THE BIRTH OF DS

DSTA

FLOWER CLASSIFICATION AND THE BIRTH OF DATA SCIENCE

CLASSIFYING IRIS FLOWERS

[Fisher, 1936]

Can flower samples be assigned to their proper sub-family purely on the basis of quantitative observation?

- Linear discriminant classification
- high-quality, annotated dataset

technique and data are interwined!

A formal description of the classification problem

Instance:

- n datapoints, each having over d-1 numerical dimensions $\mathcal{D}_1, \ldots \mathcal{D}_{d-1}$
- an expert classification function over k categories

Solution:

a linear combination $\mathcal{D}_1 \times \mathcal{D}_2 \times \ldots \mathcal{D}_{d-1} \rightarrow \mathcal{D}_d$

that **respects** the given classification.

Measure: *agreement* with the given classification.

THE IRIS DATASET

n=150 samples manually assigned by Fisher.

d=5 dimensions, four measurements (in cm) and the classification one

k=3 classes: Setosa, Versicolour and Virginica, 50 instances each, all available from scikit-learn

```
pip install scikit-learn
```

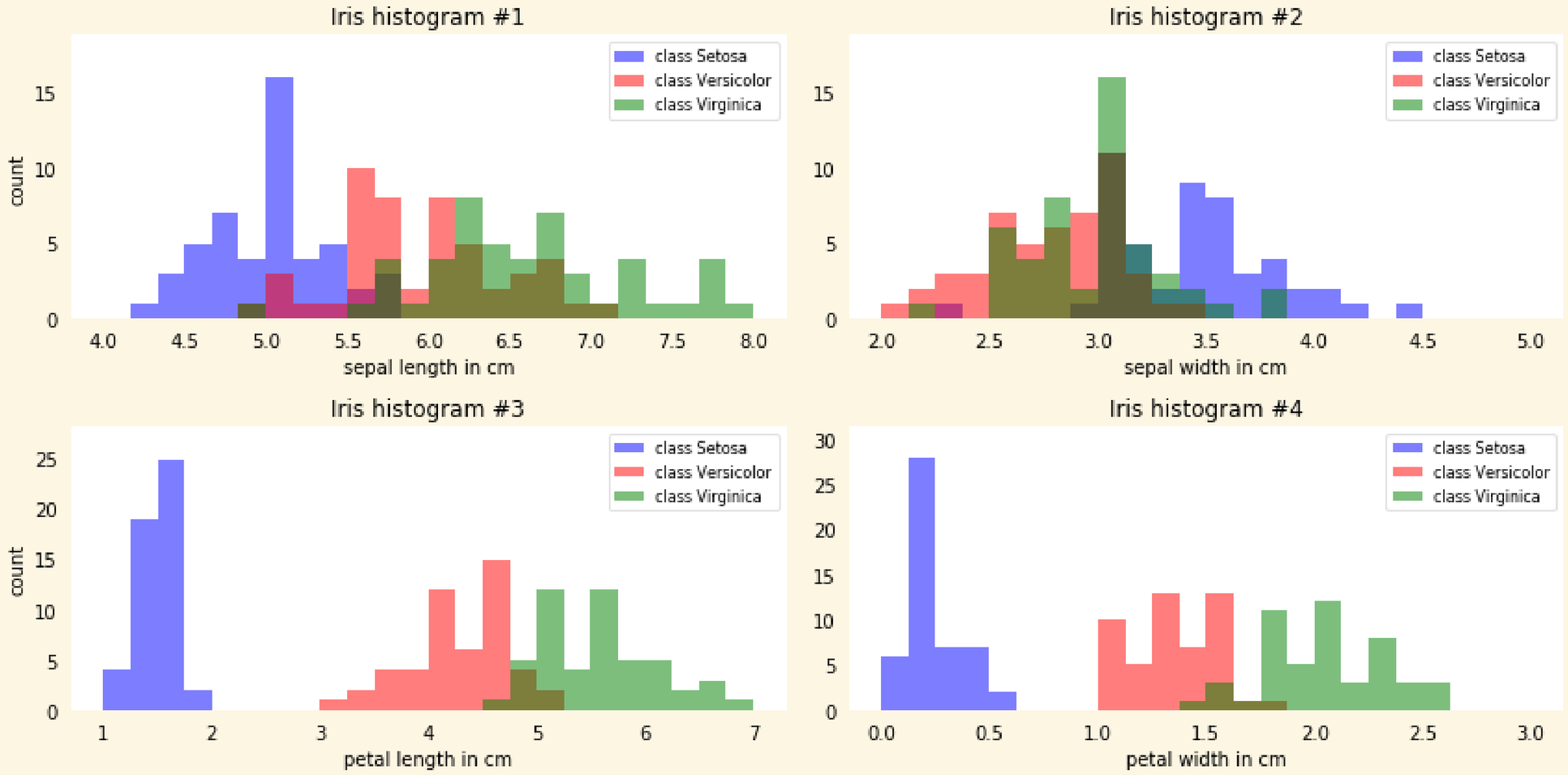
```
from sklearn import datasets

iris = datasets.load_iris()

print(iris['data'])

print(iris['target'])
```

FREQUENCY HISTOGRAM



A linear classifier corresponds to to a line drawn on the data display which creates two classification areas; more than one line is possible.

Whereas Setosa can be linearly separated, e.g., *petal_lenght <2* in the third column, the other two classes can’t be perfectly separated.

QUANTIFY AGREEMENT?

Q: Can we accept a linear combination that gives the correct answer only 19 times over 20?

A: It depends on the application.

Given two putative classifiers, which is the best?

Proposed answer:

At the same level of *precision*, (fraction of cases for which the classifier agrees with the expert classification)

prefer the one that *errs* less on the clear-cut cases.

IDEA: SUBSET SELECTION

ignore the less informative dimensions

IDEA: DIMENSION REDUCTION

Take a 2D scatterplot and map it to a line: does it improve visual classification?

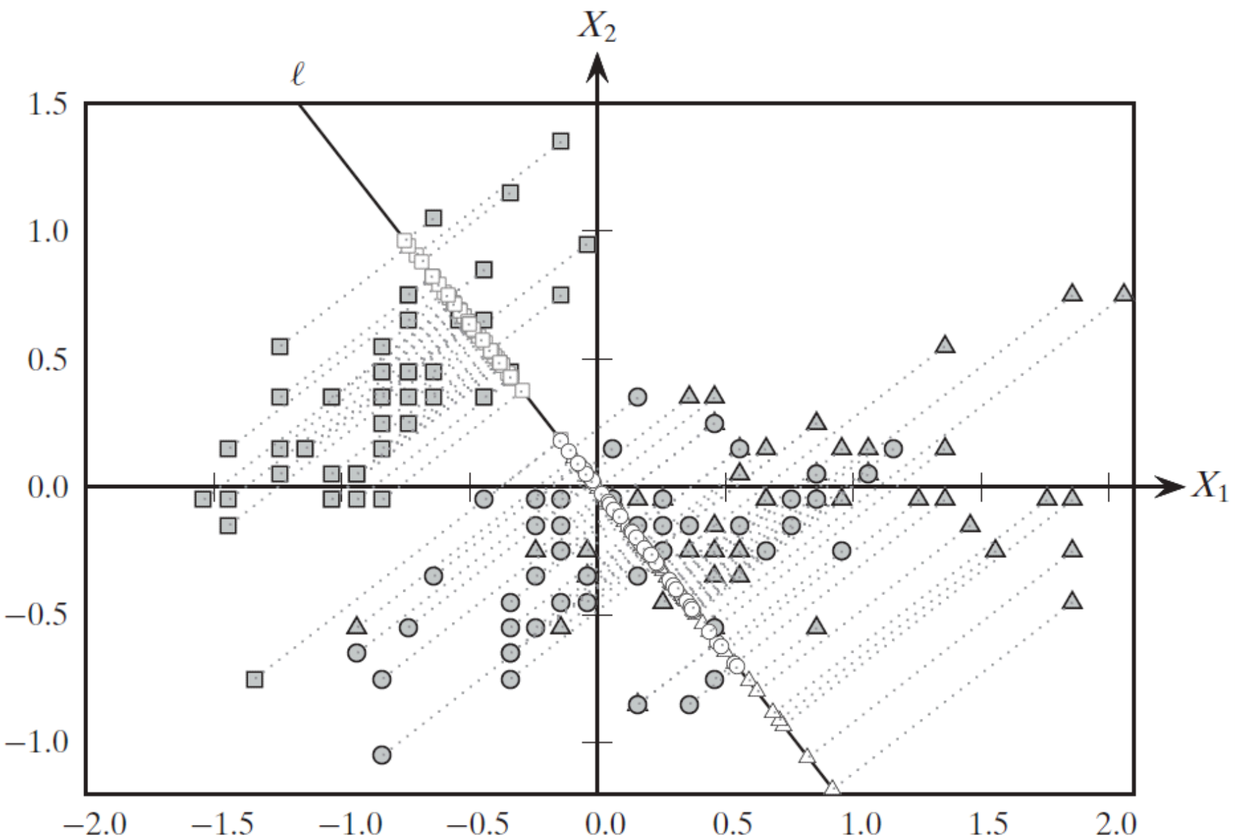


Figure 1.5. Projecting the centered data onto the line ℓ.

IDEA: SHRINKAGE

find a predictor where all predictors are used, but some are given less weight.

STUDY PLAN

This section, with the follow-up lab experience, is self-contained.

If you want more background you may read the PDF excerpt from the advanced Zaki-Meira textbook, which is available for download.

THE BIRTH OF THE NEW SCIENCE OF DATA

Fisher did not practice Statistics per se as he didn't try to estimate the distribution of tiny flowers in Canada, nor did he estimate measurement errors.

Rather, he asked whether classification could become somehow **automatic**, without the need to actually *see* the flower.