A brief introduction to the k-nearest neighbors classifier

Course: English for Academic Purposes

Student: Eduardo Henrique Basilio de Carvalho

Universidade Federal de Minas Gerais, May 13, 2025

Table of Contents I

Table of Contents II

Problem introduction

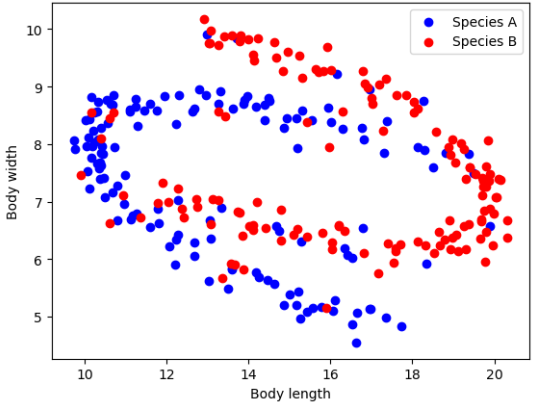
- Two species
- Count sightings of each
- Take some measurements

Training data

- Species are distinguishable by fur color
- A camera measures body length and width

Day measurements

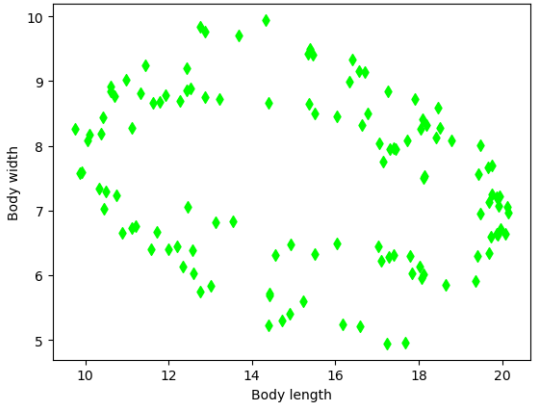
Figure: Day sightings plot



UFMG

Night measurements

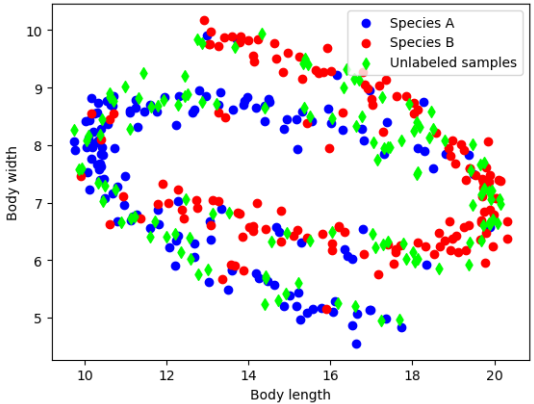
Figure: Night sightings plot



UFMG

Data superposition

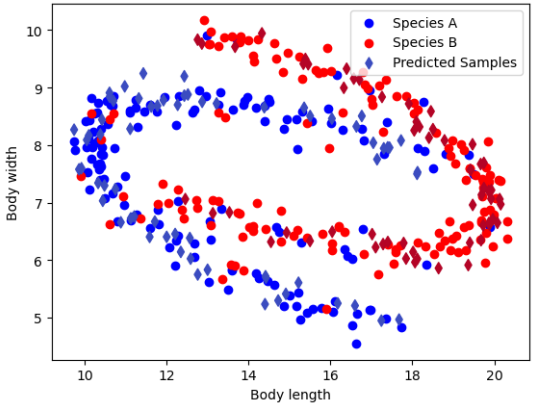
Figure: Superimposed sightings plot



UFMG

Visual prediction

Figure: Visually predicted samples



Higher dimensional data I

Table: Four-dimensional train samples

Sample	Feature 0	Feature 1	Feature 2	Feature 3	Label
0	-1.12	0.43	-1.5	0.55	1
1	1.93	-1.71	-0.75	-1.15	0
2	1.7	1.63	1.44	-0.42	1
3	-2.45	0.64	-0.48	0.17	1
4	1.14	-0.56	0.46	-1.04	1
5	-1.29	-1.58	-0.04	-2.11	0
6	-1.56	-1.13	-1.08	0.7	0
7	2.02	-0.14	-1.25	-1.96	1
8	1.37	0.01	-3.05	1.66	0

Higher dimensional data II

Table: Four-dimensional test sample

Feature 0	Feature 1	Feature 2	Feature 3	Label
-0.72	-0.41	1.21	-2.49	?

Distance

Table: Four-dimensional train samples with distances

Sample	Feature 0	Feature 1	Feature 2	Feature 3	Label	Distance
0	-1.12	0.43	-1.5	0.55	1	1.62
1	1.93	-1.71	-0.75	-1.15	0	4.47
2	1.7	1.63	1.44	-0.42	1	5.24
3	-2.45	0.64	-0.48	0.17	1	4.73
4	1.14	-0.56	0.46	-1.04	1	6.04
5	-1.29	-1.58	-0.04	-2.11	0	6.87
6	-1.56	-1.13	-1.08	0.7	0	7.34
7	2.02	-0.14	-1.25	-1.96	1	8.29
8	1.37	0.01	-3.05	1.66	0	8.99

Closeness ranking

Table: Four-dimensional train samples ranked by distances

Sample	Feature 0	Feature 1	Feature 2	Feature 3	Label	Distance	Rank
0	-1.12	0.43	-1.5	0.55	1	1.62	1
1	1.93	-1.71	-0.75	-1.15	0	4.47	2
2	1.7	1.63	1.44	-0.42	1	5.24	4
3	-2.45	0.64	-0.48	0.17	1	4.73	3
4	1.14	-0.56	0.46	-1.04	1	6.04	5
5	-1.29	-1.58	-0.04	-2.11	0	6.87	6
6	-1.56	-1.13	-1.08	0.7	0	7.34	7
7	2.02	-0.14	-1.25	-1.96	1	8.29	8
8	1.37	0.01	-3.05	1.66	0	8.99	9

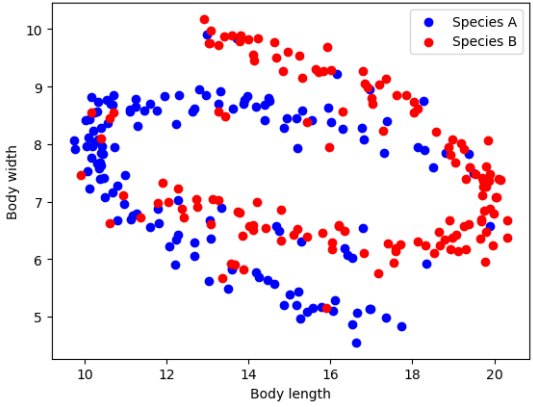
Nearest neighbor label

Table: Four-dimensional test sample labelled by its nearest neighbor

Feature 0	Feature 1	Feature 2	Feature 3	Label
-0.72	-0.41	1.21	-2.49	1

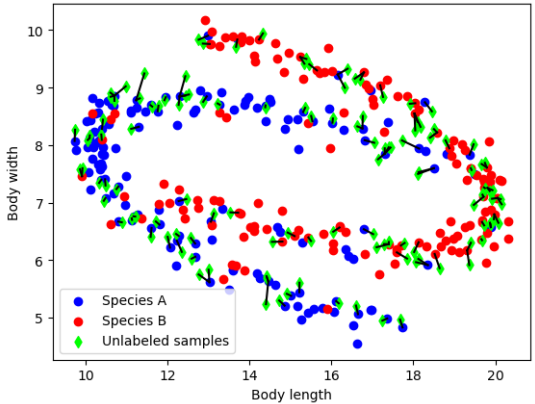
2D set recap

Figure: Recall of the 2D dataset



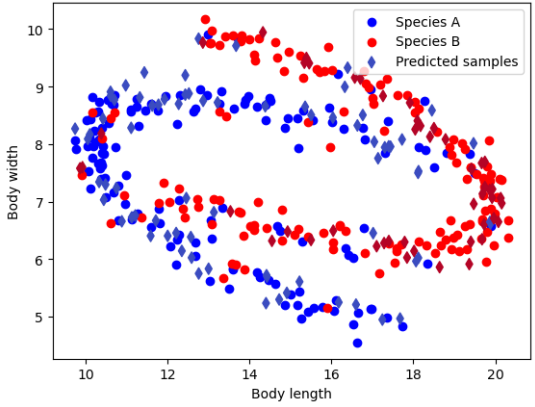
Edges to the nearest neighbor

Figure: Test samples connected to their nearest neighbor



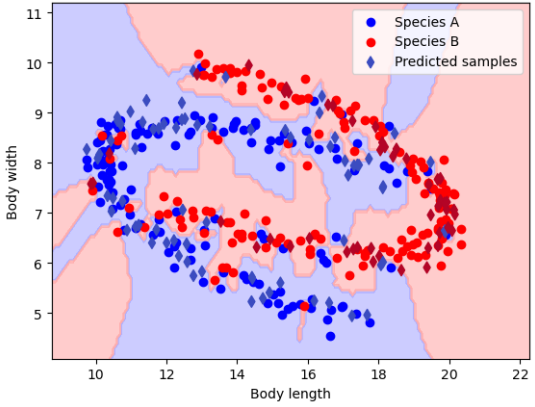
Nearest neighbor prediction

Figure: Test samples predicted by their nearest neighbor



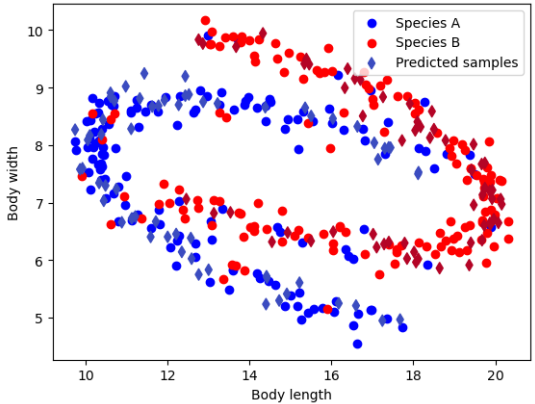
Decision boundary

Figure: Decision boundary of the nearest neighbor classifier



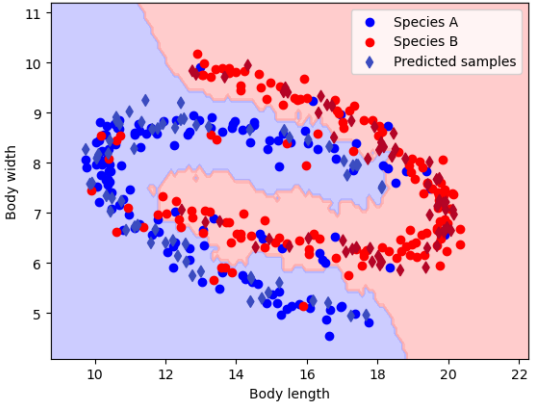
Prediction for 10-NN

Figure: Test samples predicted by their 10 nearest neighbors



Decision boundary for 10-NN

Figure: Decision boundary of the 10 nearest neighbors classifier



Dataset summary

Pima Indians Diabetes, (SMITH et al., 1988)

- 768 samples: female patients of Pima Indian heritage
- 5 features: glucose, blood pressure, skin thickness, insulin, BMI
- 2 classes: diabetes (positive) or not (negative)

Results

Table: Results for 10-fold cross-validation

k	Accuracy	Standard Deviation
1	0.67	0.05
3	0.74	0.04

Questions?

Thank you! Questions?

Tools and Theoretical background

Tools

- kNN model: (PEDREGOSA et al., 2011)
- Plotting: (HUNTER, 2007)

Theoretical background: (DUDA; HART; STORK, 2012)

References

DUDA, R.; HART, P.; STORK, D. **Pattern Classification**. [S.l.]: Wiley, 2012. ISBN 9781118586006. 24

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, IEEE Computer Society, v. 9, n. 3, p. 90–95, 2007. 24

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. 24

SMITH, J. W. et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: **Proceedings of the Annual Symposium on Computer Applications in Medical Care**. [S.l.]: IEEE Computer Society Press, 1988. p. 261–265. 21