SNS COLLEGE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION)

Approved by AICTE & Affiliated to Anna University
Accredited by NBA & Accrediated by NAAC with 'A+' Grade,
Recognized by UGC saravanampatti (post), Coimbatore-641035.

Department of Biomedical Engineering

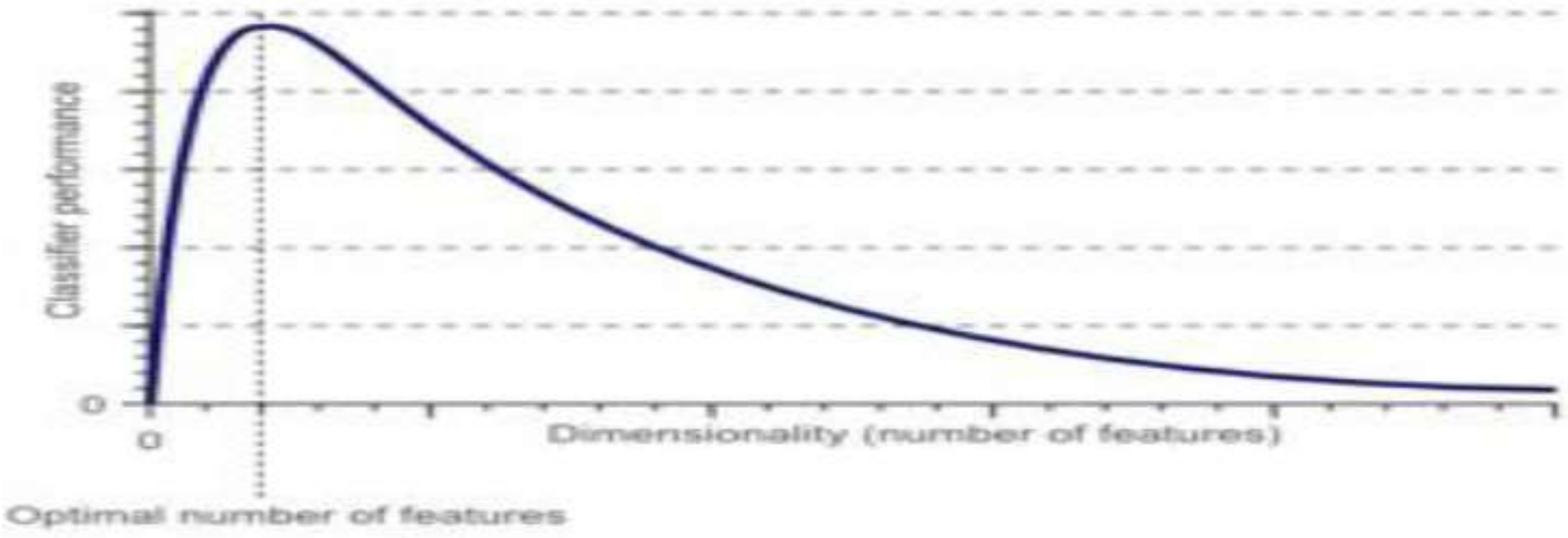
Course Name: 19ECT303 & Artificial Intelligence and machine learning

III Year : V Semester

Unit I –FUNDAMENTALS OF MACHINE LEARNING

Topic : Curse of dimensionality

Performance Vs Dimensionality



Curse of dimensionality

- The curse of dimensionality basically means that the error increases with the increase in the number of features.
- It refers to the fact that algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions.

• Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data.
• The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset.
•A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data.

• Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models.
• The difficulties related to training machine learning models due to high dimensional data are referred to as ‘Curse of Dimensionality’.
• The popular aspects of the curse of dimensionality; ‘data sparsity’and ‘distance concentration’ are discussed in the following sections.

Data Sparsity

- Training of supervised machine learning model for predicting the outcome of given input
- While training a model, the available data is used such that part of the data is used for training the model, and a part of the data is used to evaluate how the model performs on unseen data

Model generalization

Possible combinations of predictor variables and the corresponding targets

Data Concentration

pairwise distances between different samples/points in the space converging to the same value as the dimensionality of the data increases

Mitigating curse of dimensionality

Dimensionality reduction techniques fall into one of the two categories- ‘Feature selection’ or ‘Feature extraction’.

Feature selection technique

Low Variance filter- Test of variance of attributes
High correlation filter-High correlated attributes are eliminated
Multi-collinearity/VIF-High VIF features are eliminated
Feature ranking- Decision tree models
Feature selection-Adjusted R2 values

Feature extraction techniques

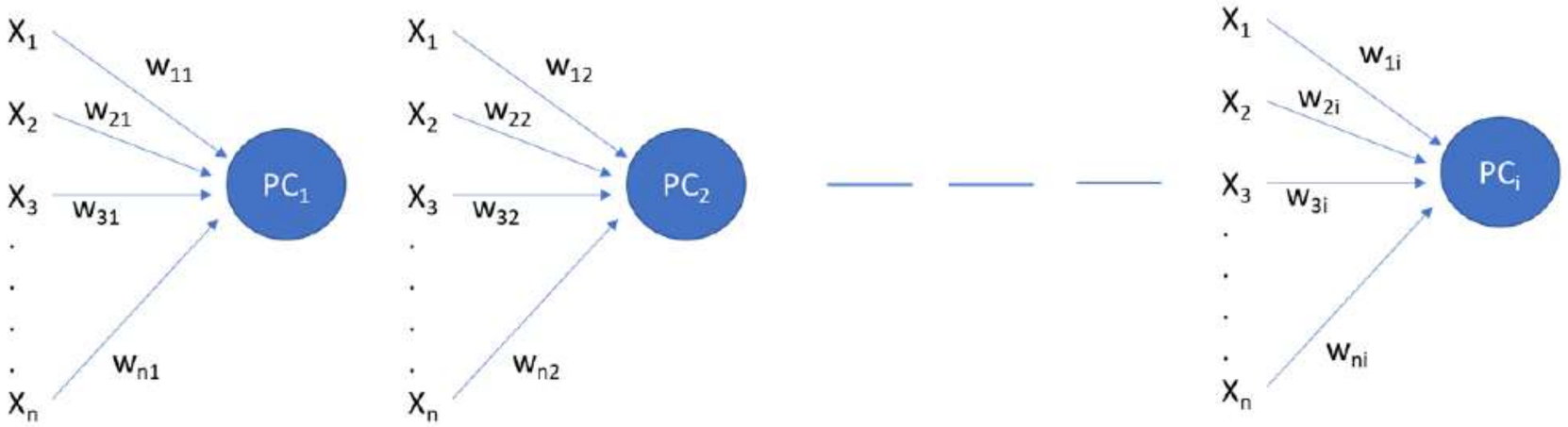
•In feature extraction techniques, the high dimensional attributes are combined in low dimensional components (PCA or ICA) or factored into low dimensional factors (FA).

• PCA- n dimension → n principle components

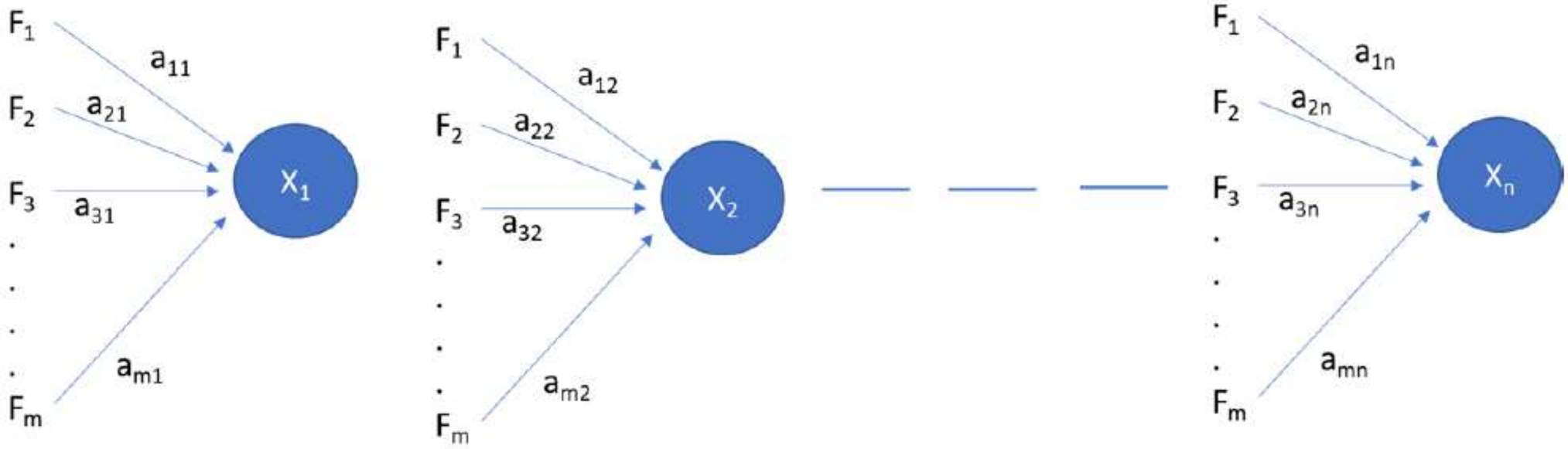
•FA- weighted linear combination of latent factors- n → m

ICA

•



Principle component analysis (i≤n)

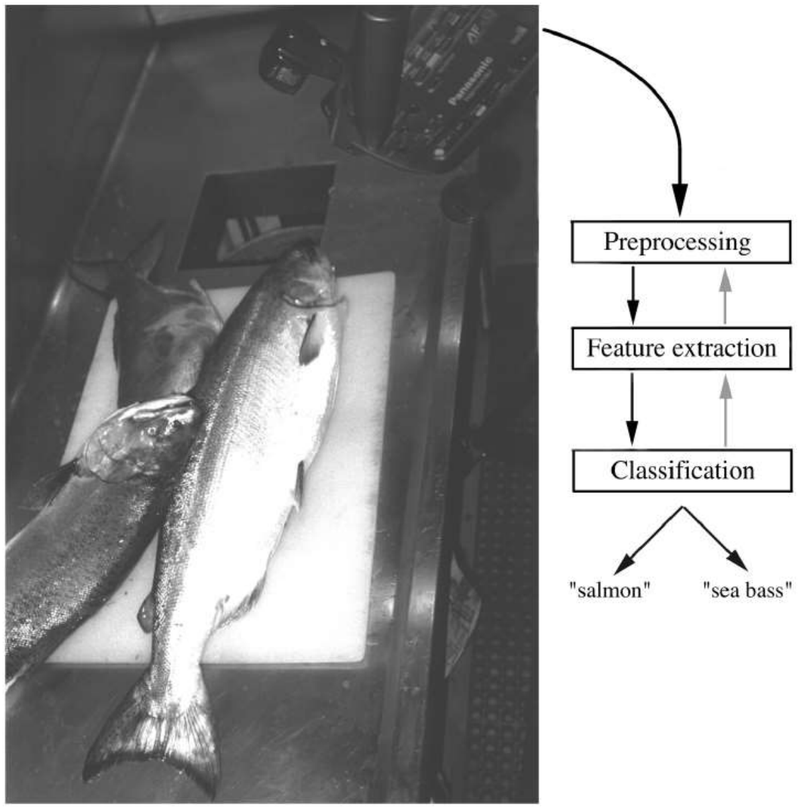


Factor analysis (m≤n)

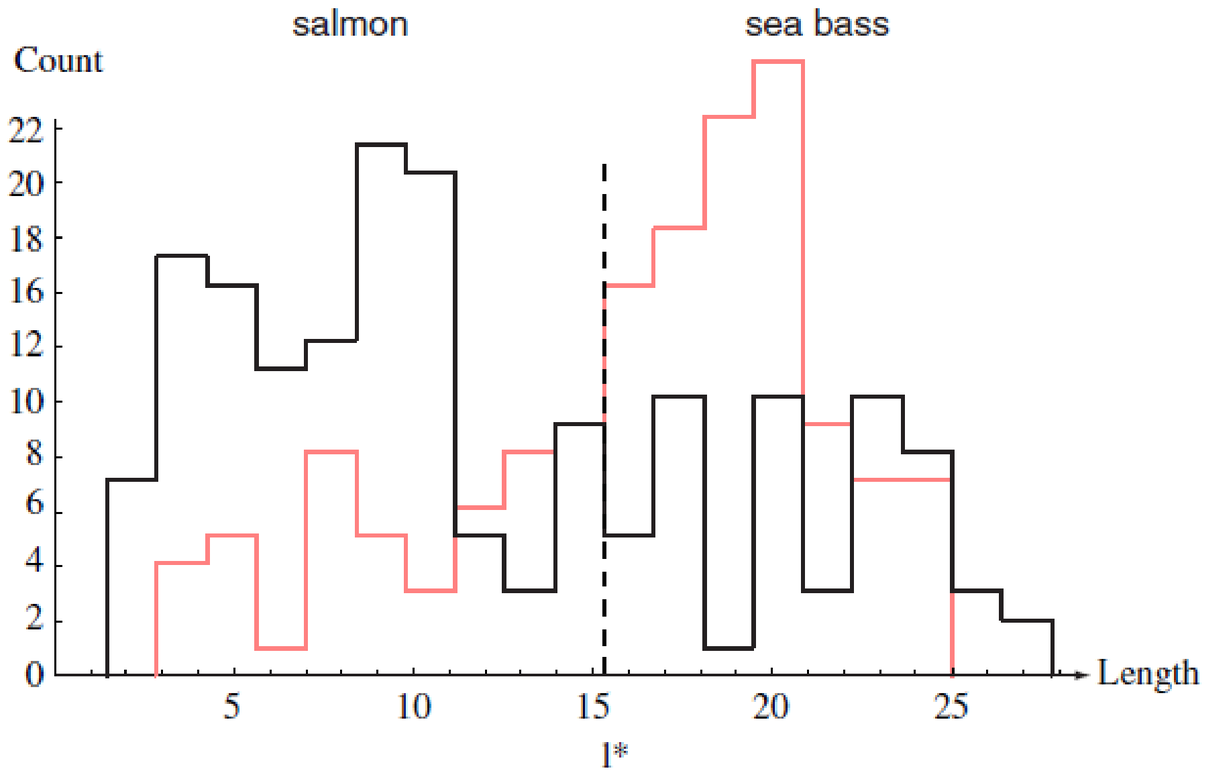
THINK AND ANSWER!!

Suppose that a fish packing plant wants to automate the process of sorting incoming fish on a conveyor belt according to species, using a machine learning process. Identify the method and features for this problem.

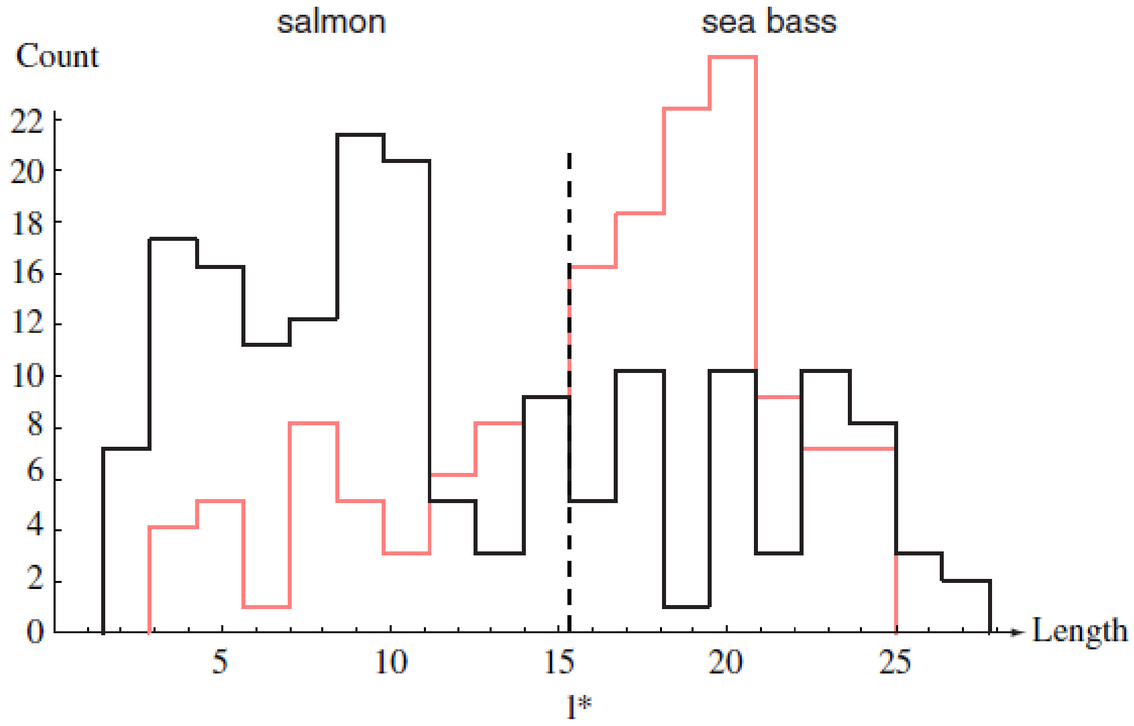
SOLUTION TO THE PROBLEM



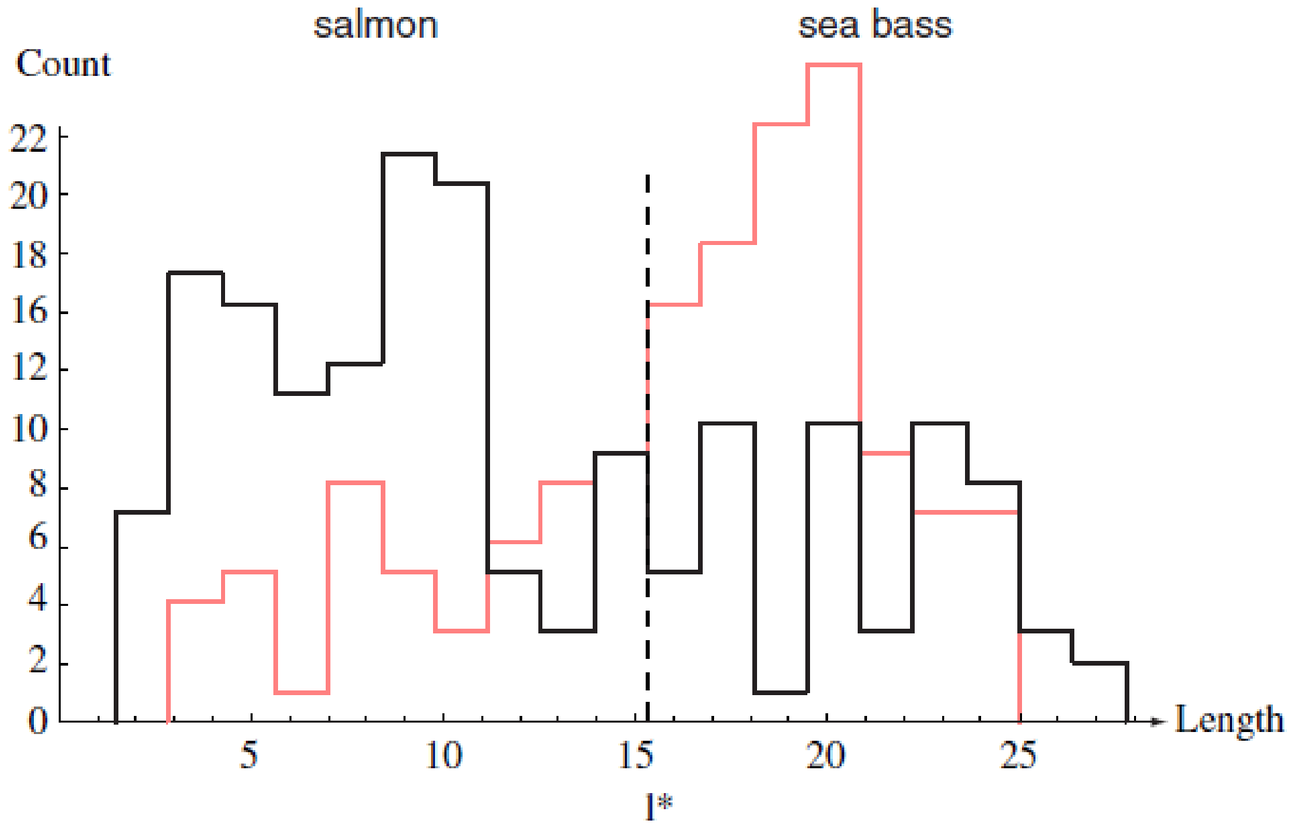
Length as the feature



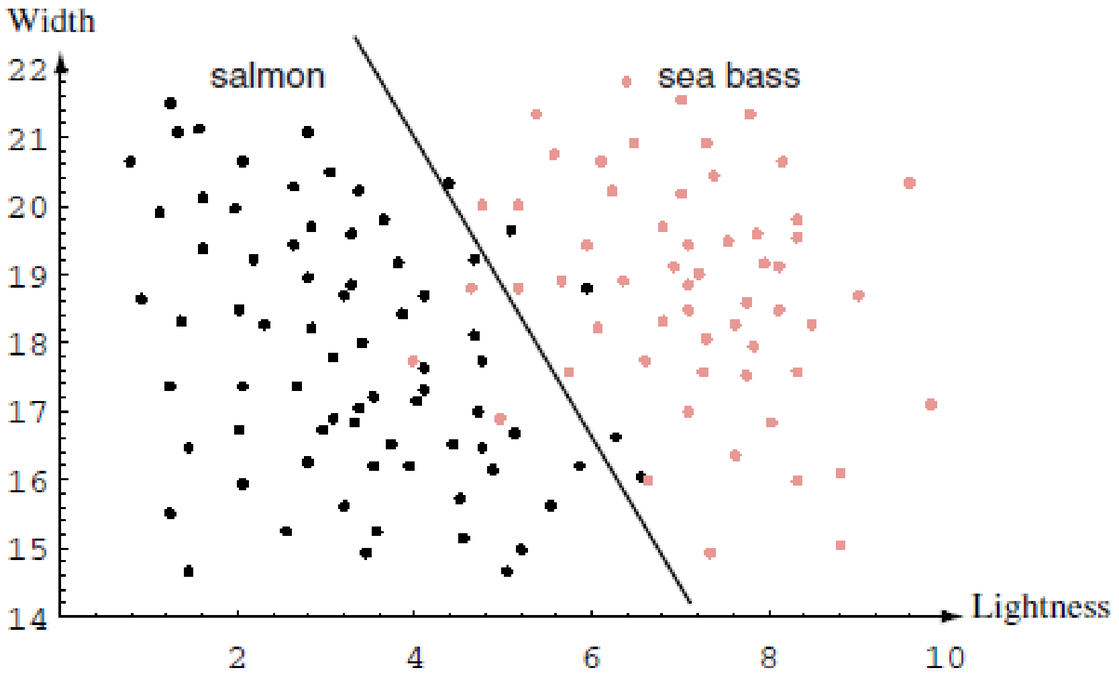
Length as the feature



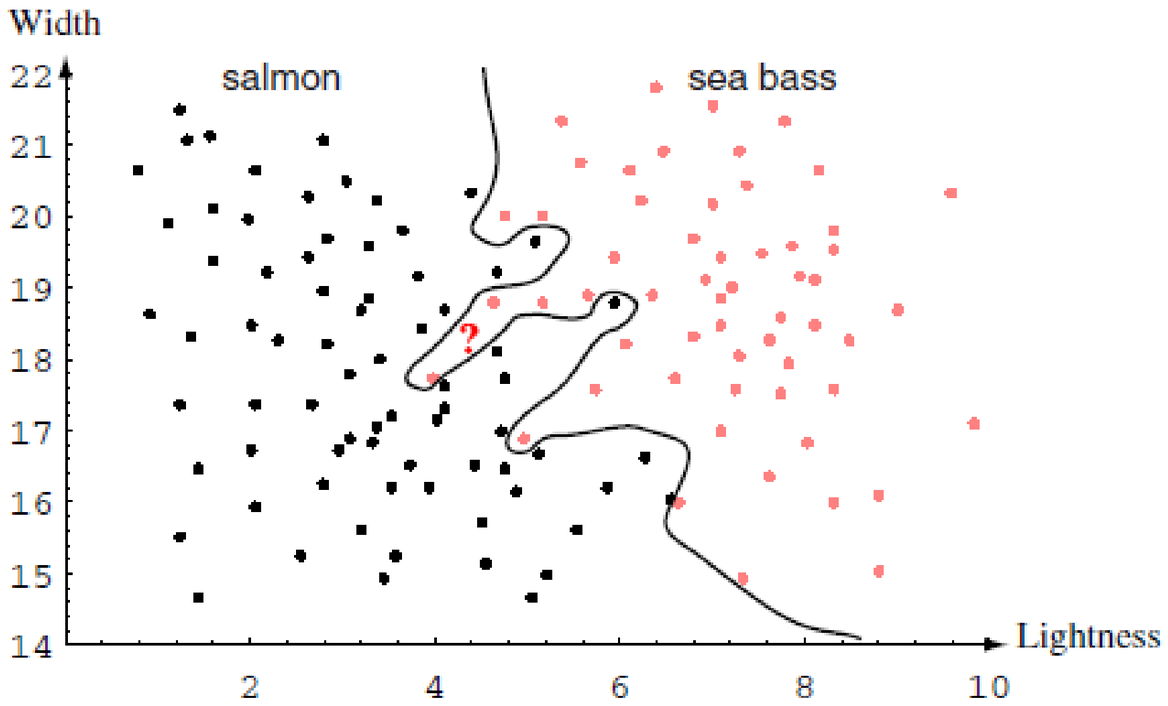
lightness as the feature



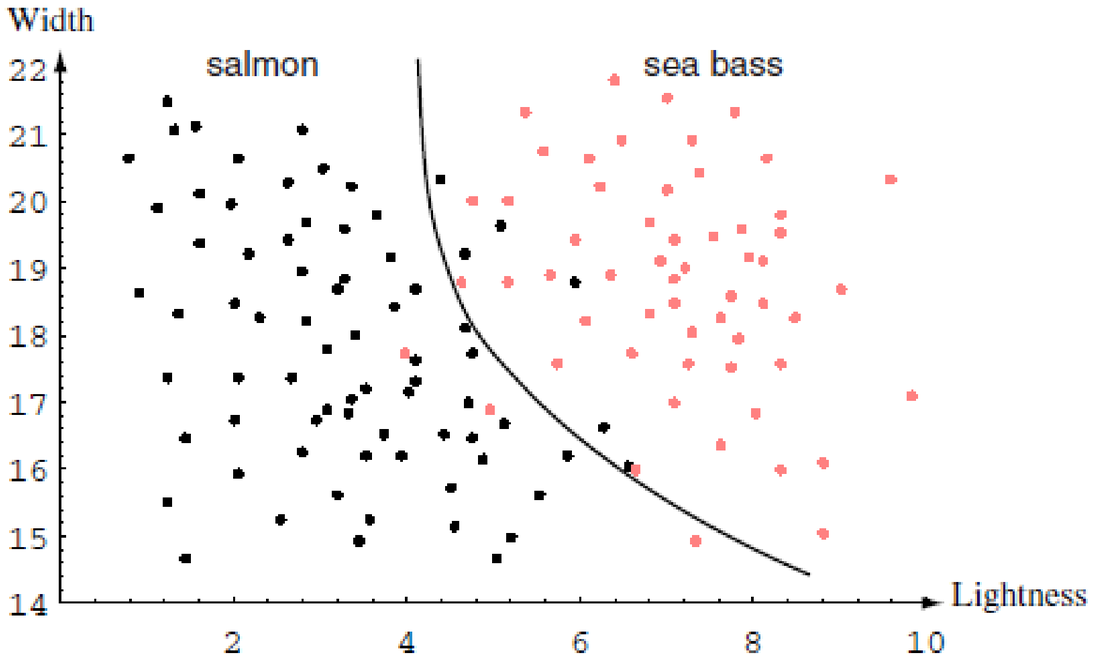
Lightness and width as the feature



Complicated feature models



optimal tradeoff between performance on the training set and simplicity of classifier



THANK YOU

QUESTIONS ??