Introduction to Machine Learning and Pattern Recognition

David Brady[1]

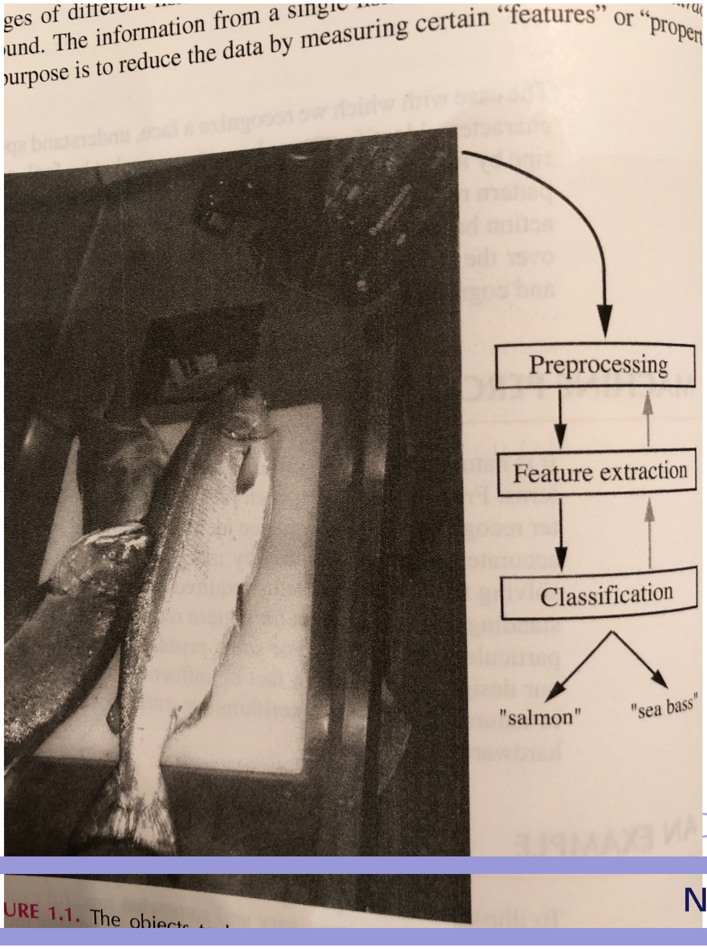
[1]ECE Department
Northeastern University

Fall 2018

The Classification Problem

Solutions to classification problems involve:

- data sensing
  - equipment, installation, data storage and retrieval
- preprocessing
  - segmentation (isolation of fish)
  - grouping (tail + torso + head)
- feature extraction
  - invariance to scale, 3D rotation, etc.
  - length, weight, lightness, etc.
- classification design: (this course)
  - bass or salmon?
  - use feature space ! (next)
  - missing data



ges of different
und. The information from a single
urpose is to reduce the data by measuring certain “features” or “propert

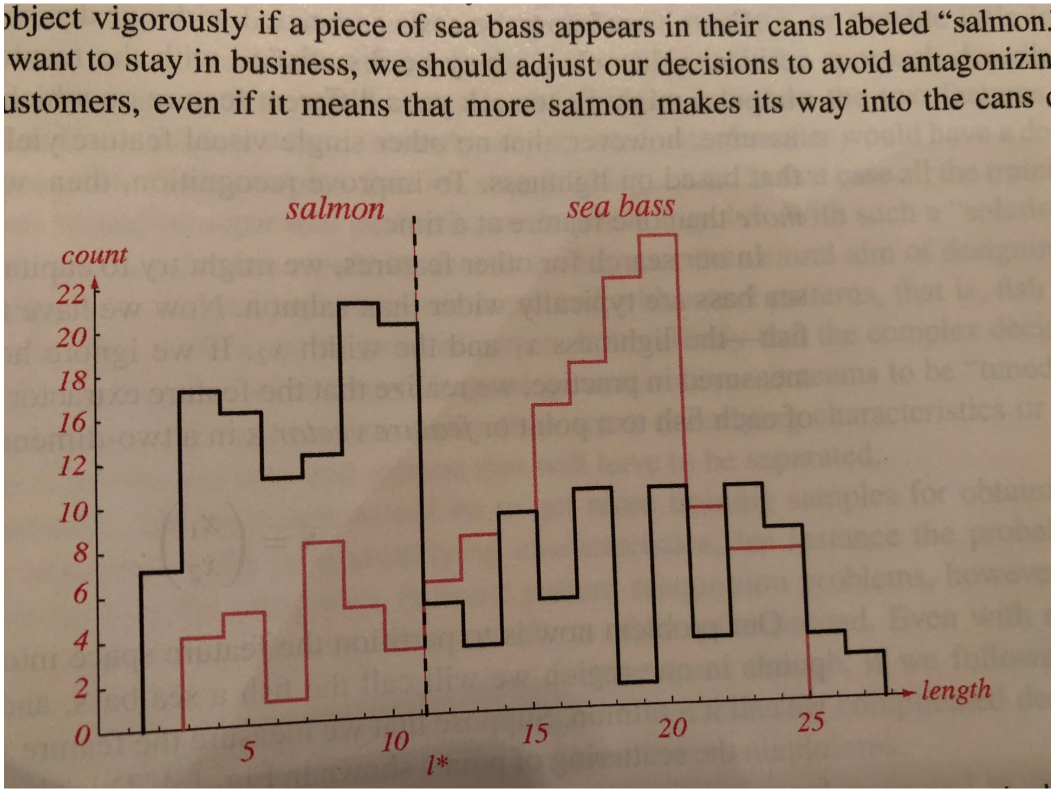
URE 1.1. The objects

Decision Boundary in Feature Space

Placing the decision boundary depends on:

- cost of each type of error
- *prior* probabilities

Why is *length* a poor feature for distinguishing fish?



bject vigorously if a piece of sea bass appears in their cans labeled “salmon. want to stay in business, we should adjust our decisions to avoid antagonizin ustomers, even if it means that more salmon makes its way into the cans

Figure: Histograms for length feature.

Feature Selection

Feature selection depends on:

- sensing noise
- classification problem
- class invariance
- training samples

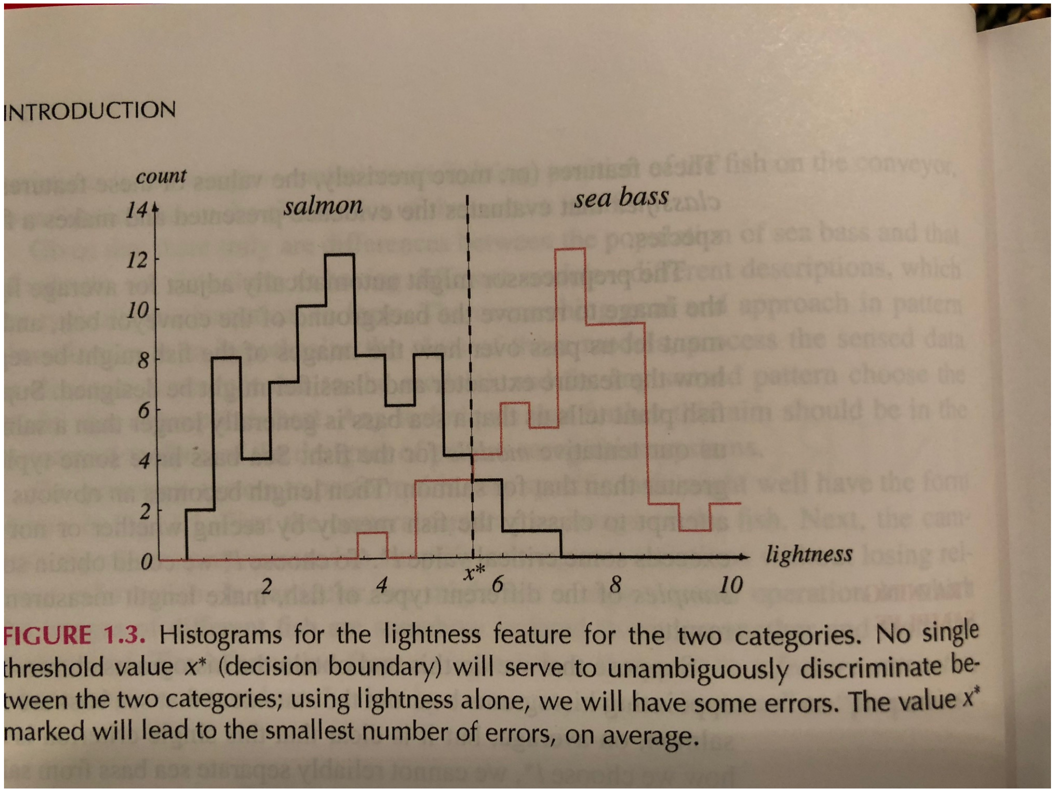

INTRODUCTION

FIGURE 1.3. Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average.

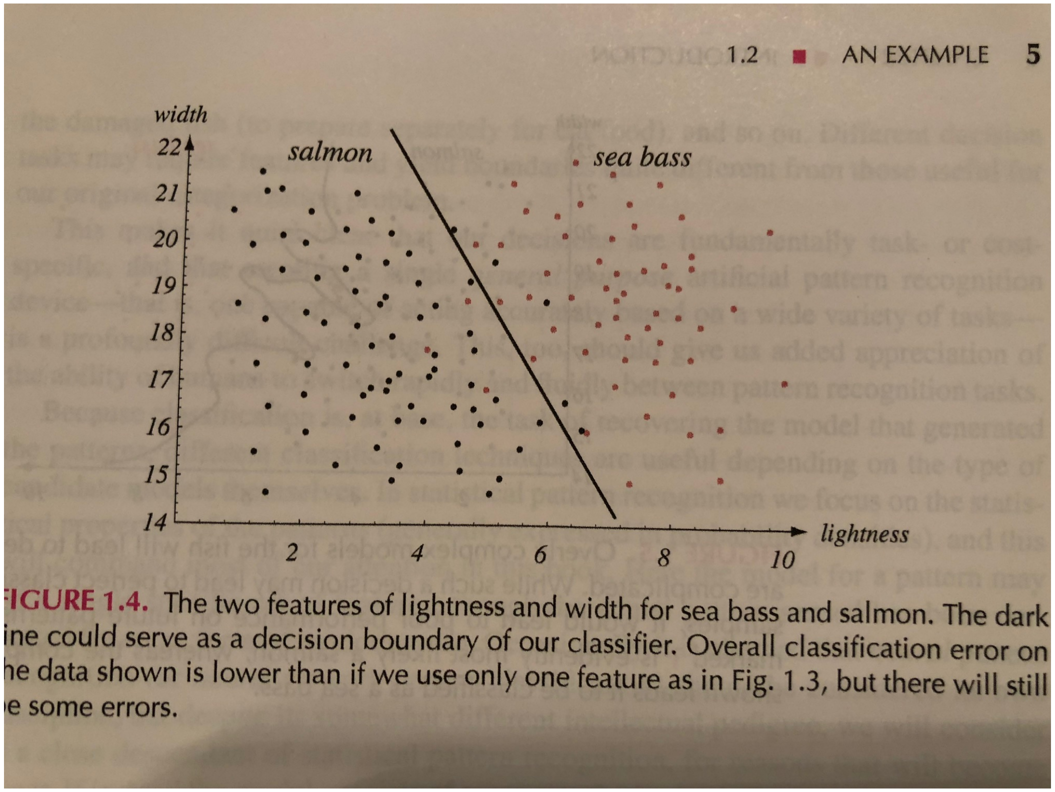
Figure: Histograms for lightness feature.

>2D Feature Space

Additional features provide:

- discrimination
- information
- *complexity*
- *increased training size*

How should you select features?



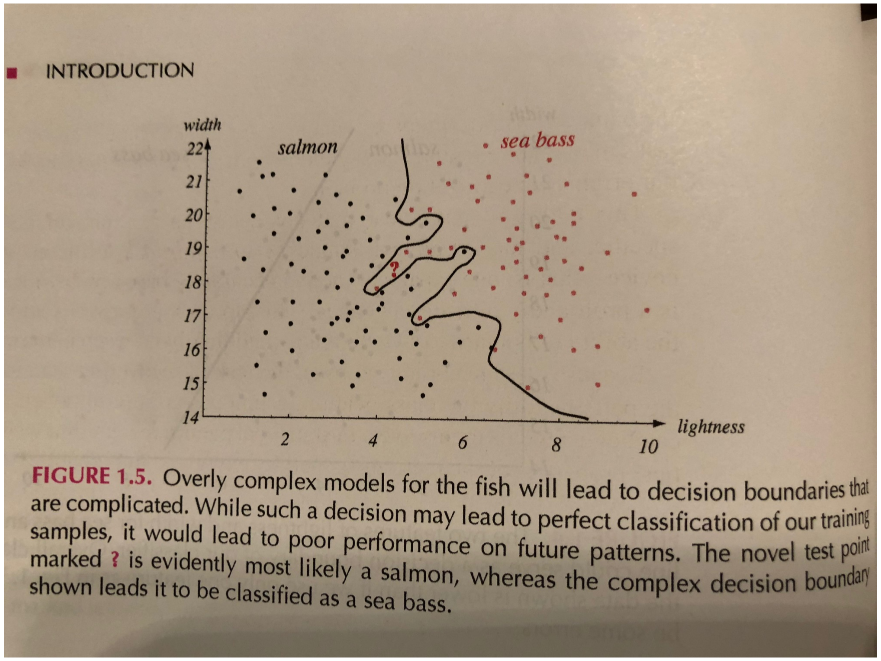
1.2 AN EXAMPLE 5

IGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark ine could serve as a decision boundary of our classifier. Overall classification error on he data shown is lower than if we use only one feature as in Fig. 1.3, but there will still e some errors.

Figure: 2D feature space with a decision boundary

Overfitting

Overly-complex boundaries follow noise!

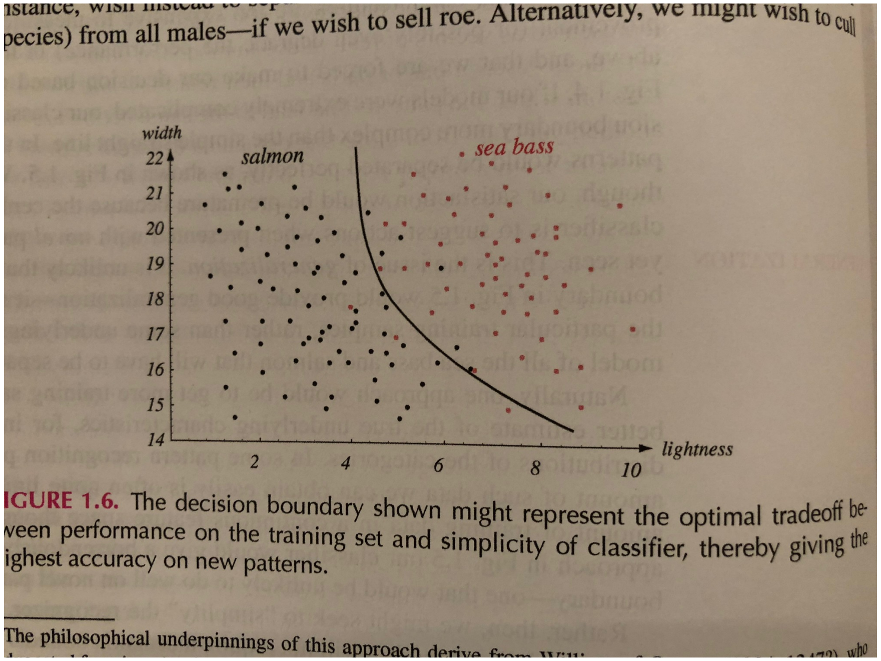


INTRODUCTION

width
22
21
20
19
18
17
16
15
14
salmon
sea bass
?
2
4
6
8
10
lightness

FIGURE 1.5. Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked ? is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be classified as a sea bass.

Use the simplest possible boundaries.



nstance, wish instead to sep

pecies) from all males—if we wish to sell roe. Alternatively, we might wish to cull

width
22
21
20
19
18
17
16
15
14
salmon
sea bass
2
4
6
8
10
lightness

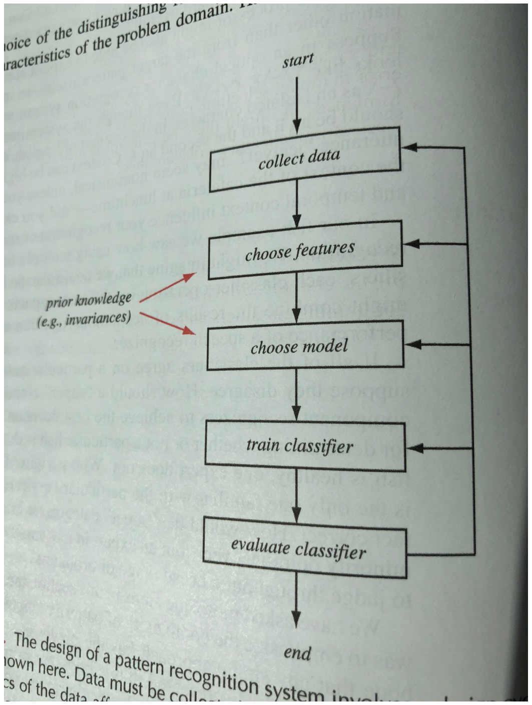
IGURE 1.6. The decision boundary shown might represent the optimal tradeoff between performance on the training set and simplicity of classifier, thereby giving the ighest accuracy on new patterns.

The philosophical underpinnings of this approach derive from

Classifier Design Cycle

Design Tweaking Issues: (improve performance, reduce complexity)

- data collection
  - relevant data for features
- feature choice
  - retain fewest
- model choice
  - connects features & hypotheses
  - prefer simpler models
- training
  - balance *overfitting* with *typicality*



- evaluation
  - criteria (error rates, costs)
  - confidence bounds on criteria (training size)
- computational complexity – as much as your budget can handle!