# DECISION TREE AND RANDOM FOREST

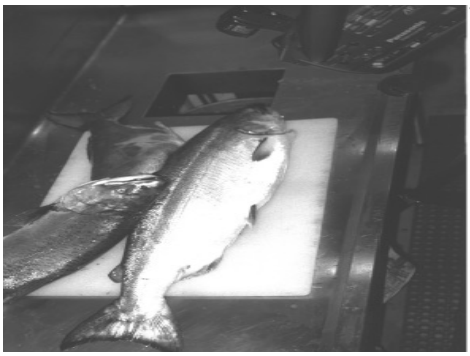
Vivek Kanhangad
Department of Electrical Engineering
Indian Institute of Technology Indore

1

## Pattern Recognition: An example

2

- A simple example: Identifying species of a fish on a conveyor belt
  - Species: Sea bass and salmon



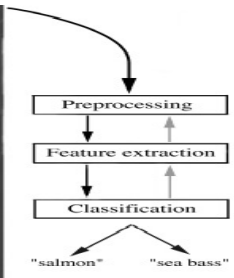


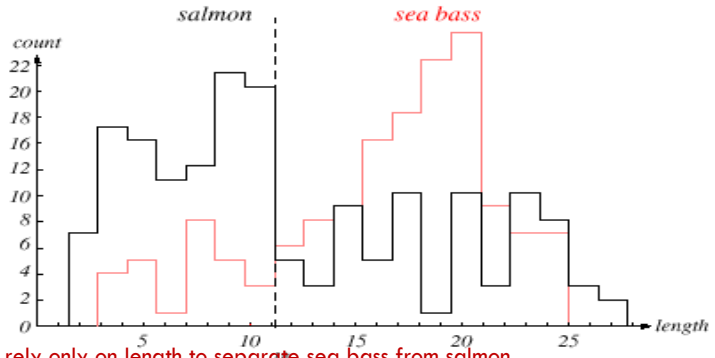

Image source: Pattern Classification by Duda, Hart and Stork

2

## Pattern Recognition: An example

3

- Single feature based classification
  - Feature: Length




We cannot rely only on length to separate sea bass from salmon

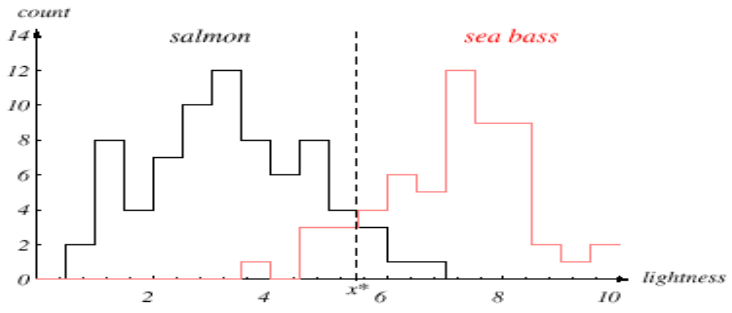
Image source: Pattern Classification by Duda, Hart and Stork

3

## Pattern Recognition: An example

4

- Single feature based classification
  - Feature: Lightness




The two classes are much better separated using lightness feature!

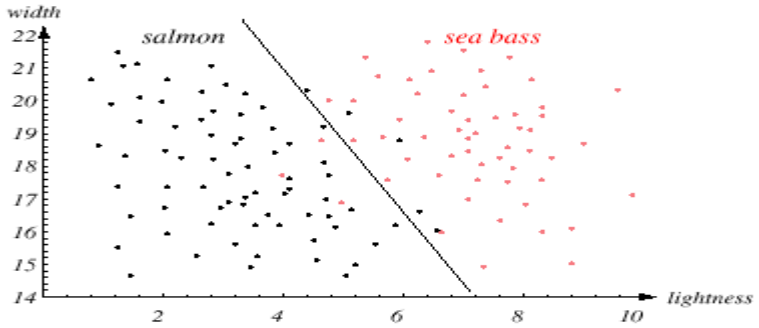
Image source: Pattern Classification by Duda, Hart and Stork

4

## Pattern Recognition: Feature Space

5

- Two features for classification




Can we improve the performance further? If yes, how?

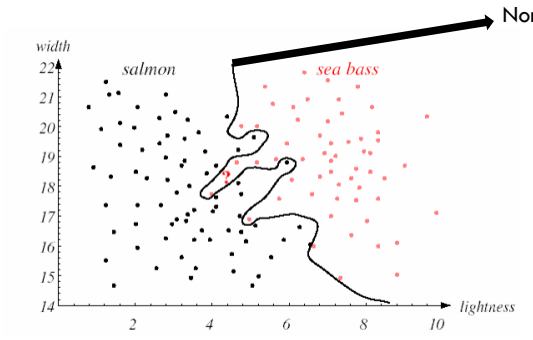
Image source: Pattern Classification by Duda, Hart and Stork

5

## Pattern Recognition: Feature Space

6

- Two features for classification




Perfect Classification! Is there a catch?

Image source: Pattern Classification by Duda, Hart and Stork

6

## Pattern Recognition: Generalization

7

- Classification Goal: Make accurate predictions for new/unseen data - Good Generalization
- The model should NOT be tuned to the specific characteristics of the training data – Overfitting
- In practice, training data is likely to contain some noise

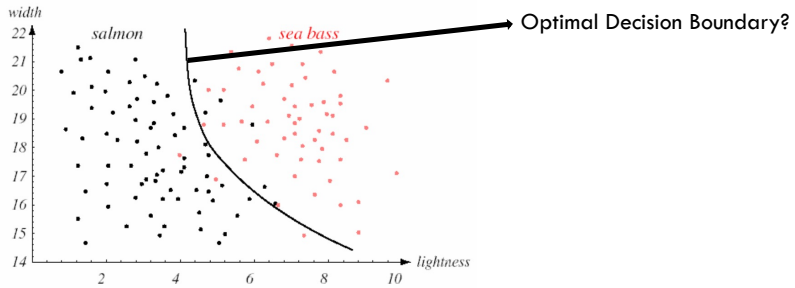
We are better off with a slightly poorer performance on the training examples if this means that our classifier will have better performance on unseen patterns.

7

## Pattern Recognition: Generalization

8

- Classification Goal: Make accurate predictions for new/unseen data - Good Generalization



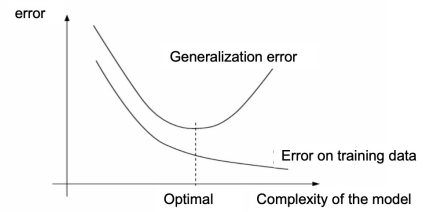

- A decision boundary that provides an optimal tradeoff between accuracy on the training set and unseen data

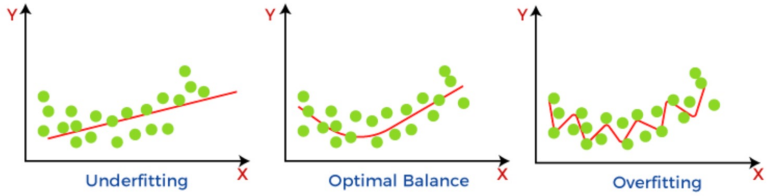
8

## Pattern Recognition: Avoid Overfitting and Achieve Optimal Tradeoff

9

- Evaluate the classifier model on unseen data – Validation Set




9

## Gini Impurity to Build Decision Trees

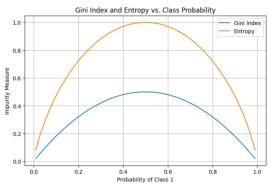
10

| | age | income | student | credit_rate | default |
|---|---|---|---|---|---|
| 0 | youth | high | no | fair | no |
| 1 | youth | high | no | excellent | no |
| 2 | middle_age | high | no | fair | yes |
| 3 | senior | medium | no | fair | yes |
| 4 | senior | low | yes | fair | yes |
| 5 | senior | low | yes | excellent | no |
| 6 | middle_age | low | yes | excellent | yes |
| 7 | youth | medium | no | fair | no |
| 8 | youth | low | yes | fair | yes |
| 9 | senior | medium | yes | fair | yes |
| 10 | youth | medium | yes | excellent | yes |
| 11 | middle_age | medium | no | excellent | yes |
| 12 | middle_age | high | yes | fair | yes |
| 13 | senior | medium | no | excellent | no |

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

$$Gini_A(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2)$$

$$\triangle Gini(A) = Gini(D) - Gini_A(D)$$



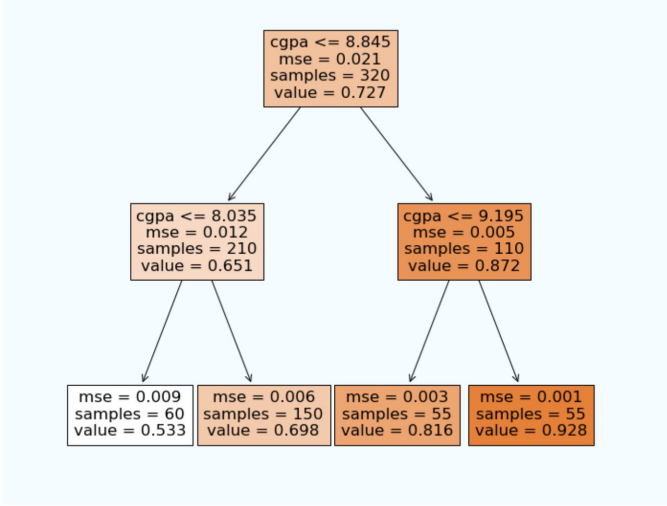


Gini Impurity for Credit Rating is 0.429

10

# Decision Tree for Regression

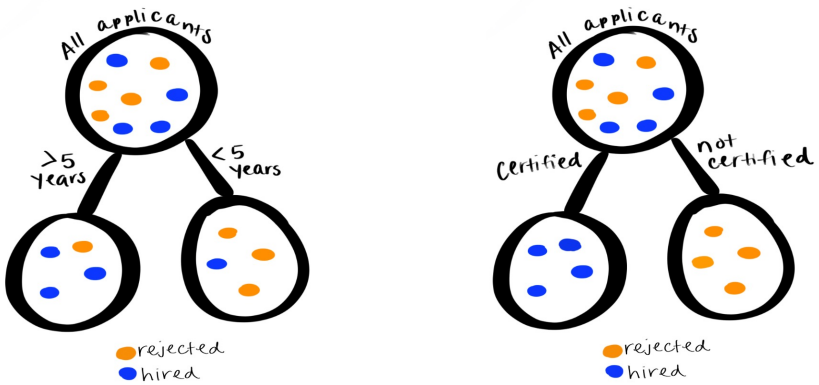
11



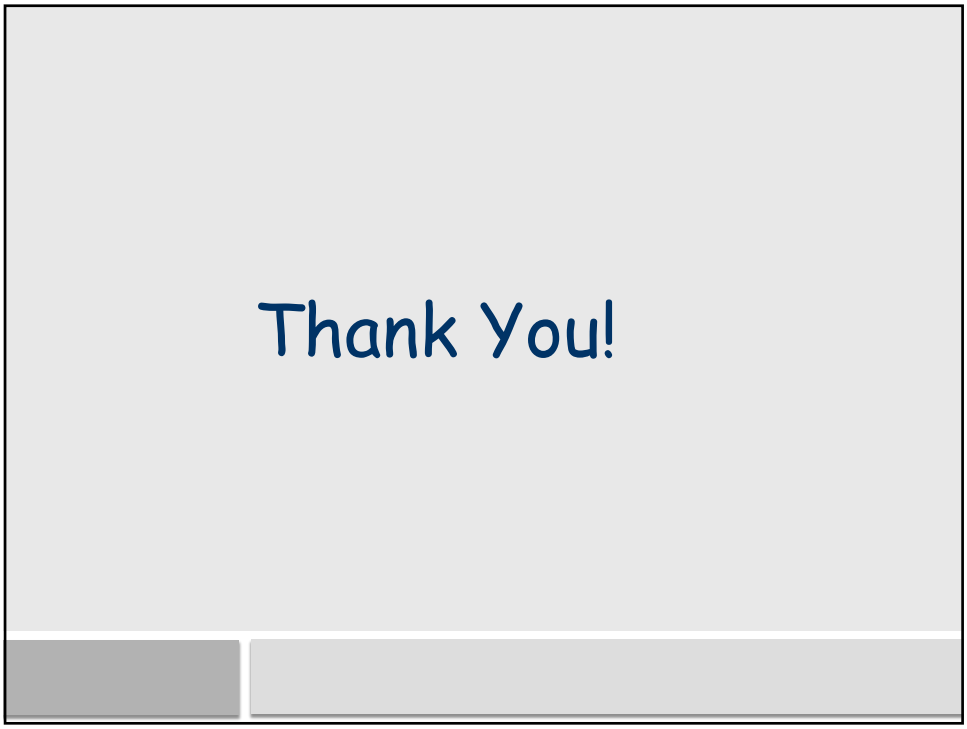

11

# Feature Importance – Gini Impurity

12




12




13