Probabilistic Methods in Bioinformatics

Pabitra Mitra
pabitra@cse.iitkgp.ernet.in

Probability in Bioinformatics

- Classification
  - Categorize a new object into a known class
  - Supervised learning/predictive analysis
- Regression
  - Supervised prediction of continuous valued variables
- Clustering
  - Extract homogenous groups in population
  - Unsupervised learning/exploratory analysis
- Sequence analysis
- Relation and graph structure analysis

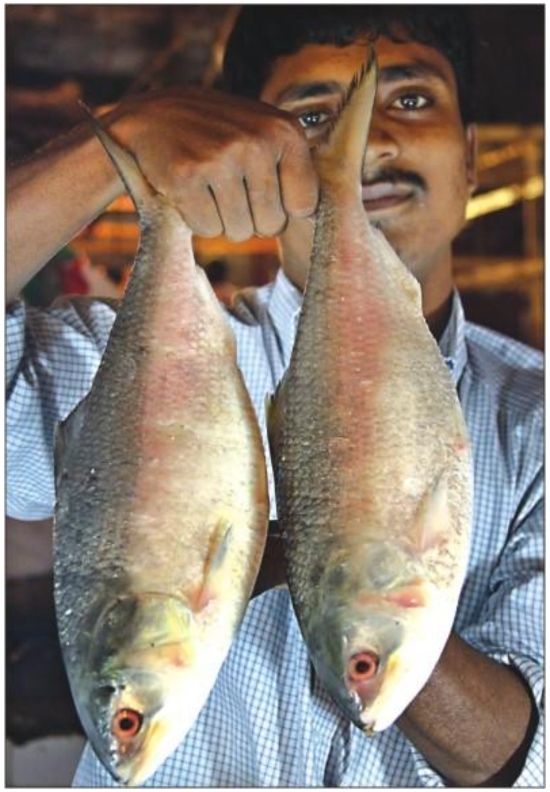
Probabilistic Algorithms

- Classification
  - Bayes classification, graphical models
- Regression
  - Logistic regression
- Clustering
  - Gaussian mixture models
- Sequence analysis
  - Markov models, hidden markov models, conditional random fields
- Relation analysis
  - Markov processes, graph structure analysis

many many more…..

A Simple Species Classification Problem

- Measure the *length* of a fish, and decide its <u>class</u>
  - Hilsa or Tuna

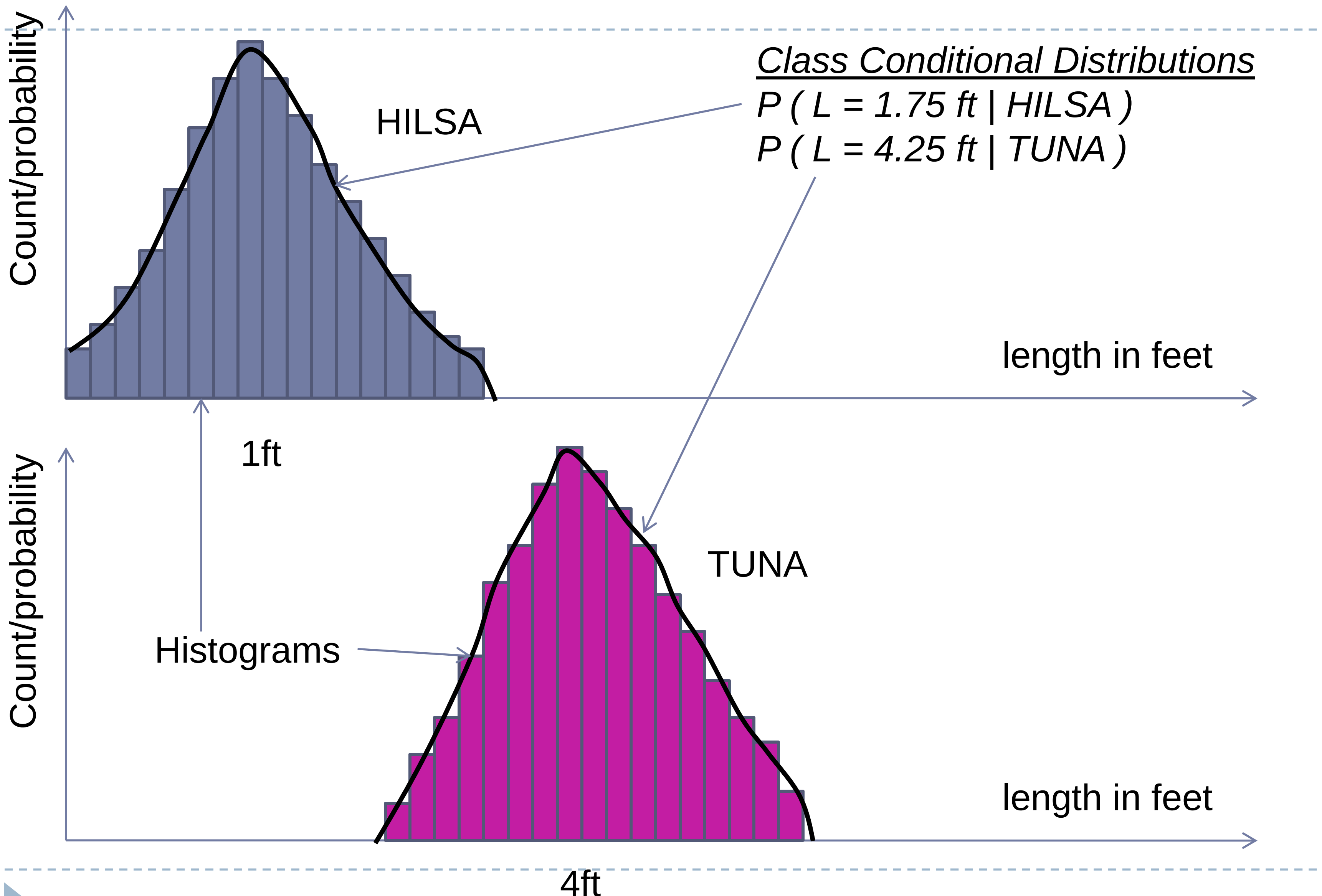


Collect Statistics …

Population for Class Hilsa

Population for Class Tuna

Distribution of "Fish Length"

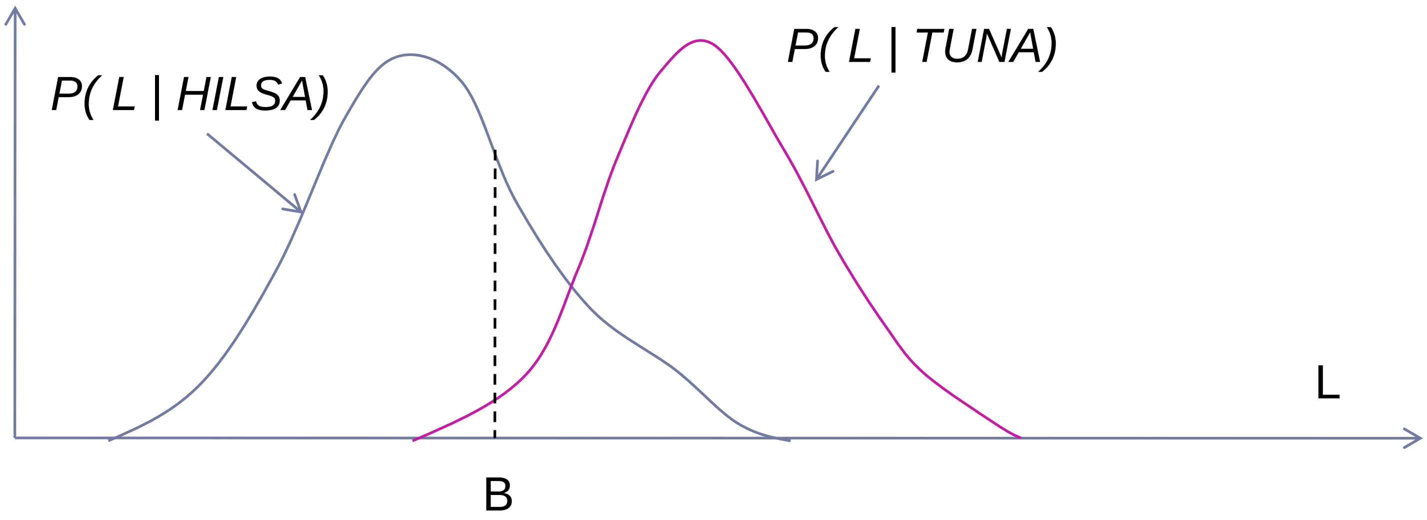


Decision Rule

- If length L ≤ B
  - HILSA
- ELSE
  - TUNA

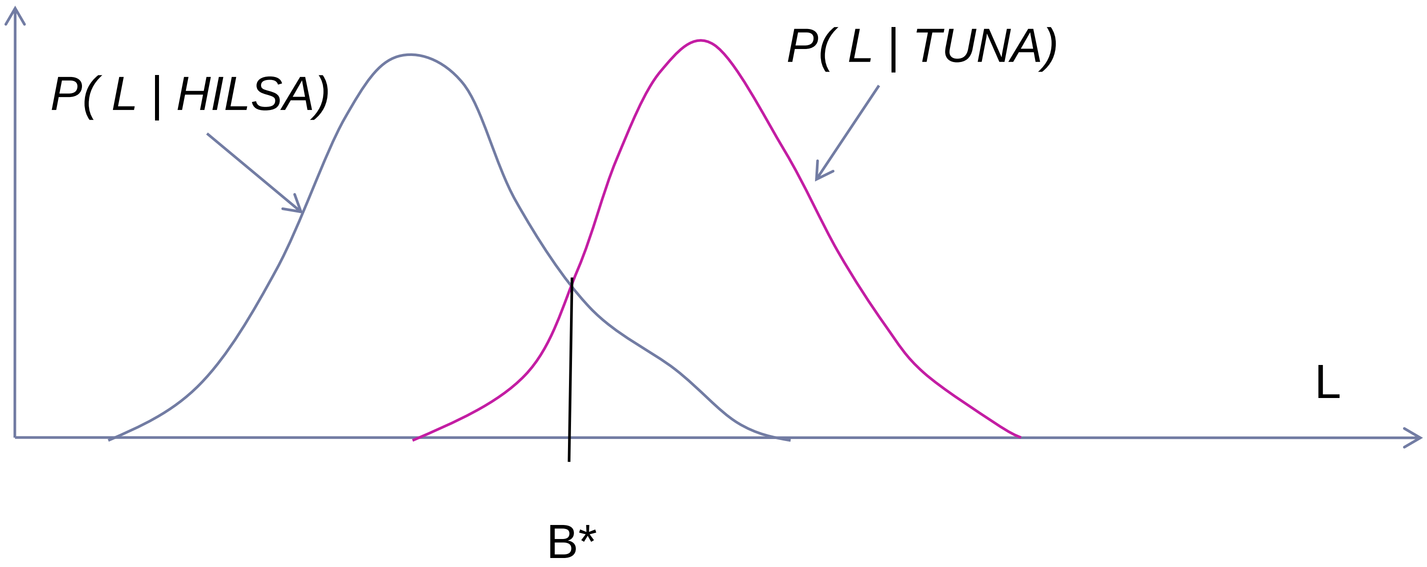
- What should be the value of B (“boundary” length) ?
  - Based on population statistics

Error of Decision Rule



Errors: Type 1 + Type 2,
Type 1: Actually Tuna, Classified as Hilsa (area under pink curve to the left of a B)
Type 2: Actually Hilsa, Classified as Tuna (area under blue curve to the right of a B)

Optimal Decision Rule



B*: Optimal Value of B, (Optimal Decision Boundary)

Minimum Possible Error

P (B* | HILSA) = P (B* | TUNA)

If Type 1 and Type 2 errors have different costs : optimal boundary shifts

Species Identification Problem

- Measure lengths of a (sizeable) population of Hilsa and Tuna fishes
- Estimate Class Conditional Distributions for Hilsa and Tuna classes respectively
- Find Optimal Decision Boundary B* from the distributions
- Apply Decision Rule to classify a newly caught (and measured) fish as either Hilsa or Tuna
  - (with minimum error probability)

Location/Time of Experiment

- Calcutta in Monsoon
  - More Hilsa few Tuna
- California in Winter
  - More Tuna less Hilsa

- Even a 2ft fish is likely to be Hilsa in Calcutta (2000 Rs/Kilo!),
- a 1.5ft fish may be Tuna in California

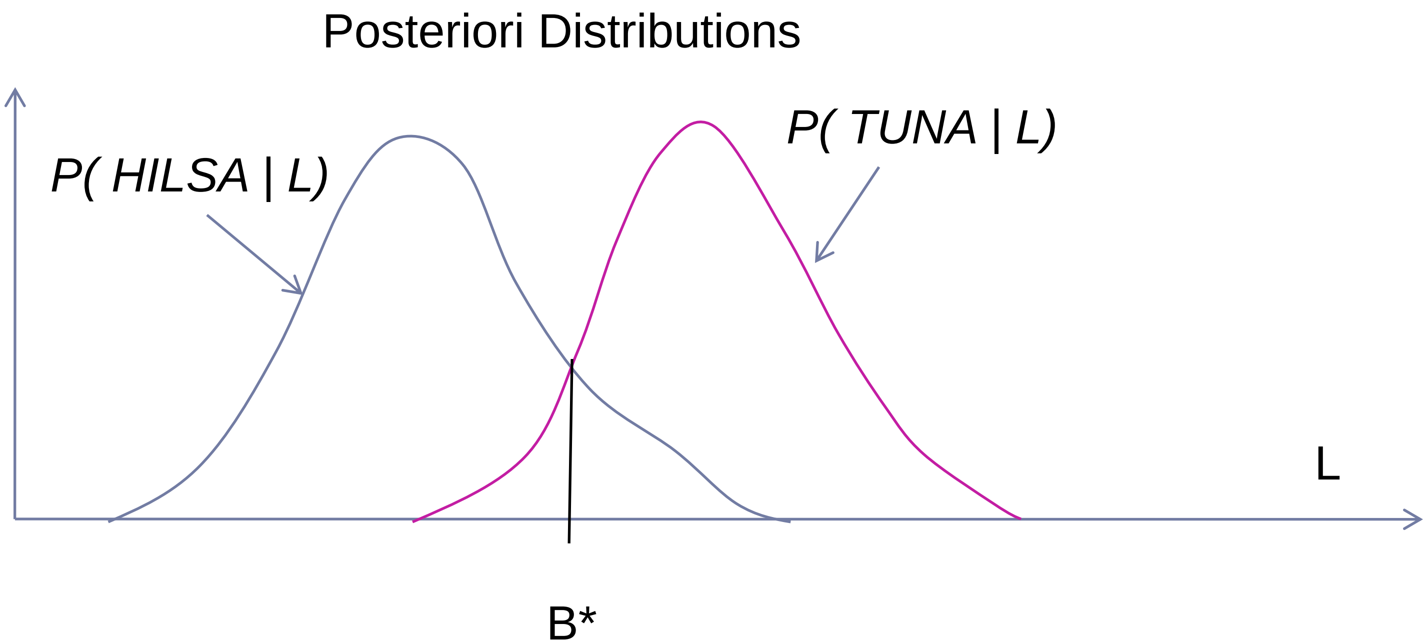
Apriori Probability

- Without measuring length what can we guess about the class of a fish
  - Depends on location/time of experiment
    - Calcutta : Hilsa, California: Tuna
- Apriori probability: *P(HILSA), P(TUNA)*
  - Property of the frequency of classes during experiment
    - Not a property of length of the fish
  - Calcutta: *P(Hilsa) = 0.90, P(Tuna) = 0.10*
  - California: *P(Tuna) = 0.95, P(Hilsa) = 0.05*
  - London: *P(Tuna) = 0.50, P(Hilsa) = 0.50*
- Also a determining factor in class decision along with class conditional probability

Classification Decision

- We consider the product of *Apriori* and *Class conditional* probability factors
- *Posteriori probability (Bayes rule)*
  - *P(HILSA | L = 2ft) = P(HILSA) x P(L=2ft | HILSA) / P(L=2ft)*
  - *Posteriori ≈ Apriori x Class conditional*
  - *denominator is constant for all classes*
- Apriori: Without any measurement - based on just location/time – what can we guess about class membership (estimated frm size of class populations)
- Class conditional: Given the fish belongs to a particular class what is the probability that its length is *L=2ft* (estimated from population)
- Posteriori: Given the measurement that the length of the fish is *L=2ft* what is the probability that the fish belongs to a particular class (obtained using Bayes rule from above two probabilities).
  - Useful in decision making using evidences/measurements.

Bayes Classification Rule (Bayes Classifier)

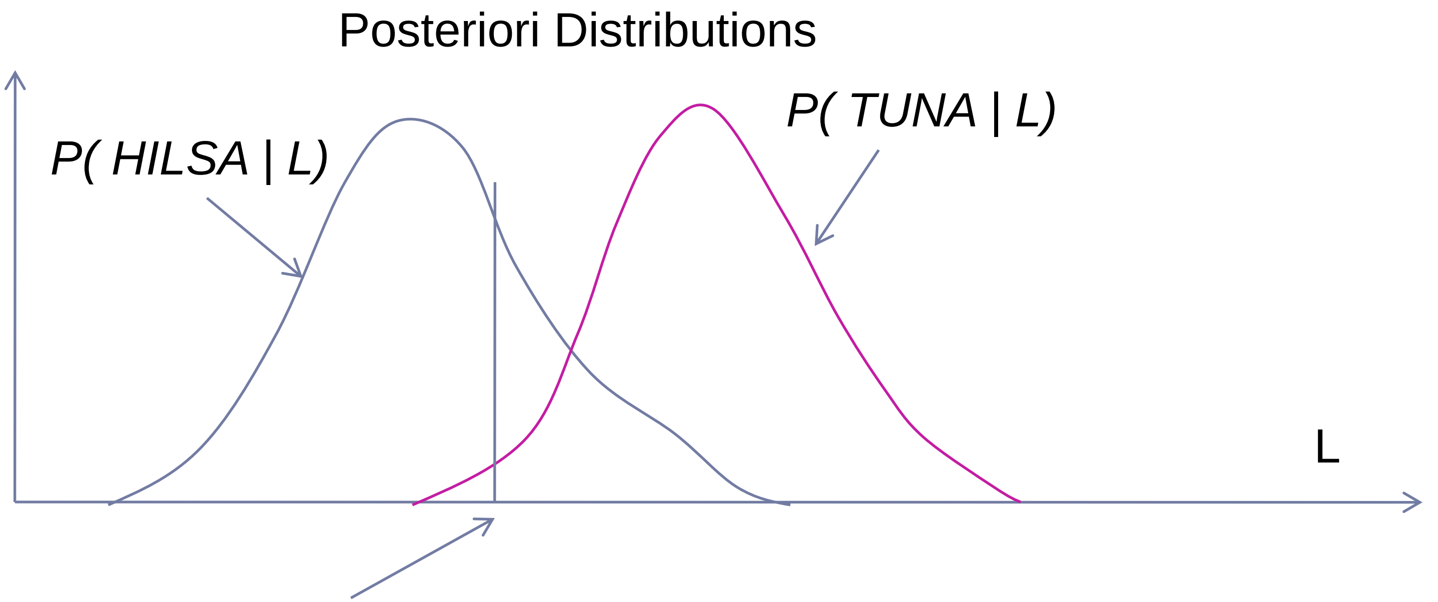


B*: Optimal Value of B, (Bayes Decision Boundary)

$$\boxed{P\,(\,HILSA \,|\, L = B^*\,) = P\,(\,TUNA \,|\, L = B^*)}$$

Minimum error probability: Bayes error

MAP Representation of Bayes Classifier

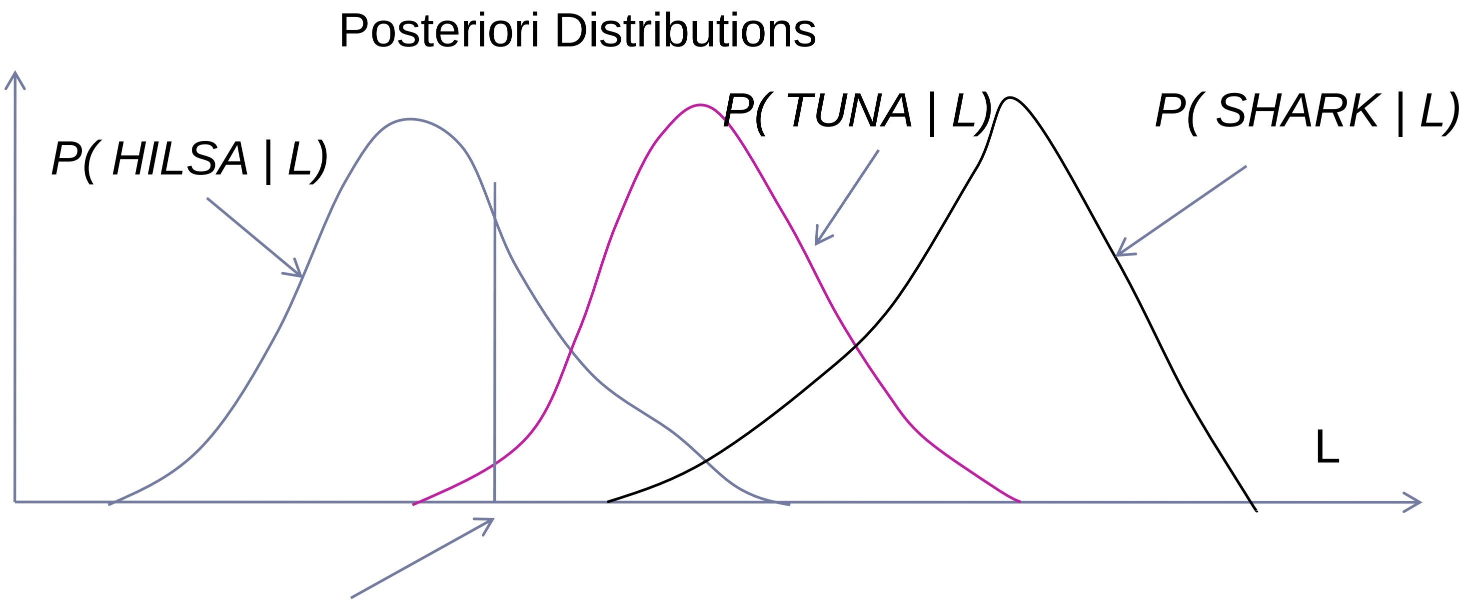


Hilsa has higher posteriori probability than Tuna for this length

Instead of finding decision boundary B*, state classification rule as:

Classify an object in to the class for which it has the highest posteriori prob.
(MAP: Maximum Aposteriori Probability)

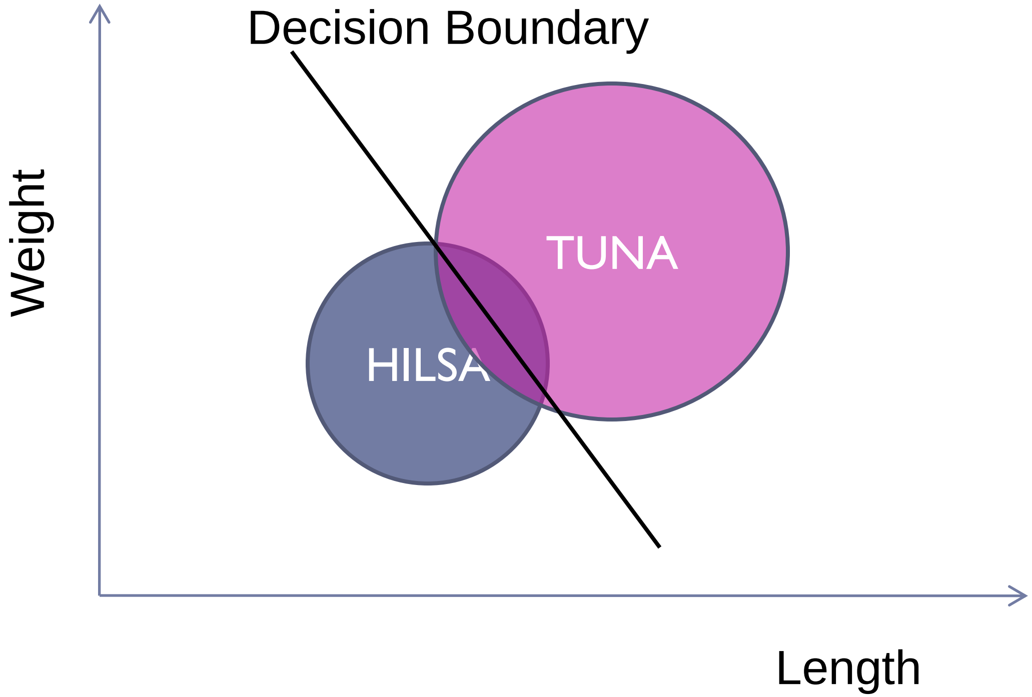
MAP Multiclass Classifier



Hilsa has highest posteriori probability among all classes for this length

Classify an object in to the class for which it has the highest posteriori prob.
(MAP: Maximum Aposteriori Probability)

Multivariate Bayes Classifier

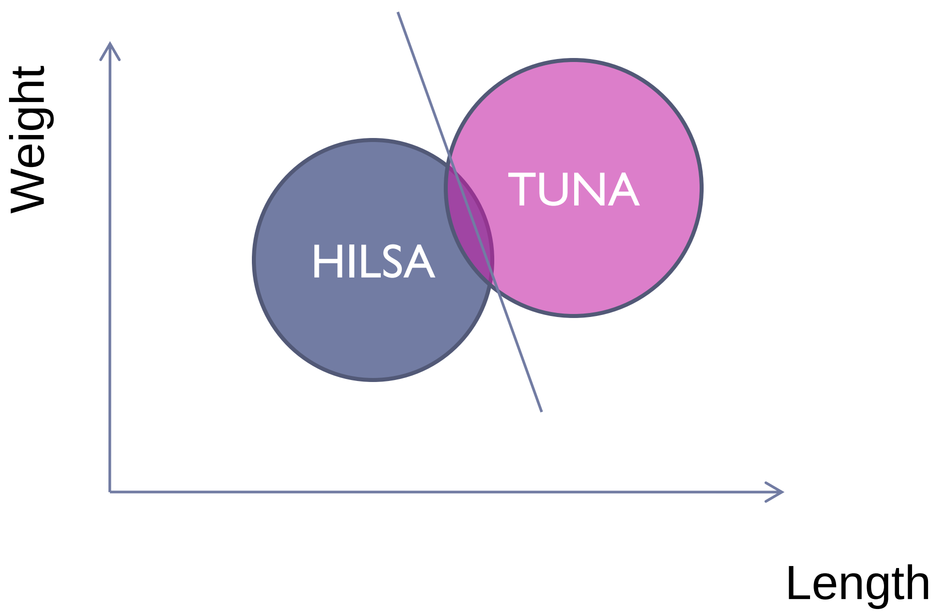


•Feature or Attribute Space

•Class Seperability

Decision Boundary: Normal Distribution

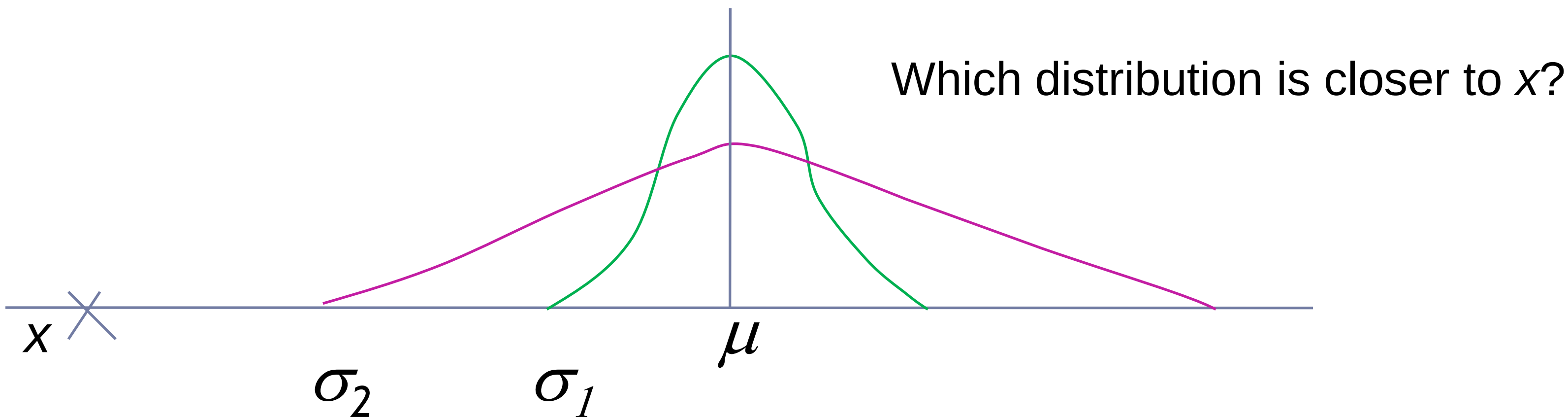
- Two spherical classes having different means, but same variance (diagonal covariance matrix with same variances)



Decision Boundary: Perpendicular bisector of the mean vectors

Distances

- Two vectors: Euclidean, Minkowski etc
- A vector and a distribution: Mahalanobis, Bhattacharya

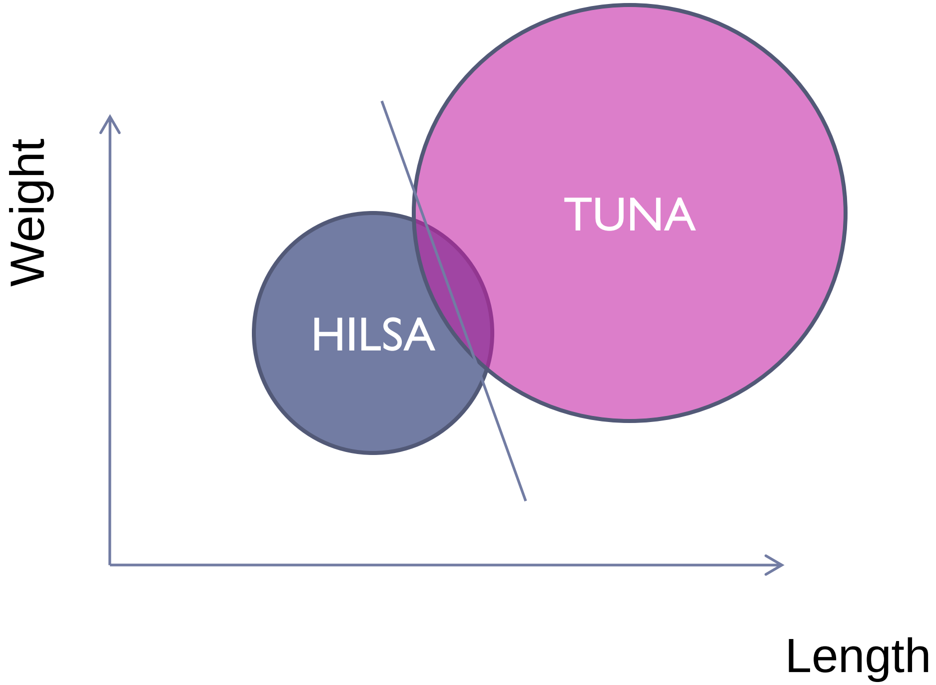


$$d_M = \frac{(x-\mu)^2}{\sigma}, d_M = (X-\mu)\Sigma^{-1}(X-\mu)^T$$

- Between two distributions: Kullback-Liebler Divergence

Decision Boundary: Normal Distribution

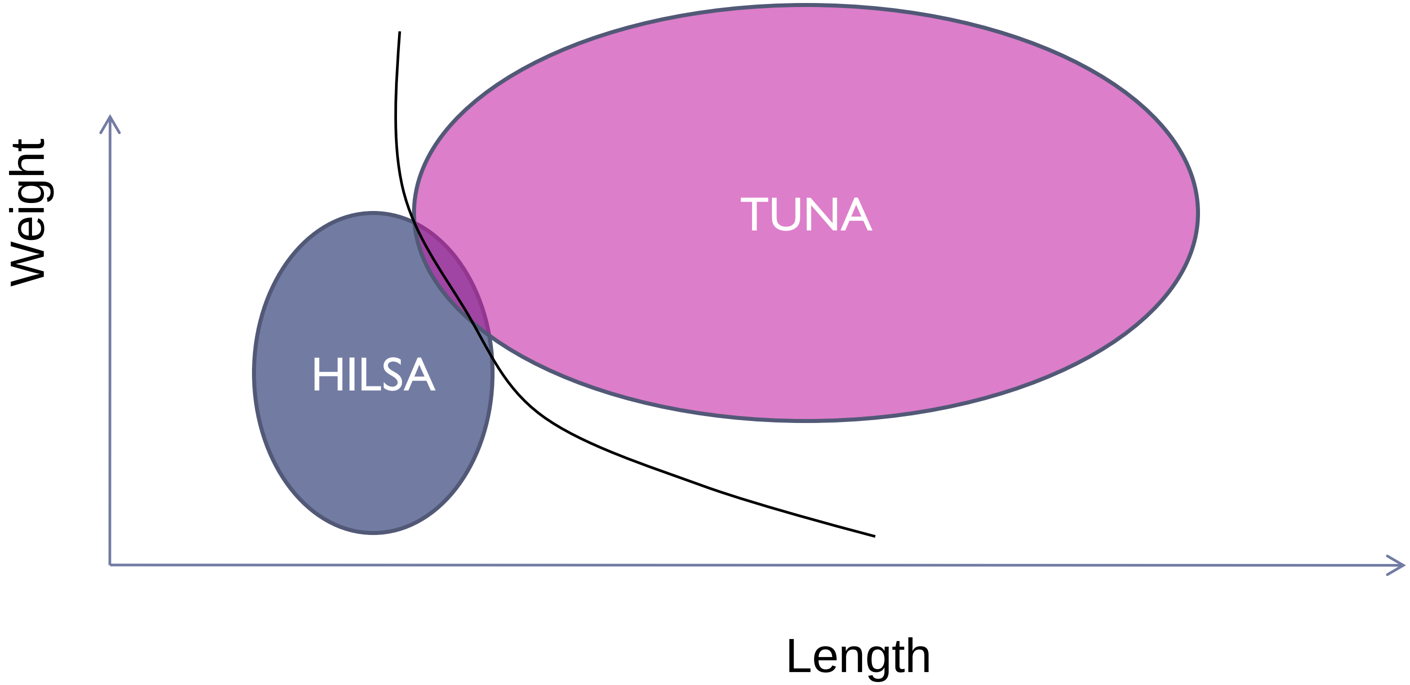
- Two spherical classes having different means and variances (diagonal covariance matrix with different variances)



Boundary: Locus of equi-Mahalanobis distance points from the class distributions. (still a straight line)

Decision Boundary: Normal Distribution

- Two elliptical classes having different means and variances (general covariance matrix with different variances)



Class Boundary: Parabolic

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$
- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Estimating Multivariate Class Distributions

- Sample size requirement
  - In a small sample: difficult to find a Hilsa fish whose length is 1.5ft <u>and</u> weight is 2 kilos, as compared to that of just finding a fish whose length is 1.5ft
  - $P(L=1.5, W=2 \mid Hilsa)$, $P(L=1.5 \mid Hilsa)$
  - Curse of dimensionality
- Independence Assumption
  - Assume length and weight are independent
  - $P(L=1.5, W=2 \mid Hilsa) = P(L=1.5 \mid Hilsa) \times P(W=2 \mid Hilsa)$
  - Joint distribution = product of marginal distributions
  - Marginals are easier to estimate from a small sample

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j) P(A_2 | C_j) \ldots P(A_n | C_j)$
  - Can estimate $P(A_i | C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \Pi P(A_i | C_j)$ is maximal.

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption

k

  - Two-way split: $(A < v)$ or $(A > v)$
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|c)$

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

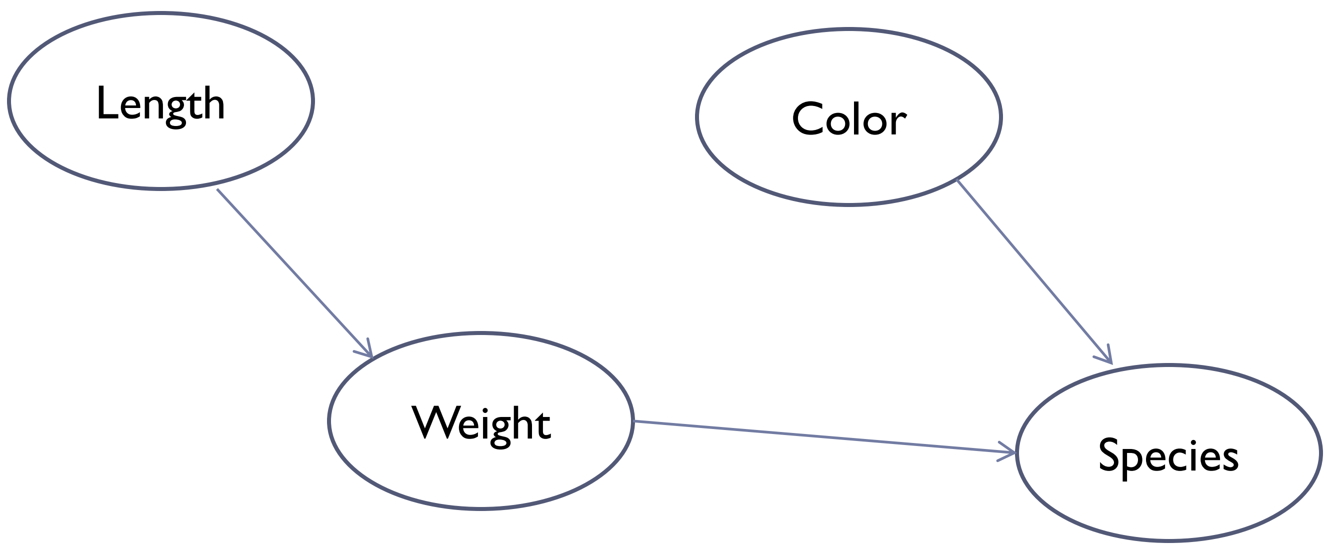
m: parameter

Bayes Classifier (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Length and weight of a fish are not independent

Bayesian Belief Network

- A directed acyclic probablistic graphical model that captures dependence among the attributes



Nodes: Variable/Attributes/Class
Directed edges: Causality
Absence of edge: independence

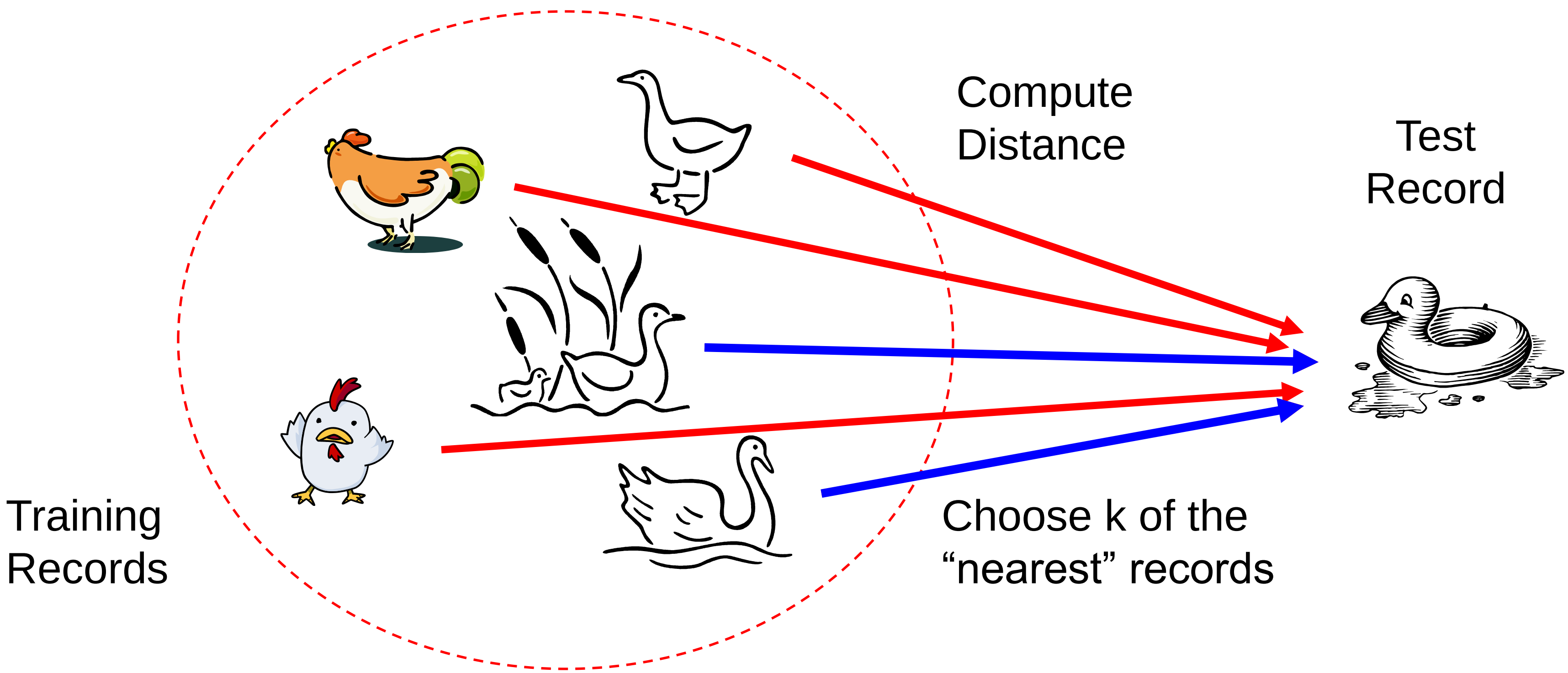
Network structure: domain knowledge
Joint probabilities: from data

Nonparametric Statistics

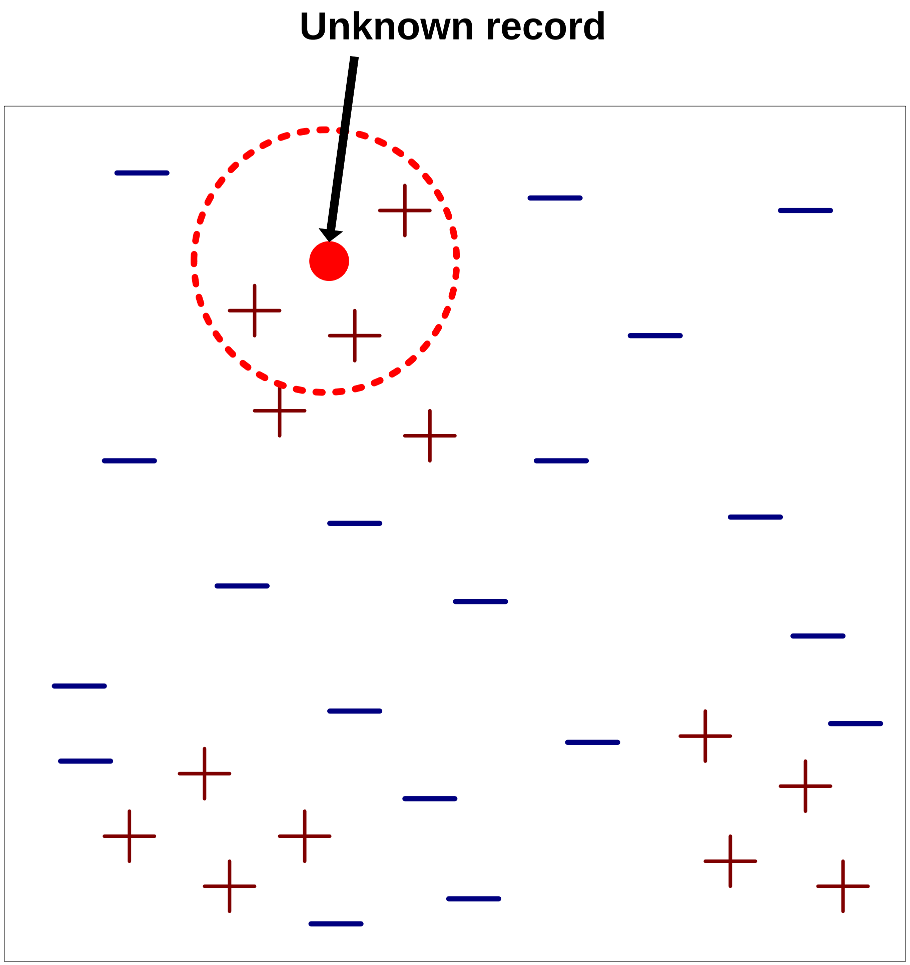
- Do not assume parametric data distribution/model
- Take decisions based on given sample
- Bayesian statistics vs frequentist statistics

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



Nearest-Neighbor Classifiers



- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Nearest Neighbor Classification

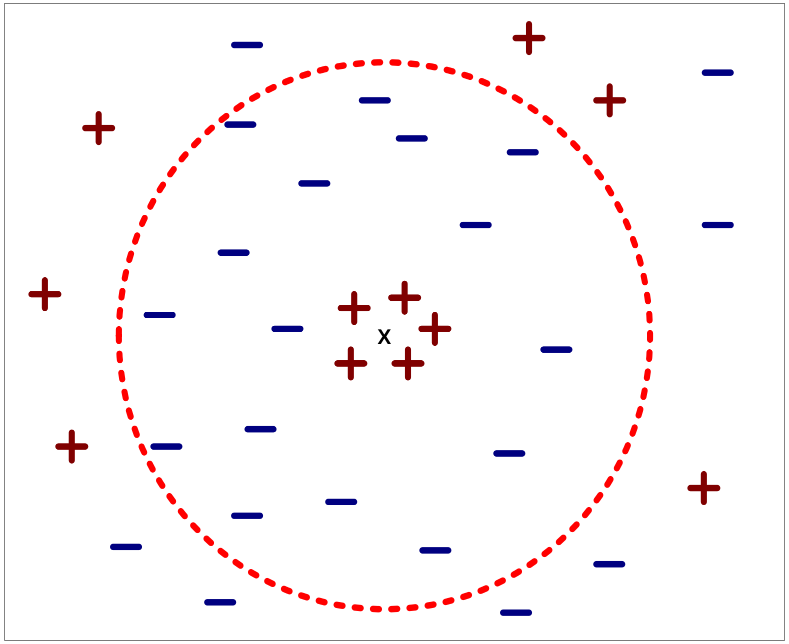
- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

Nearest Neighbor Classification…

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



Nearest neighbor Classification…

- k-NN classifiers are lazy learners
  - It does not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

DNA Coding Segment Identification

- Classes: Coding – noncoding segment
- Attributes/features: sequence information
- Complex interdependence among attributes

Microarray Data Analysis

- Classes: Disease classes
- Attributes/features: gene expression levels
- Large number attributes, fewer samples

Protein Secondary Structure Prediction

- Classes: α-helix, coil etc
- Attributes/features: length, amino acid sequence, hydrophobicty, shape, ions
- Complex class distributions

Protein Interaction Prediction

- Classes: Binary
- Attributes/features: protein properties
- Presence of domain knowledge

References

- Pattern Classification, Duda, Hart and Stork, Wiley, 2010
- Slides on data mining by Vipin Kumar

Questions!