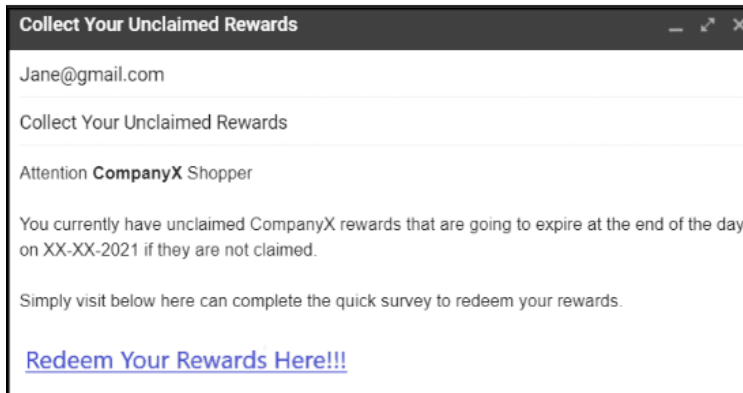
Data points:  $\boldsymbol{x} \in \mathfrak{R}^n$

Features:  $\boldsymbol{x} = (x_1, \dots, x_n)$

Labels:  $y \in \{0, \dots, m - 1\}$

$$TR = \{(\boldsymbol{x}^{(j)}, y^{(j)})\}_j; TE = \{(\boldsymbol{x}^{(j)}, y^{(j)})\}_j; VA = \{(\boldsymbol{x}^{(j)}, y^{(j)})\}_j$$

# Data representation: notation



Data points:  $\mathbf{x} \in \mathbb{R}^{50000}$

Features:  $\mathbf{x} = (x_1, \dots, x_{50000})$



Dear John Smith:

As part of our ongoing effort to provide better services and support, we would like to request your feedback via a short online survey. It should only take about 15 minutes to complete.

The survey is active for a limited time only, so please respond as soon as possible. This survey is hosted by an external company (VendorName), so the link below does not lead to our website. Your responses will be subject to [Amazon's Privacy Notice](#).

If you have any concerns about the authenticity of this email or to find out more about Amazon's survey program, please visit the [Amazon Customer Service help page](#).

In order to take the survey, please click here:  
<http://www1.vendordomain.com/study/rjyl/index.html?var=2739974>

Thank you very much for your time and effort!

Sincerely,

The Kindle Marketing Team

We hope you enjoyed receiving this message. However, if you'd rather not receive future e-mails of this sort from Amazon.com, please visit the opt-out link below.

<http://www.amazon.com/gp/qss/q/1pLvnyIpGCNHTOEZRXXRJ-2TzZBYS00ucDzsdMUeRtSI>

Please note that this message was sent to the following e-mail address: email-qa-blast+testing@amazon.com

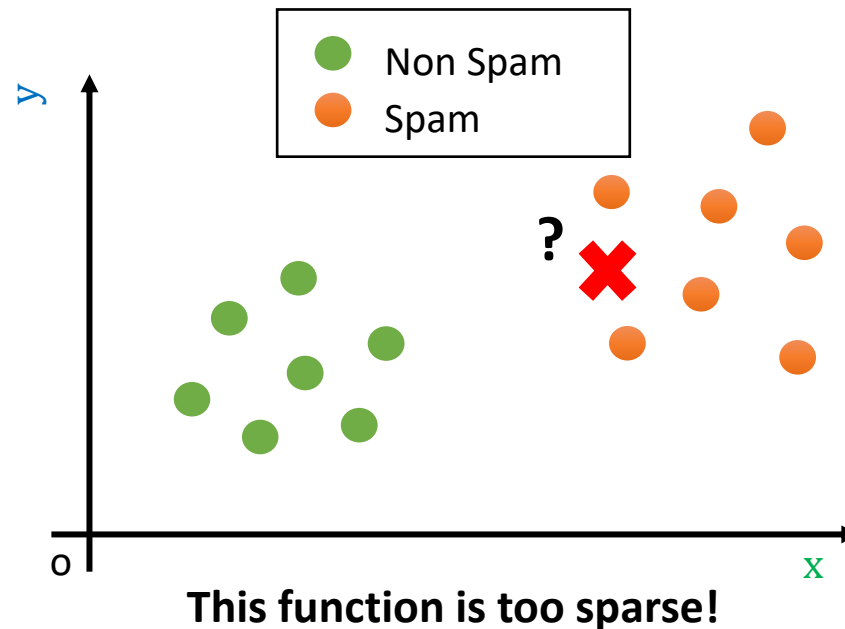
Copyright 2012 Amazon.com, Inc. or its affiliates. All Rights Reserved. Amazon, Amazon.com and the Amazon.com logo are registered trademarks of Amazon.com, Inc and its affiliates. Amazon.com, 410 Terry Avenue N, Seattle, WA 98109.

Labels:  $y \in \{0, 1\}$

# Classification By Retrieval: Nearest Neighbor

**Idea:** The value of the target function for a new query is estimated from the known value(s) of the nearest training example(s)

$$f(\mathbf{x}) = y \text{ s.t. } (\mathbf{x}, y) \in TR$$

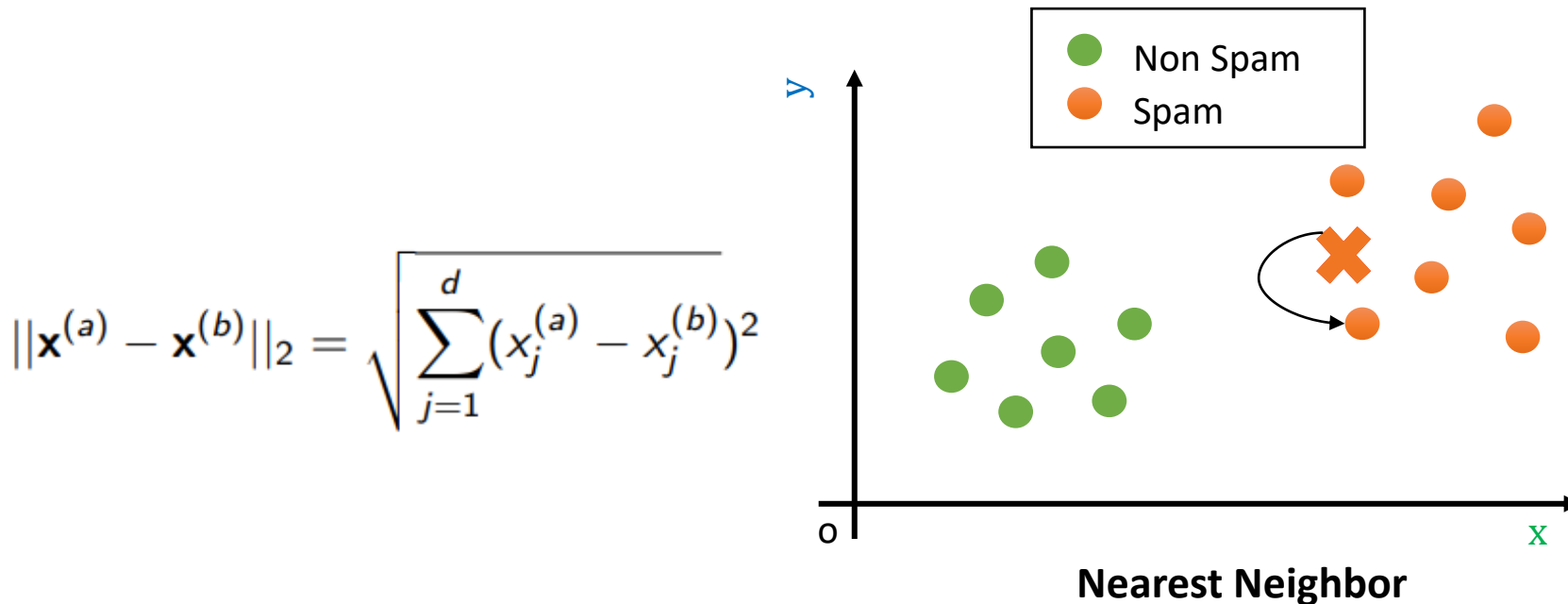




# Classification By Retrieval: Nearest Neighbor

**Idea:** The value of the target function for a new query is estimated from the known value(s) of the nearest training example(s)

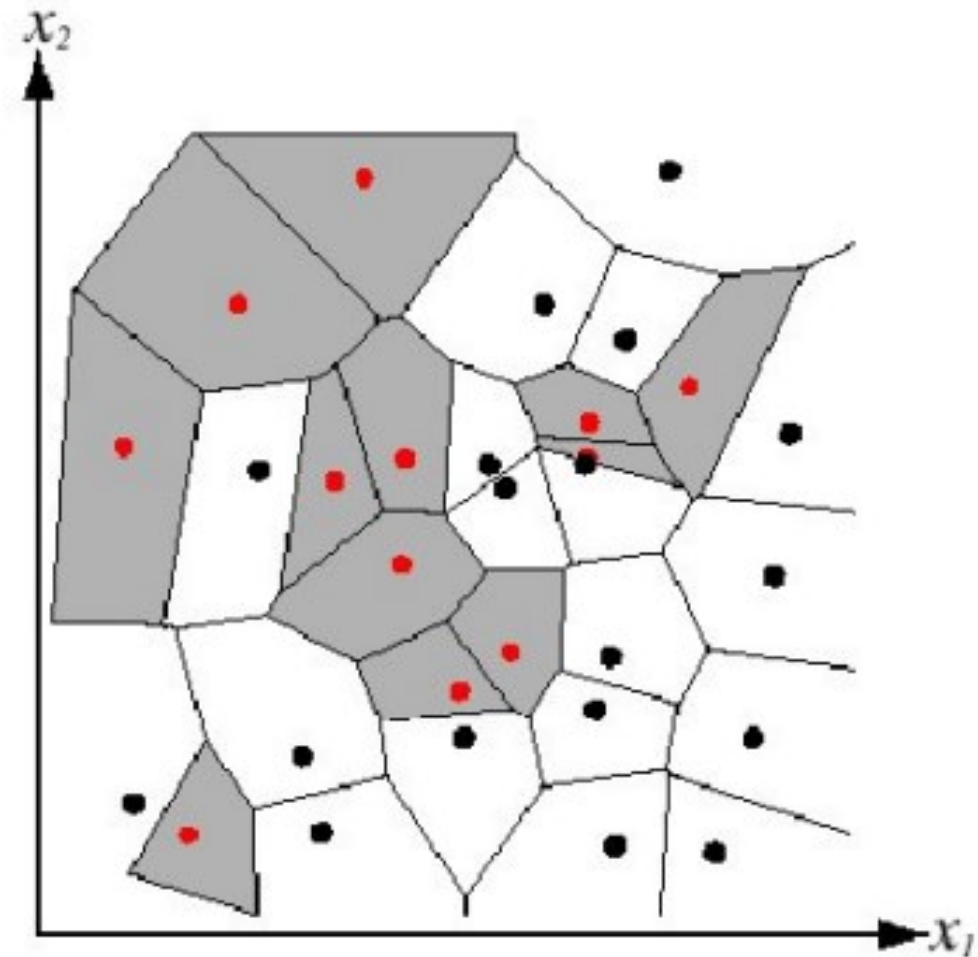
$$f(\bar{\mathbf{x}}) = \arg_y \min\{d(\bar{\mathbf{x}}, \mathbf{x}) | (\mathbf{x}, y) \in TR\}$$



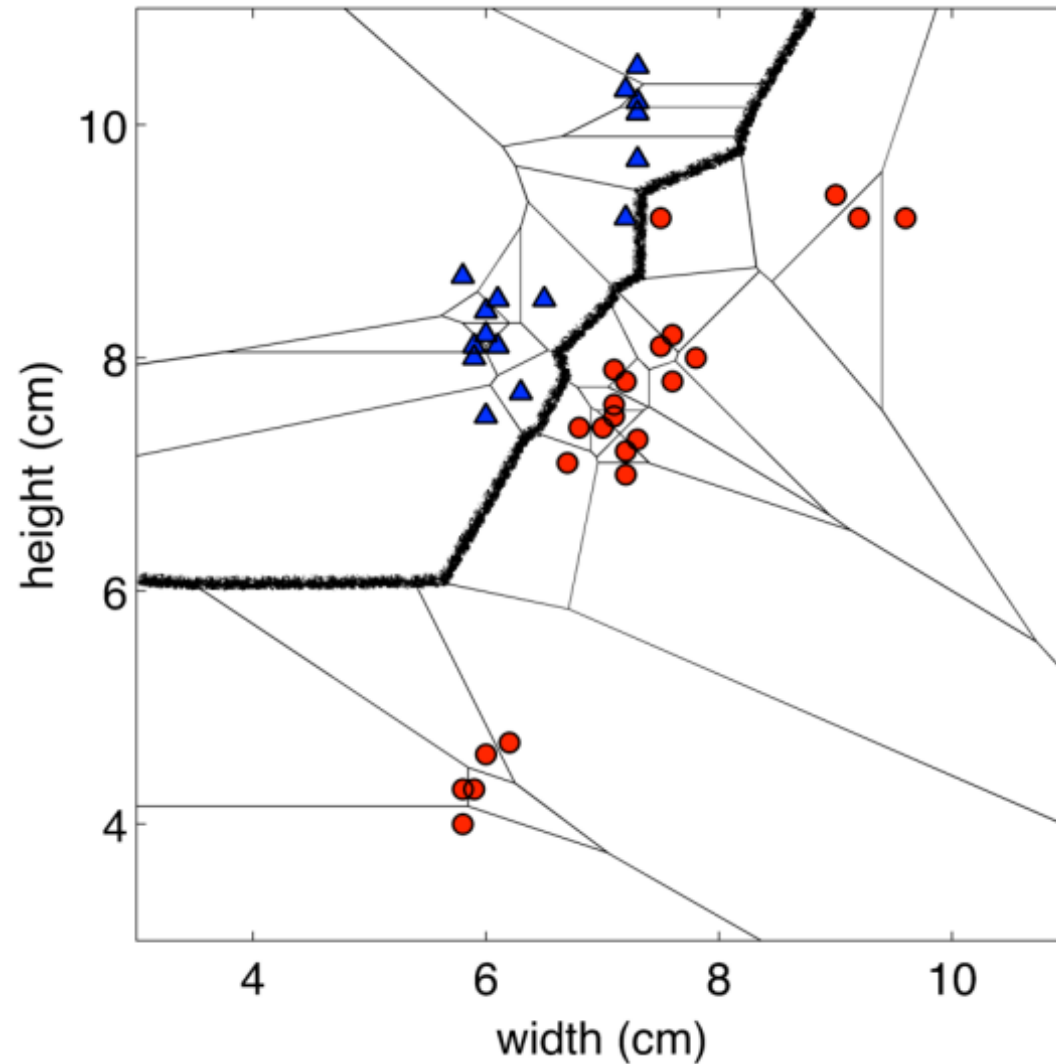
# Nearest Neighbor: Decision Boundaries

It does not explicitly compute decision boundaries, but these can be inferred

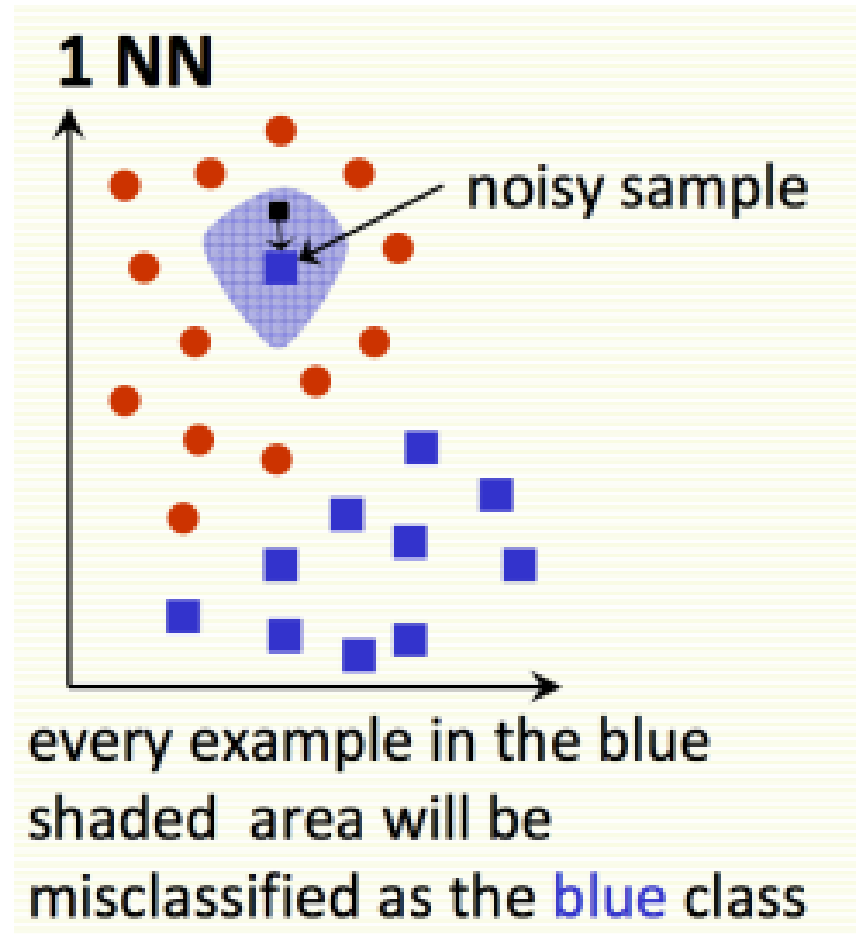
Voronoi Diagram visualization:



# Nearest Neighbors: Decision Boundaries



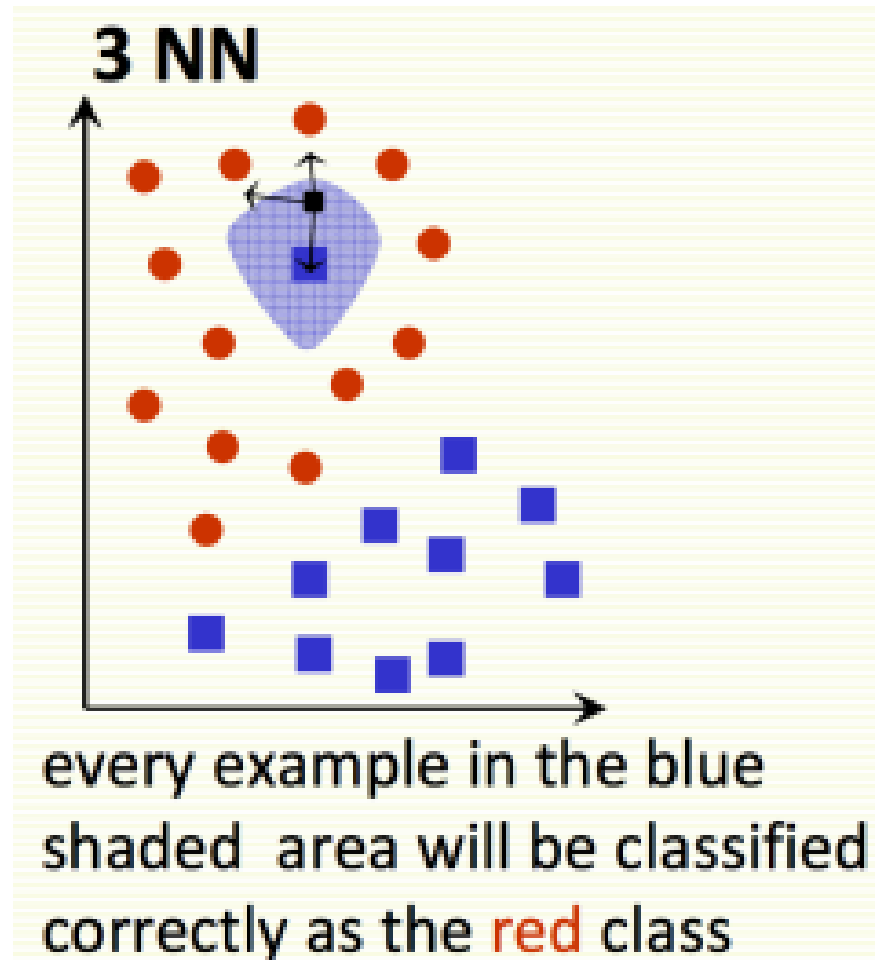
# Nearest Neighbors



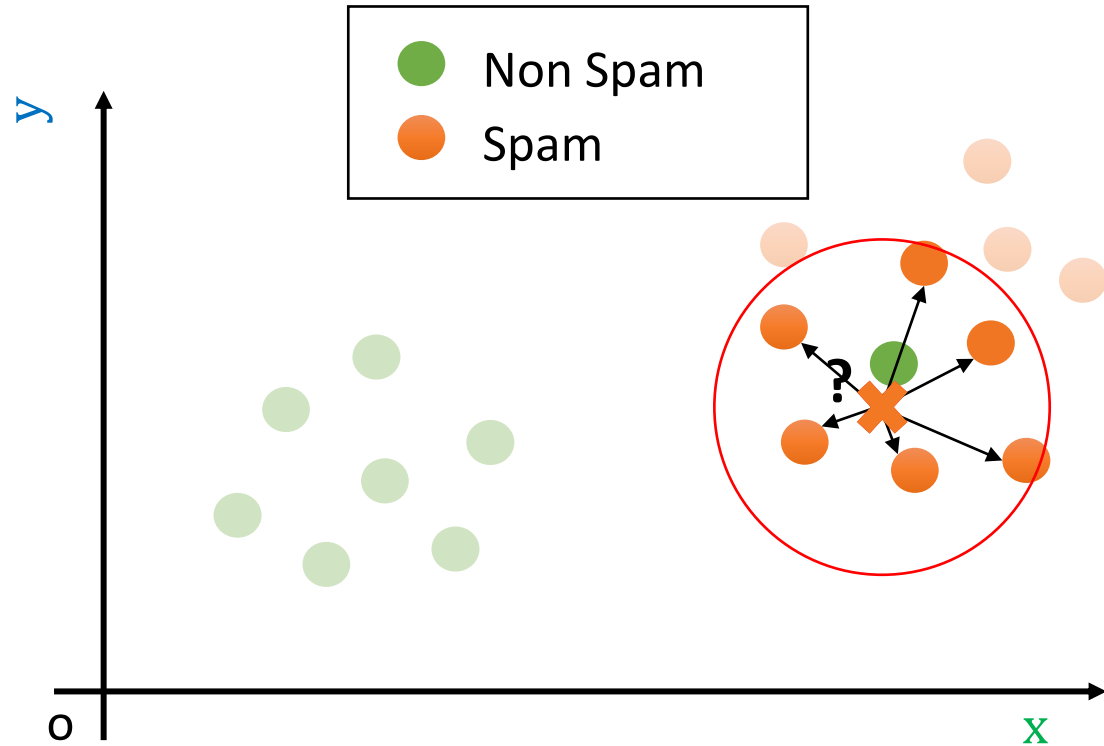
Nearest neighbors is sensitive to mis-labeled data (“class noise”). Solution?



# K-Nearest Neighbors



# K-Nearest Neighbors



$$N(\bar{\mathbf{x}}; TR, \epsilon) = \{(\mathbf{x}, y) \in TR \mid d(\bar{\mathbf{x}}, \mathbf{x}) \leq \epsilon\}$$

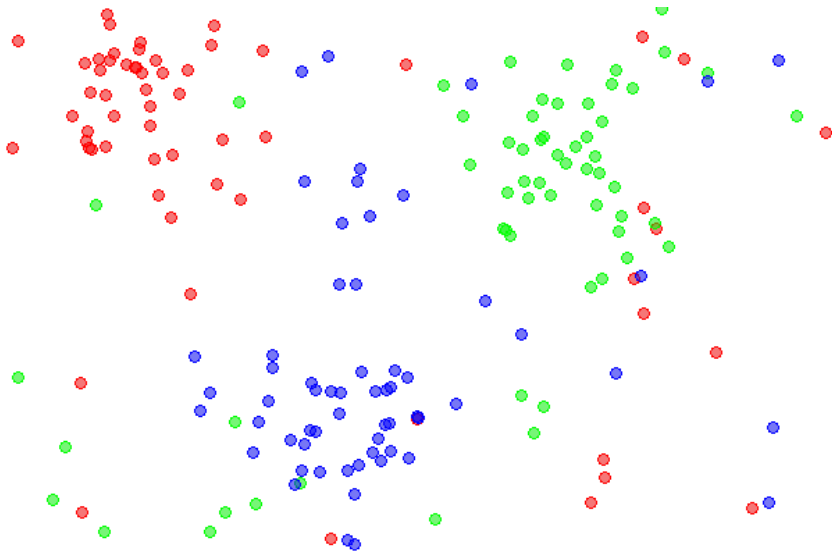
$$\epsilon_K(\bar{\mathbf{x}}; TR, K) = \max\{\epsilon \text{ s. t. } |\{d(\bar{\mathbf{x}}, \mathbf{x}) \leq \epsilon\}| \leq K, \forall \epsilon \in \mathbb{R}\}$$

$$N(\bar{\mathbf{x}}; TR, K) = \{(\mathbf{x}, y) \in TR \mid d(\bar{\mathbf{x}}, \mathbf{x}) \leq \epsilon_K(\bar{\mathbf{x}}; TR, K)\}$$

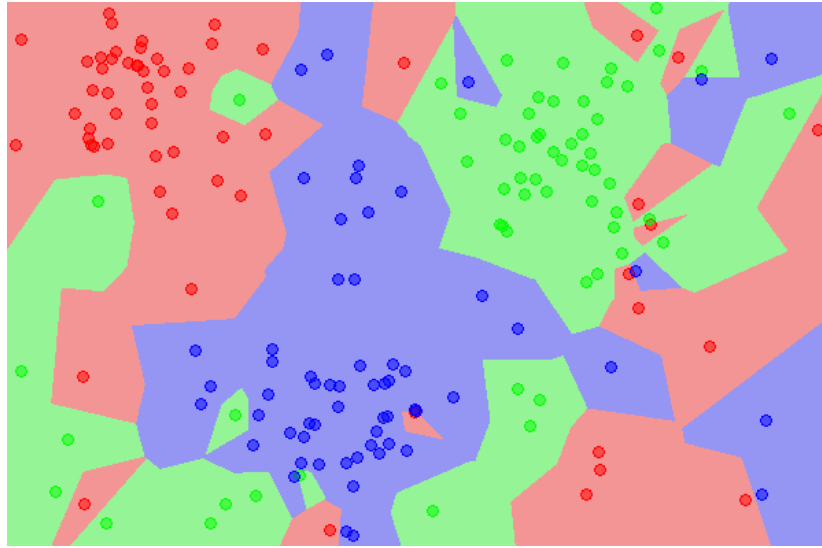
$$f(\bar{\mathbf{x}}) = \text{mode}\{y \mid (\mathbf{x}, y) \in N(\bar{\mathbf{x}}; TR, K)\}$$

# Classification Map/Decision Boundary

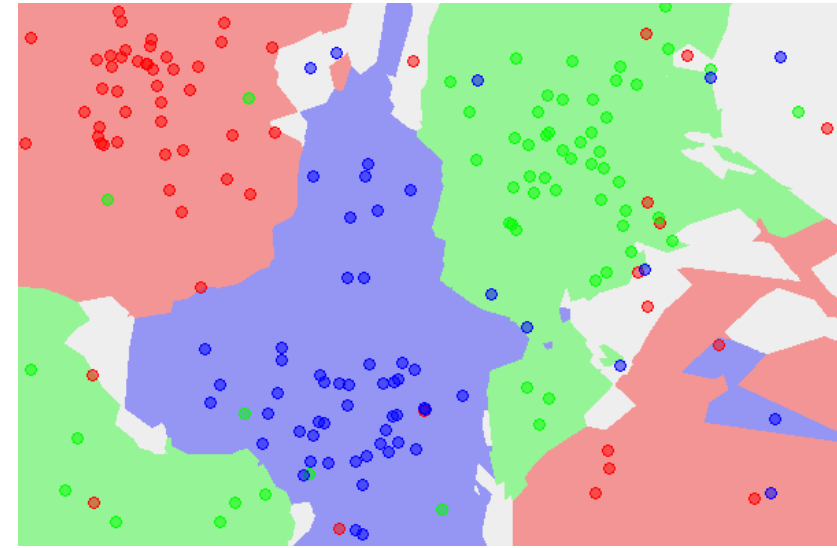
**Input Data**



**1-NN Classifier**



**5-NN Classifier**



# NN is Powerful!

Image Collection





# Classification: Evaluating Performance

Ground truth labels:  $Y = \{y^{(i)}\}_i$

Predicted labels  $\hat{Y} = \{\hat{y}^{(i)}\}_i = \{f(\mathbf{x}^{(i)})\}_i$

Performance measure:  $P(Y, \hat{Y}) \rightarrow \mathbb{R}$

# Accuracy

$$Accuracy(Y, \hat{Y}) = \frac{|\{i \mid y^{(i)} = \hat{y}^{(i)}\}|}{|Y|}$$

**not good for imbalanced datasets (example)**

# Confusion Matrix

Type 1: False Positives (FP)

Type 2: False Negatives (FN)

True Positives (TP)

True Negatives (TN)

CONFUSION MATRIX		PREDICTED LABELS	
		POSITIVE	NEGATIVE
TRUE LABELS	POSITIVE	TP	FN
	NEGATIVE	FP	TN

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

more informative with imbalanced datasets

# Precision and Recall

CONFUSION MATRIX		PREDICTED LABELS	
		POSITIVE	NEGATIVE
TRUE LABELS	POSITIVE	TP	FN
	NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP}$$

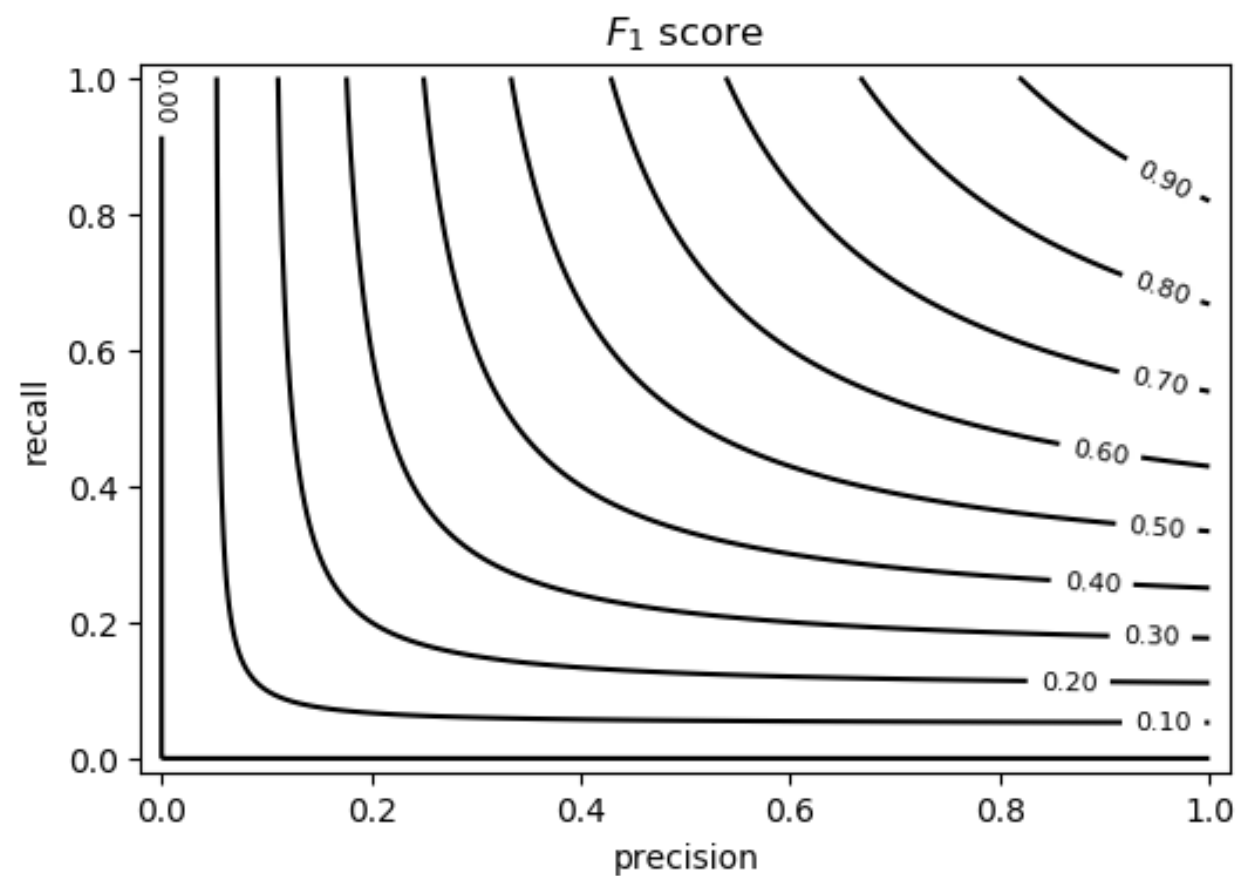
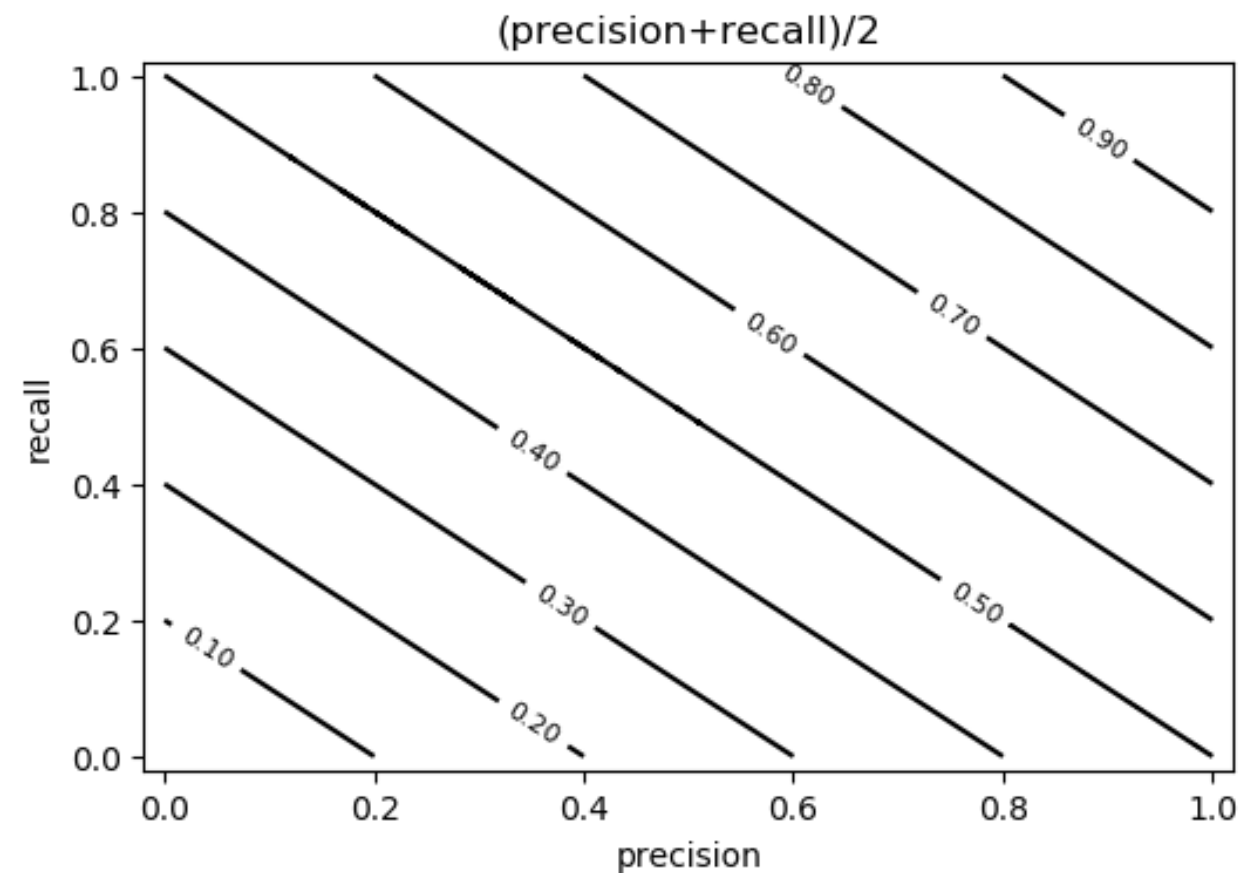
$$Recall = \frac{TP}{TP + FN}$$

summarizes the confusion matrix



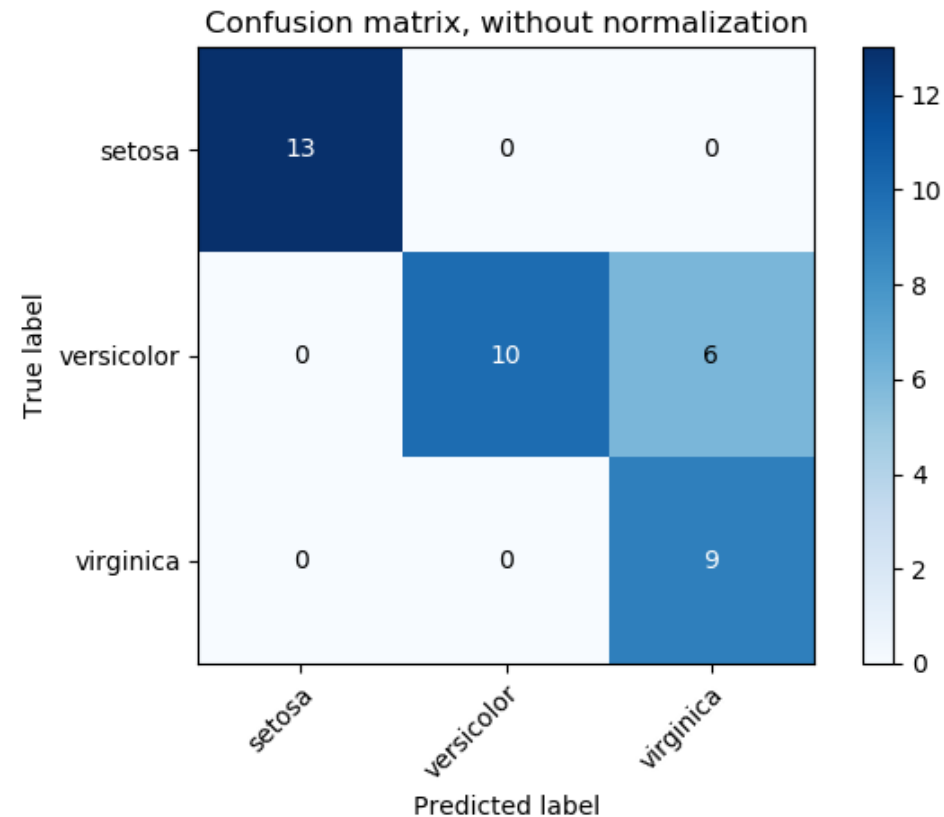
# $F_1$ Score

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



# Confusion Matrix for Multi-Class Classification

$M_{ij}$ : number of elements of class  $i$  classified as belonging to class  $j$



# References/Optional Readings

- Nearest Neighbor: Section 2.5.2 of [1]
- Evaluation Measures for Classification:  
[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)

[1] Bishop, Christopher M. *Pattern recognition and machine learning*.  
springer, 2006. <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>