UNIVERSITÀ degli STUDI di CATANIA
DIPARTIMENTO DI MATEMATICA E INFORMATICA

Social Media Data Analysis 2023/2024

Nearest Neighbors

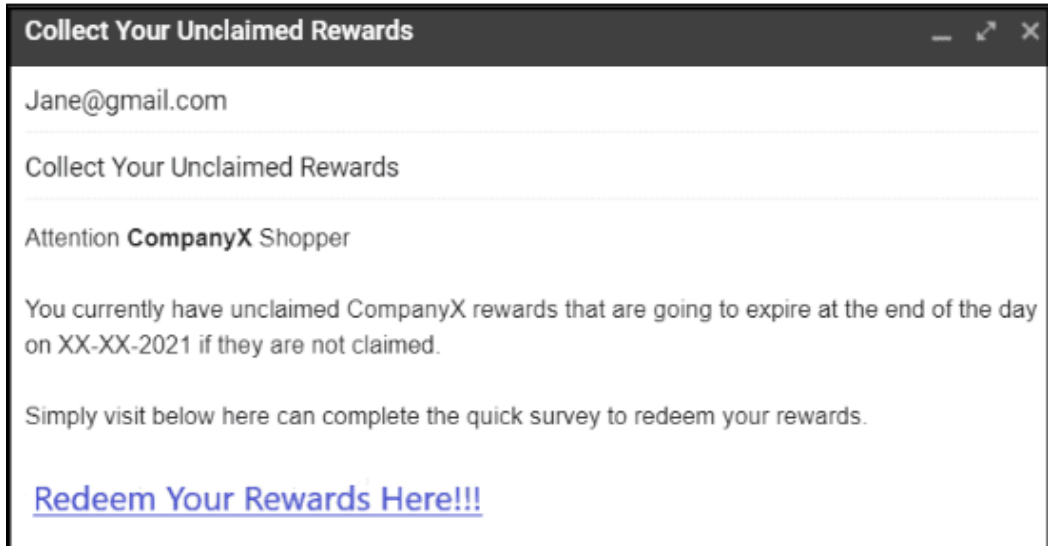
Francesco Ragusa
francesco.ragusa@unict.it
https://iplab.dmi.unict.it/ragusa/

IMAGE PROCESSING LABORATORY
https://iplab.dmi.unict.it/fpv/

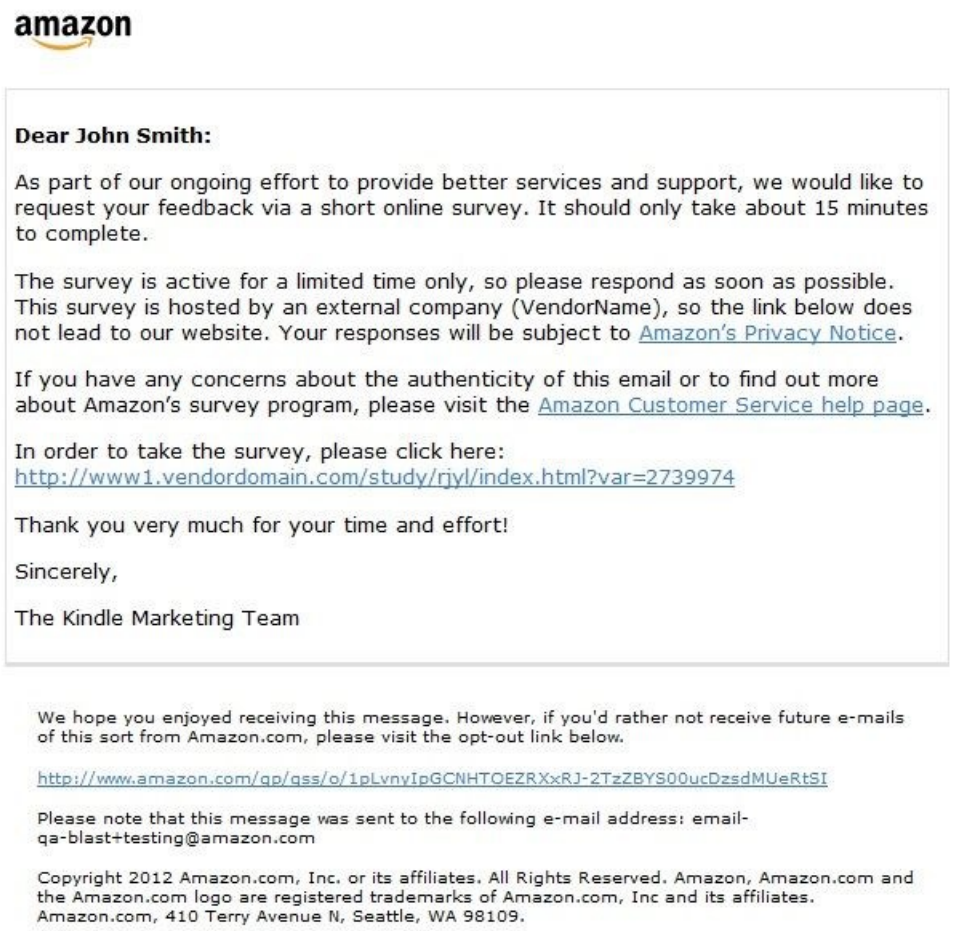
Tasks: Classification

vs.

Tasks: Classification



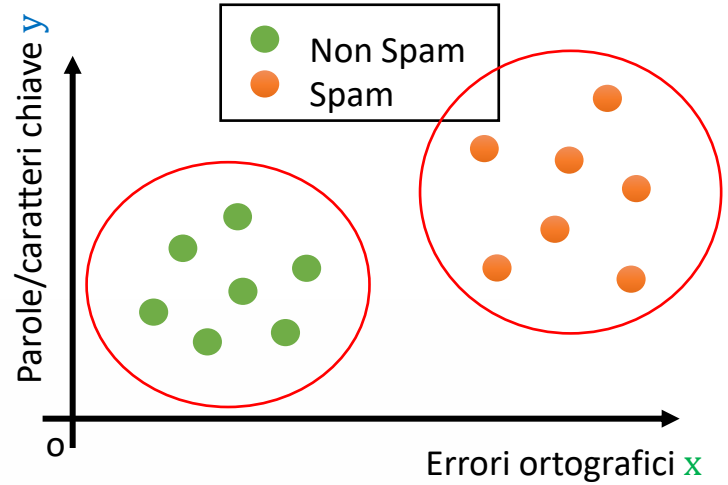
vs.



Tasks: Classification

Tasks: Classification

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



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

Deal of the week! Buy now!
Rolex w4tchs - $100
Med1cine (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

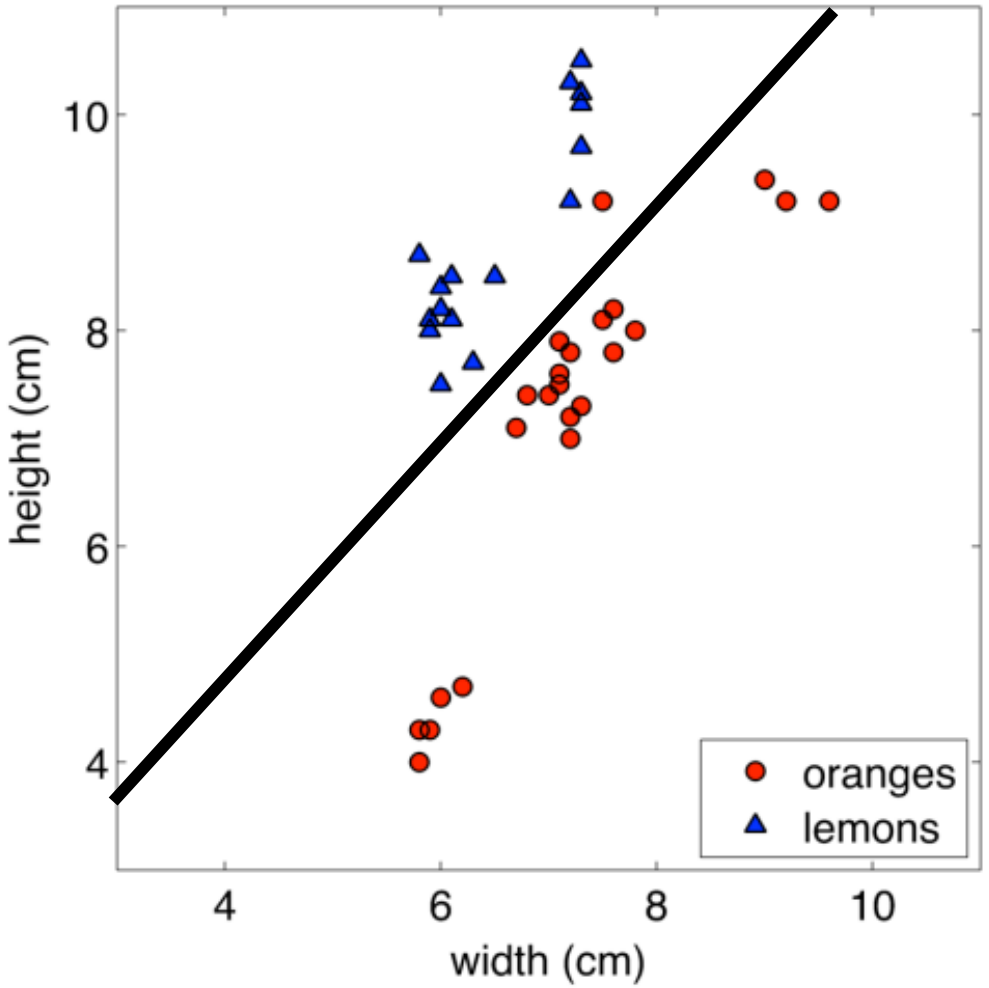
Parametric and Non-parametric models

Tasks: Classification

Parametric and Non-parametric models

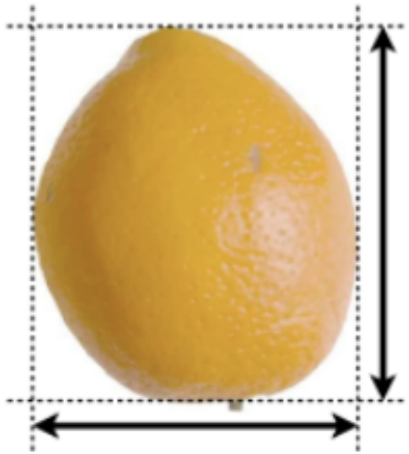
- **Distance**
- **Non-linear decision boundaries**

Classification: Oranges and Lemons



Linear decision boundary:

$$w_0 + w_1 x_1 + w_2 x_2$$



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

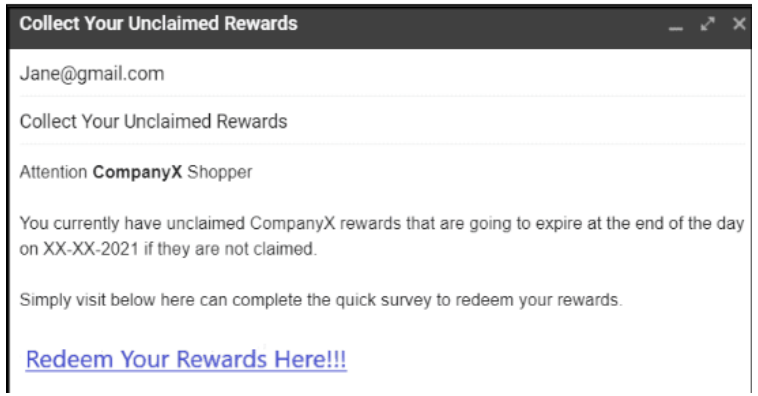
Data points: $\boldsymbol{x} \in \mathfrak{R}^n$

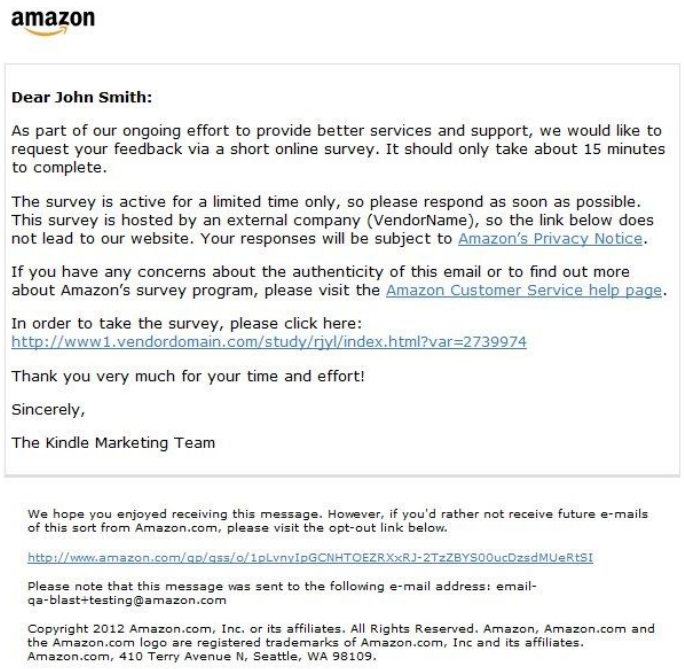
Features: $\boldsymbol{x} = (x_1, \ldots, x_n)$

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

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

Data representation: notation





Data points: $\boldsymbol{x} \in \Re^{50000}$

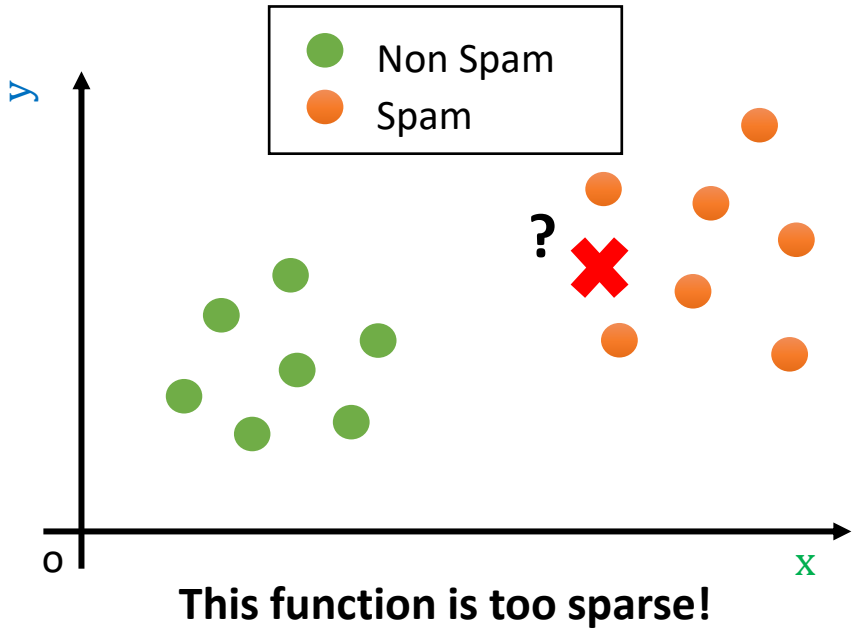
Features: $\boldsymbol{x} = (x_1, \ldots, x_{50000})$

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)

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



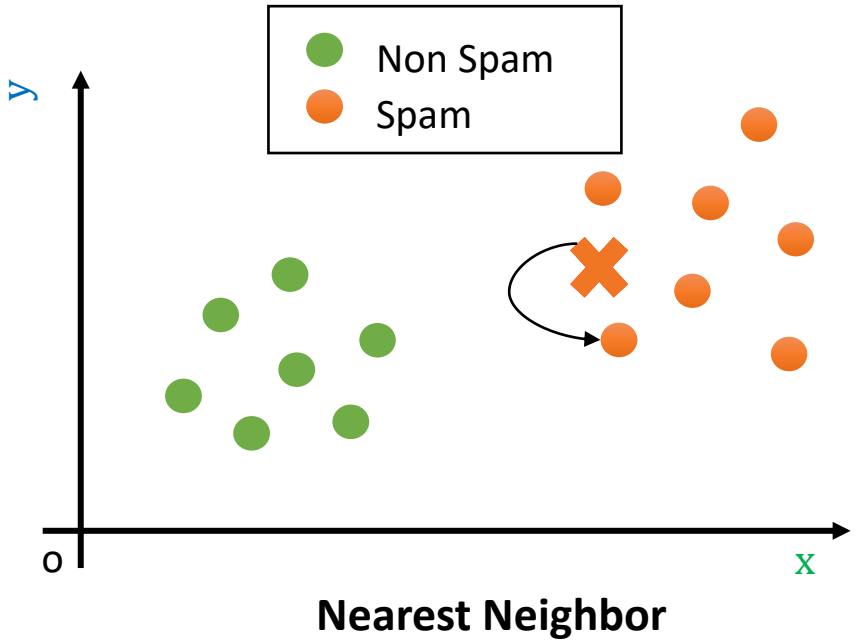
This function is too sparse!

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)

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

$$||\mathbf{x}^{(a)} - \mathbf{x}^{(b)}||_2 = \sqrt{\sum_{j=1}^{d} (x_j^{(a)} - x_j^{(b)})^2}$$

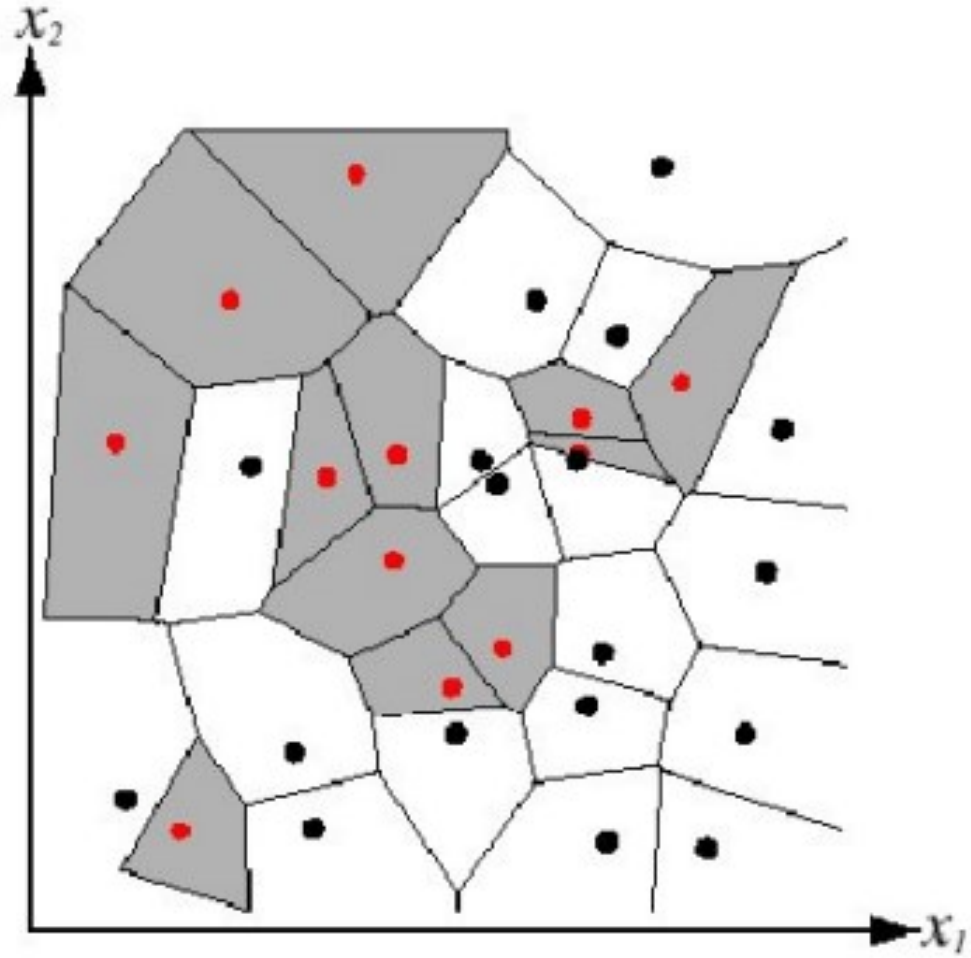


Nearest Neighbor

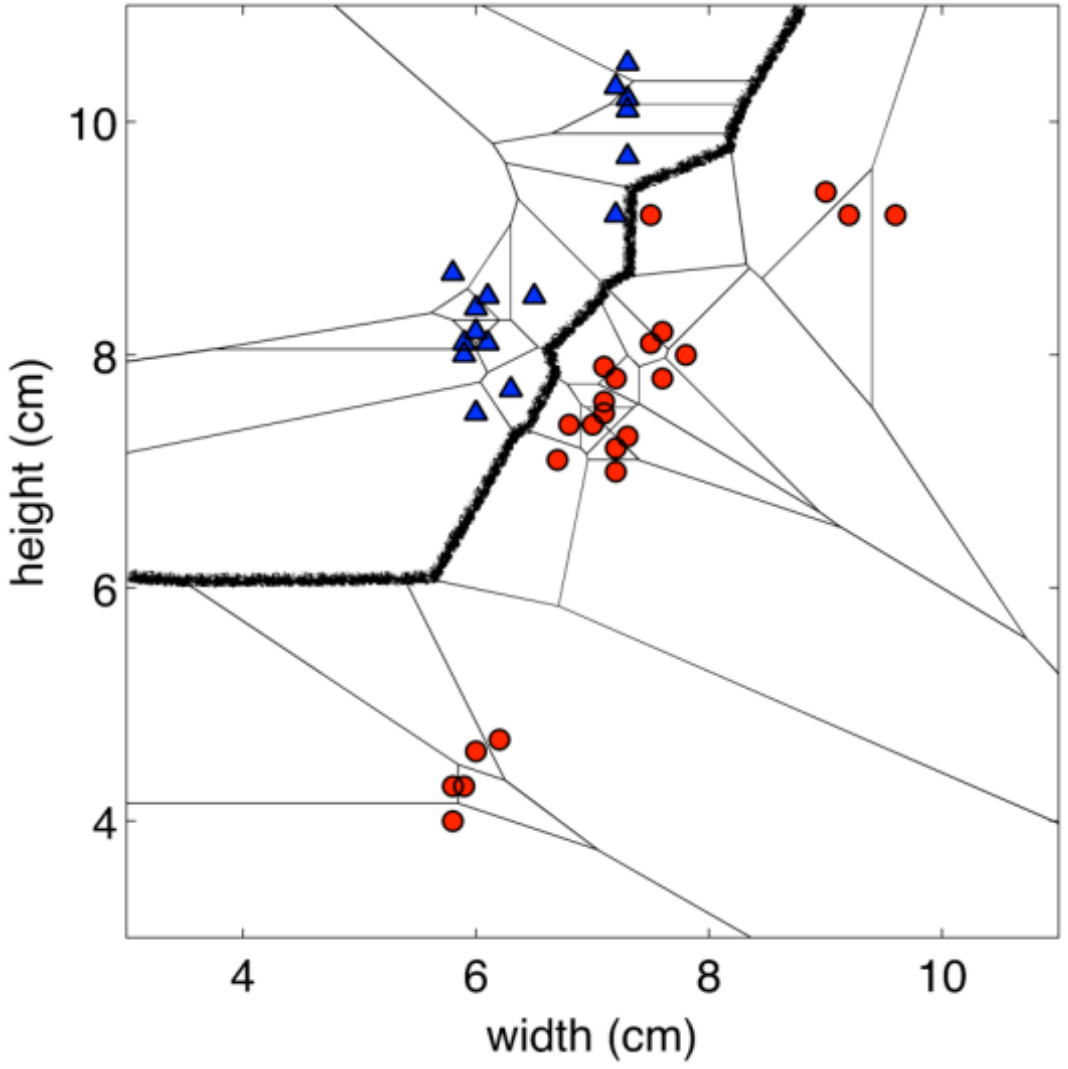
Nearest Neighbor: Decision Boundaries

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

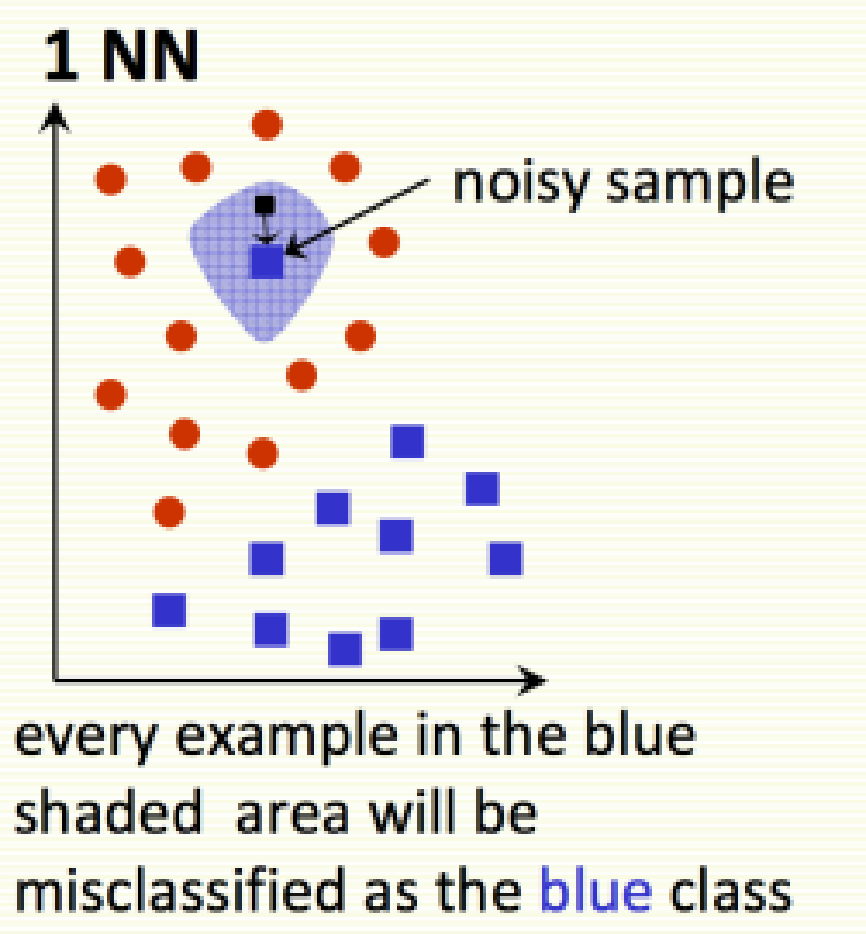
Voronoi Diagram visualization:



Nearest Neighbors: Decision Boundaries

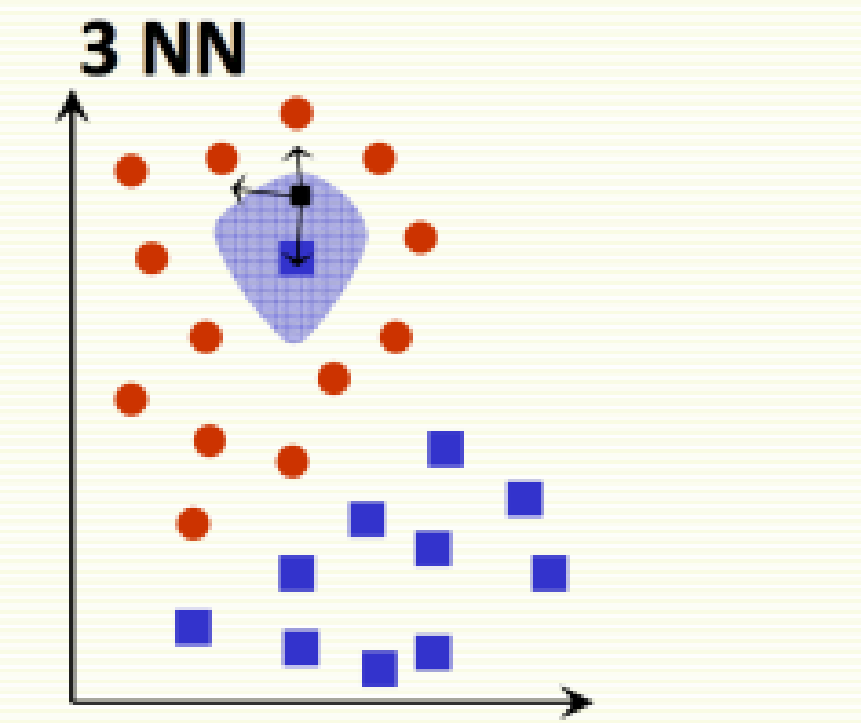


Nearest Neighbors

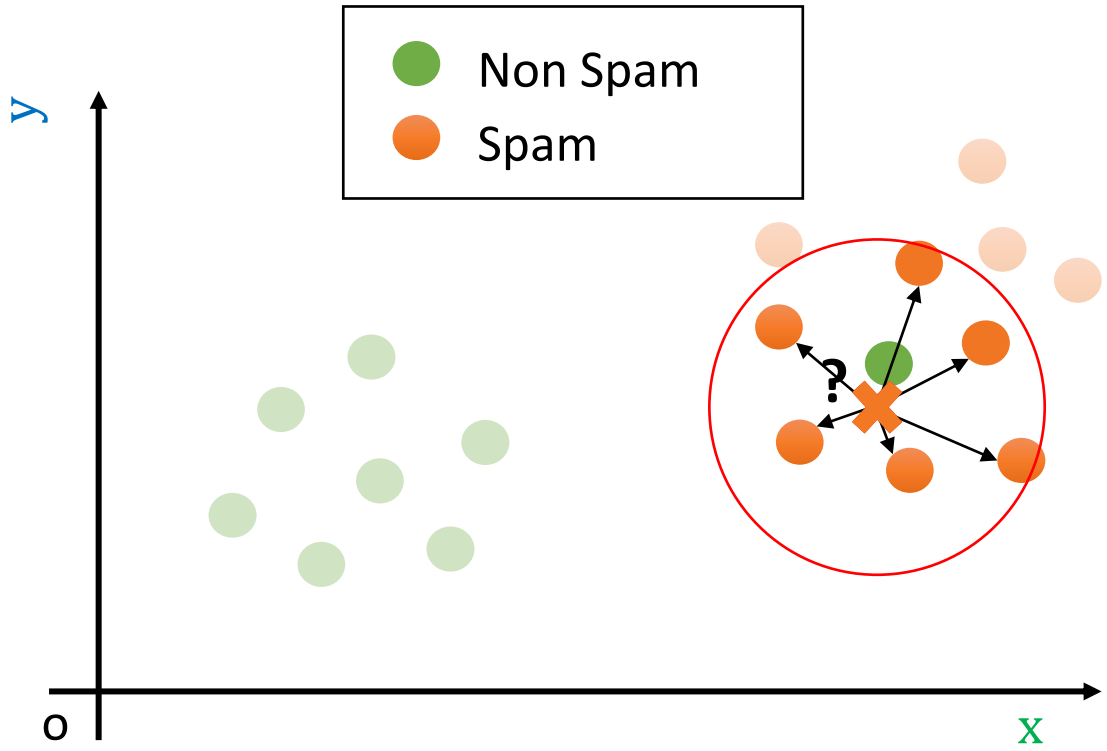


Nearest neighbors is sensitive to mis-labeled data (“class noise”). Solution?

K-Nearest Neighbors



K-Nearest Neighbors



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

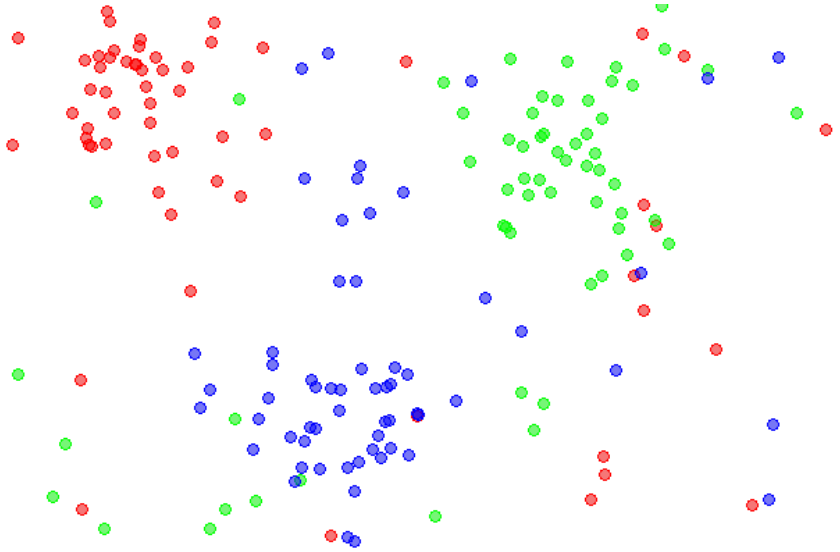
$$\epsilon_K(\overline{\boldsymbol{x}}; TR, K) = max\{\varepsilon \; s.t. \; |\{d(\overline{\boldsymbol{x}}, \boldsymbol{x}) \leq \varepsilon\}| \leq K, \forall \varepsilon \in \Re\}$$

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

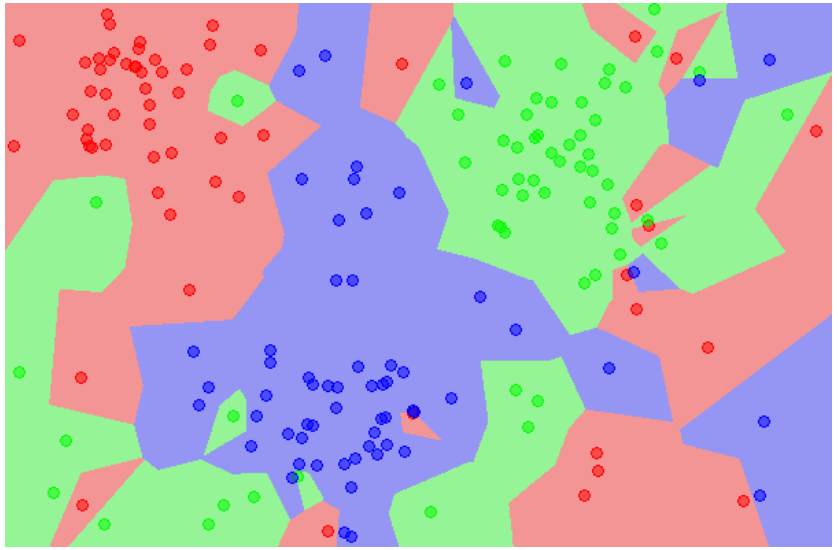
$$f(\overline{\boldsymbol{x}}) = mode\{y|(\boldsymbol{x}, y) \in N(\overline{\boldsymbol{x}}; TR, K)\}$$

Classification Map/Decision Boundary

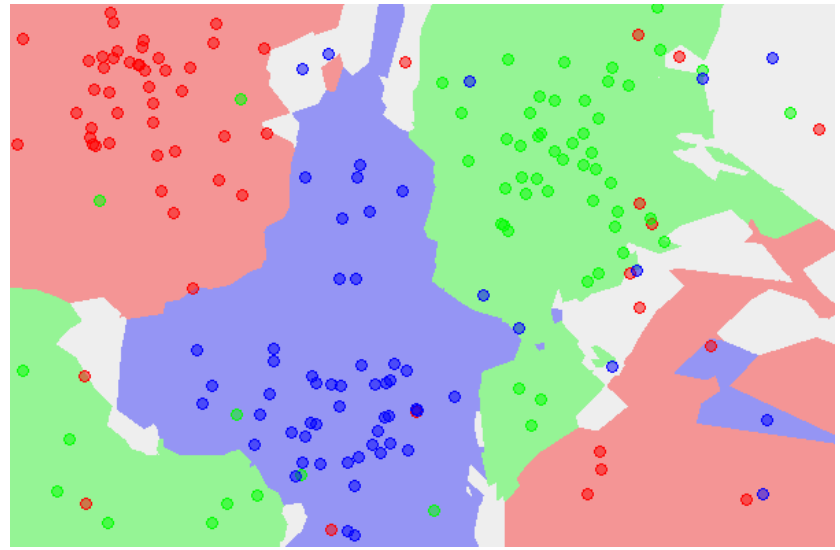
Input Data



1-NN Classifier

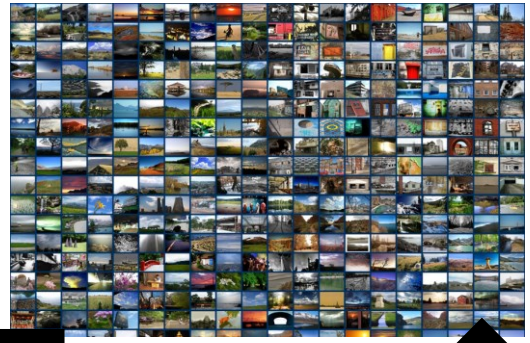


5-NN Classifier



NN is Powerful!

Image Collection



Classification: Evaluating Performance

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

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

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

Accuracy

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

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

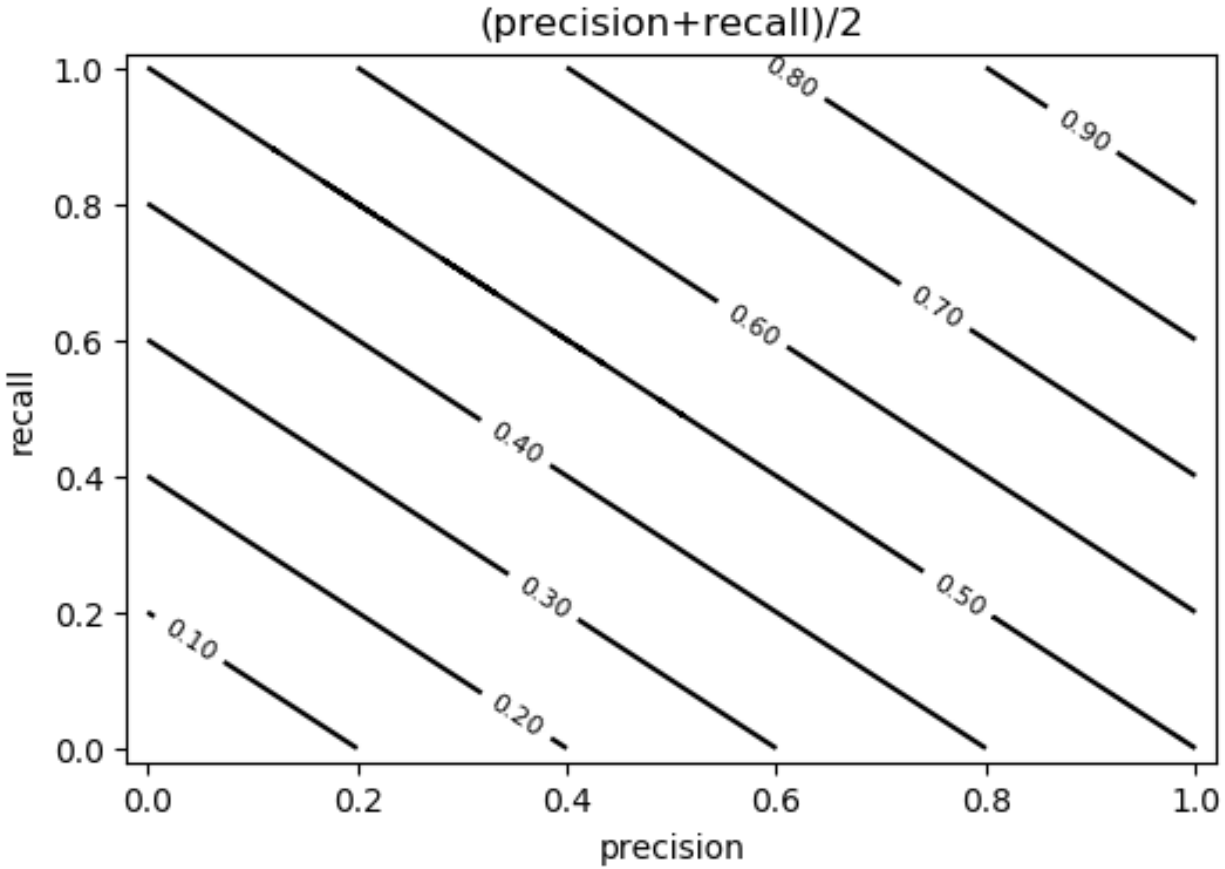
$$Precision = \frac{TP}{TP + FP}$$

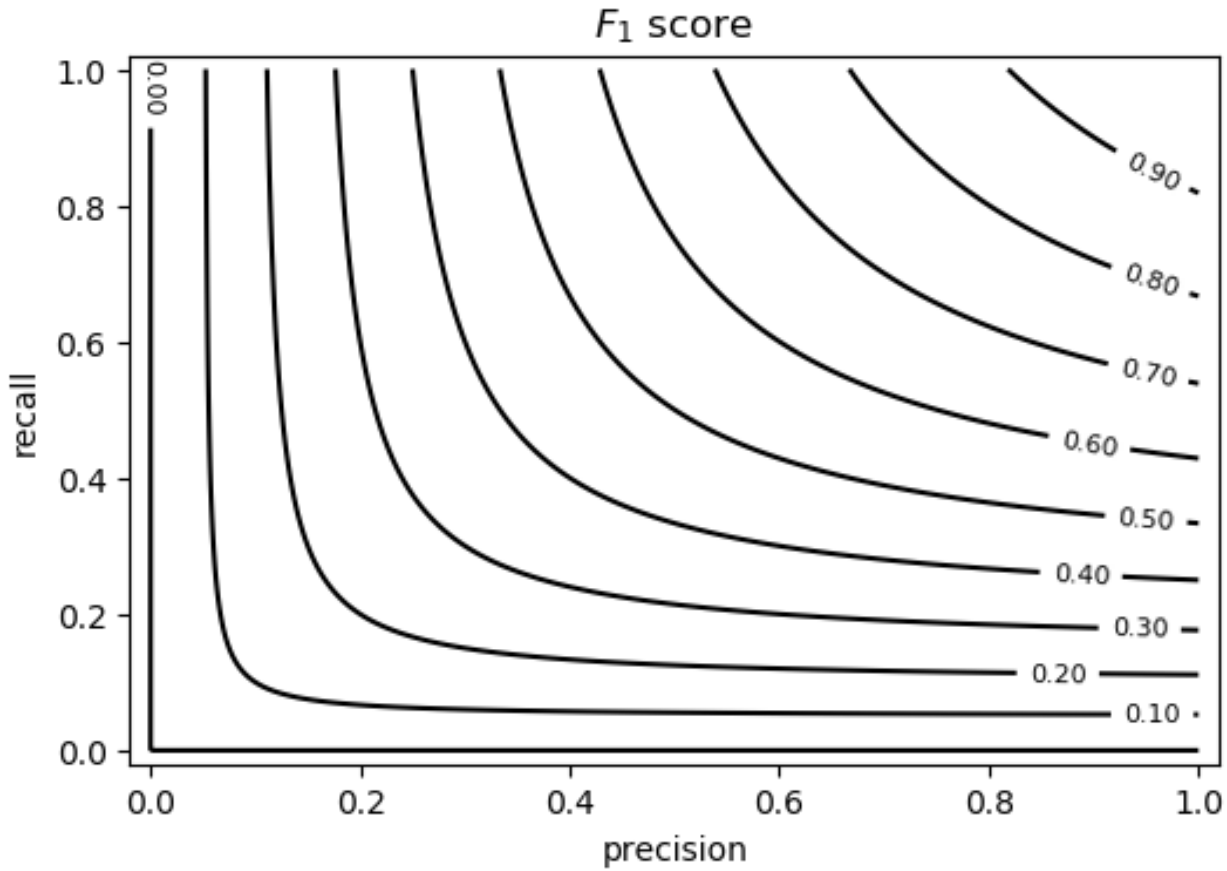
$$Recall = \frac{TP}{TP + FN}$$

summarizes the confusion matrix

F_1 Score

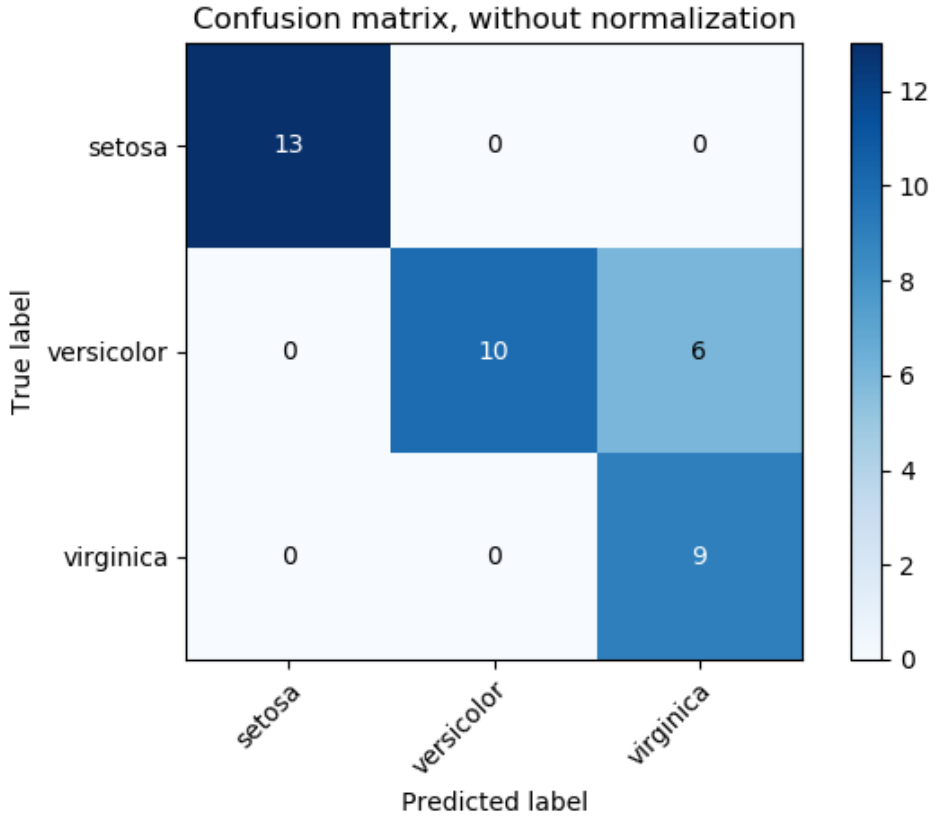
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$





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 Classifcation: 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