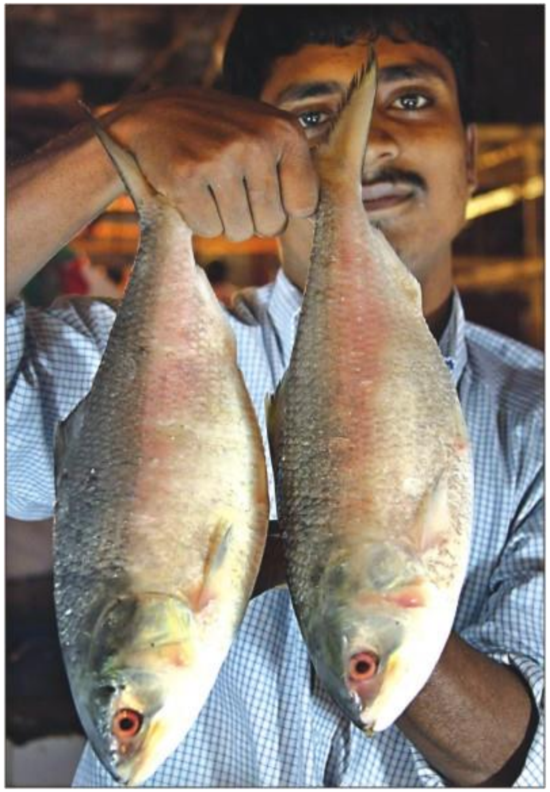
Bayesian Classification

A Simple Species Classification Problem

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

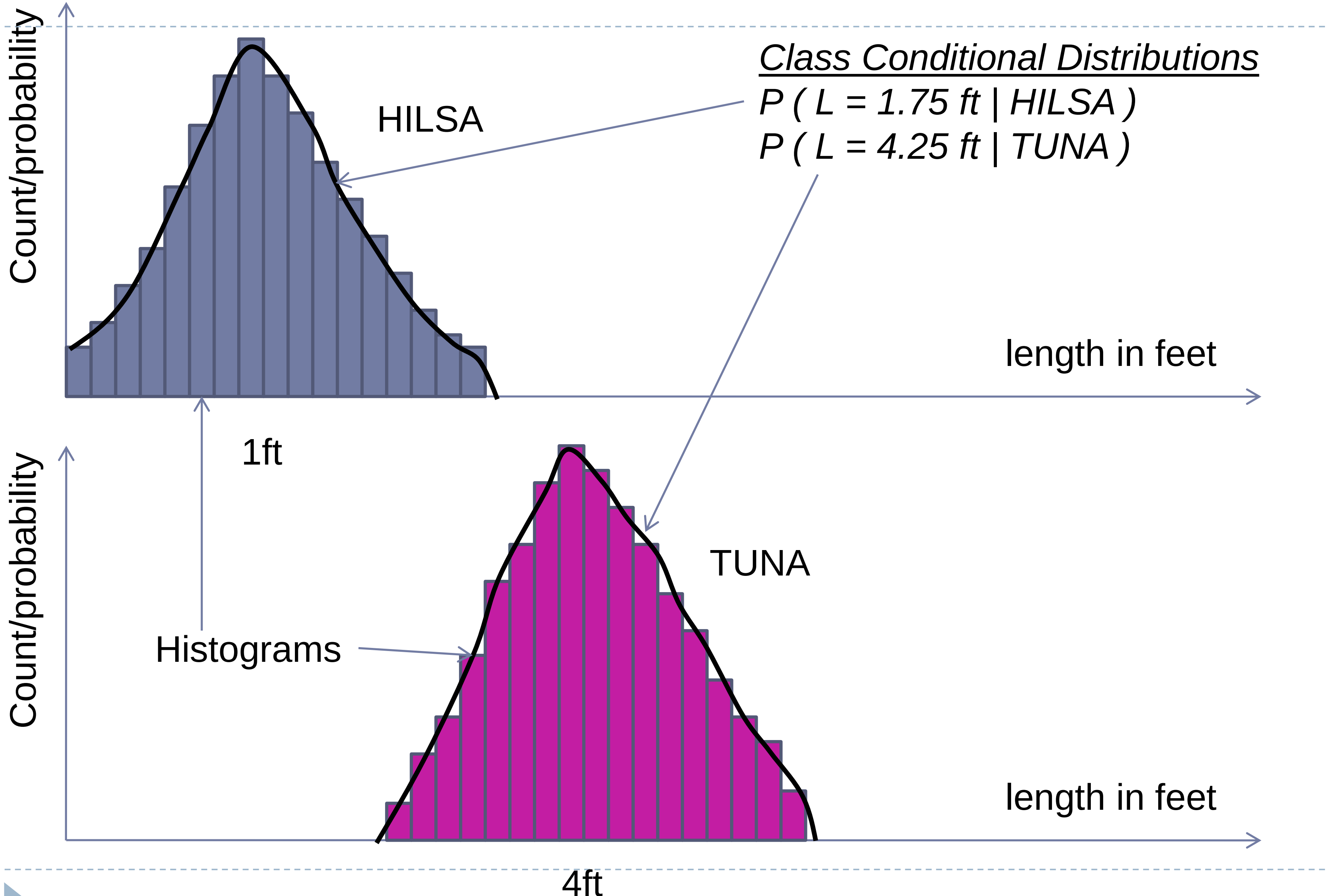


Collect Statistics …

Population for Class Hilsa

Population for Class Tuna

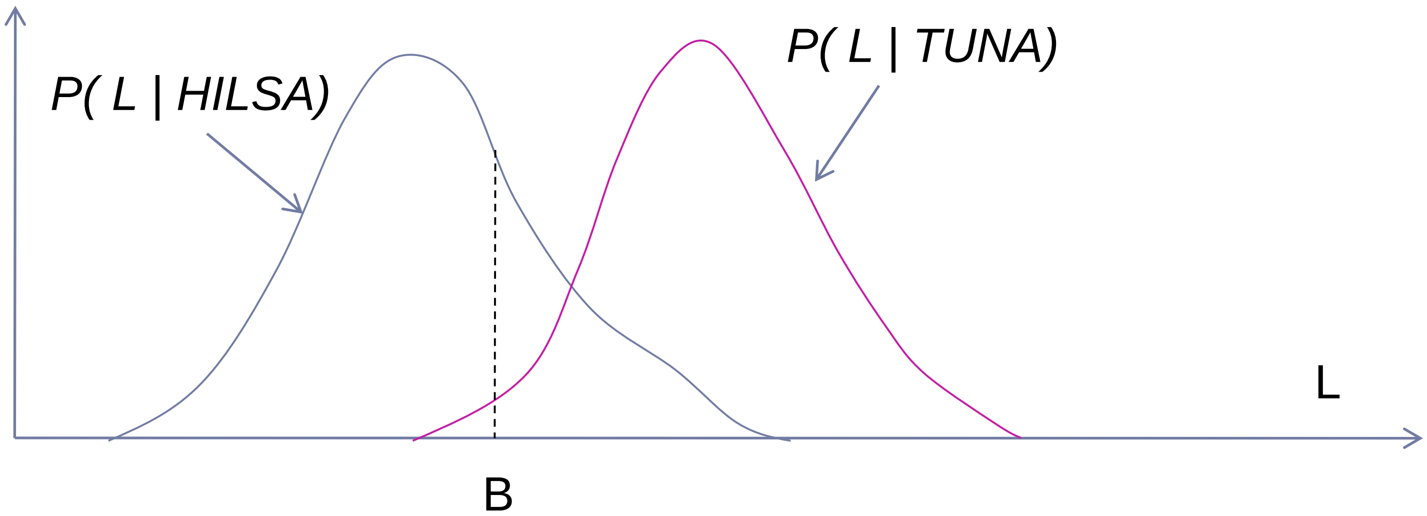
Distribution of “Fish Length”



Decision Rule

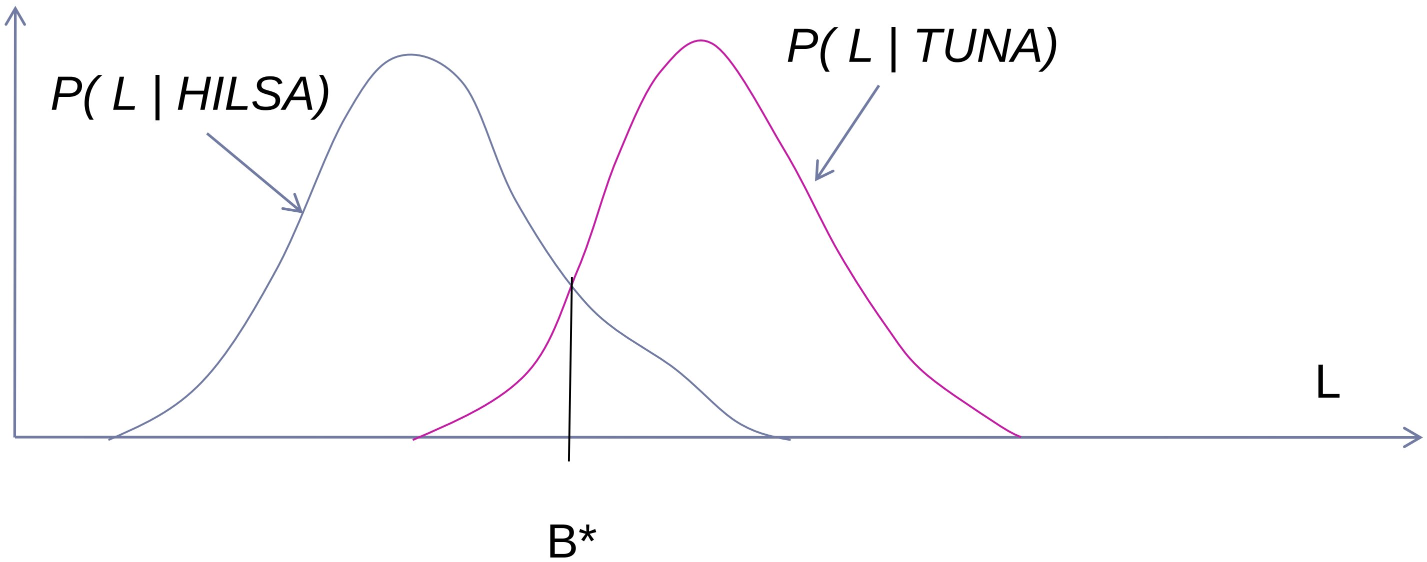
- If length $L \leq B$
  - HILSA
- ELSE
  - TUNA
- What should be the value of B (“boundary” length) ?
  - Based on population statistics

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

$$\boxed{P (B^* \mid HILSA) = P (B^* \mid TUNA)}$$

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

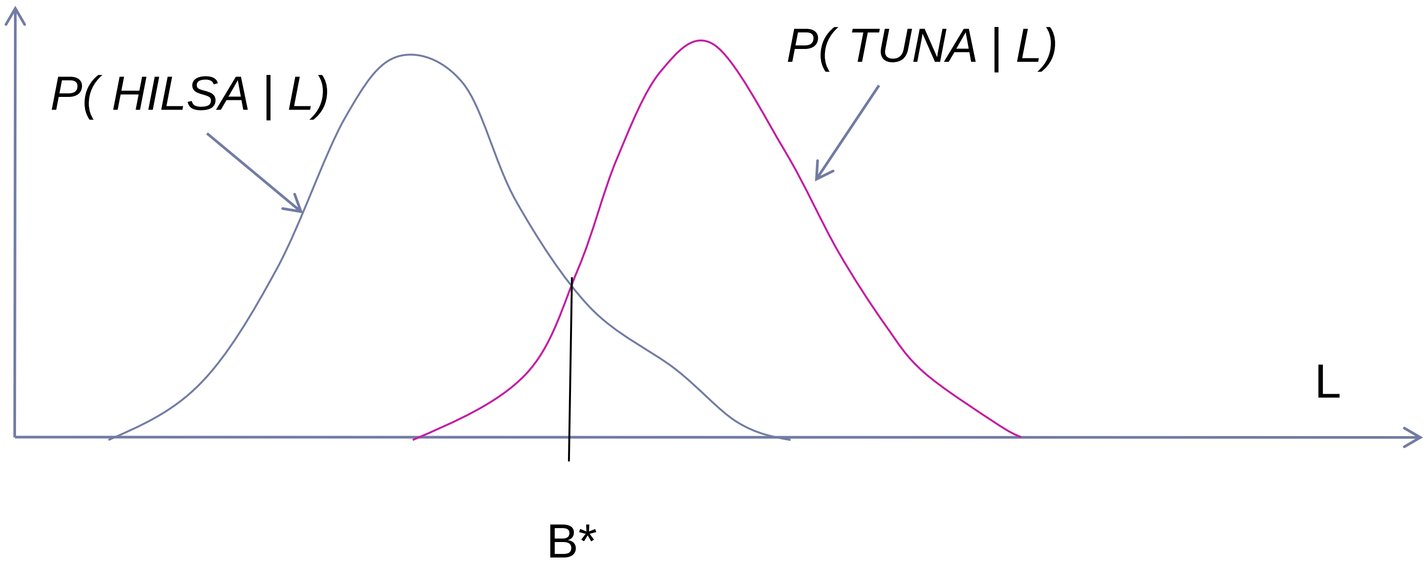
- Without measuring length what can we guess about the class of a fish
  - Depends on location/time of experiment
    - Calcutta : Hilsa, California: Tuna
- Apriori probability: *P(HILSA), P(TUNA)*
  - Property of the frequency of classes during experiment
    - Not a property of length of the fish
  - 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)

Posteriori Distributions

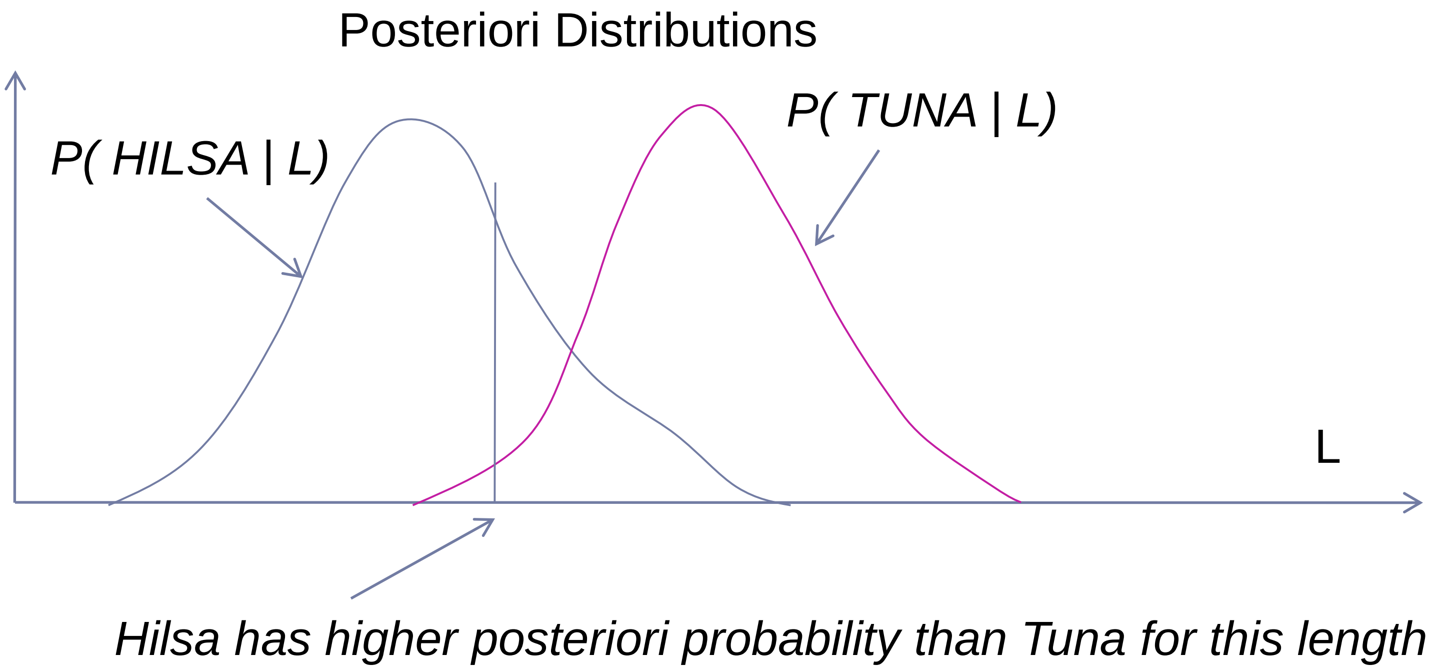


B*: Optimal Value of B, (Bayes Decision Boundary)

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

Minimum error probability: Bayes error

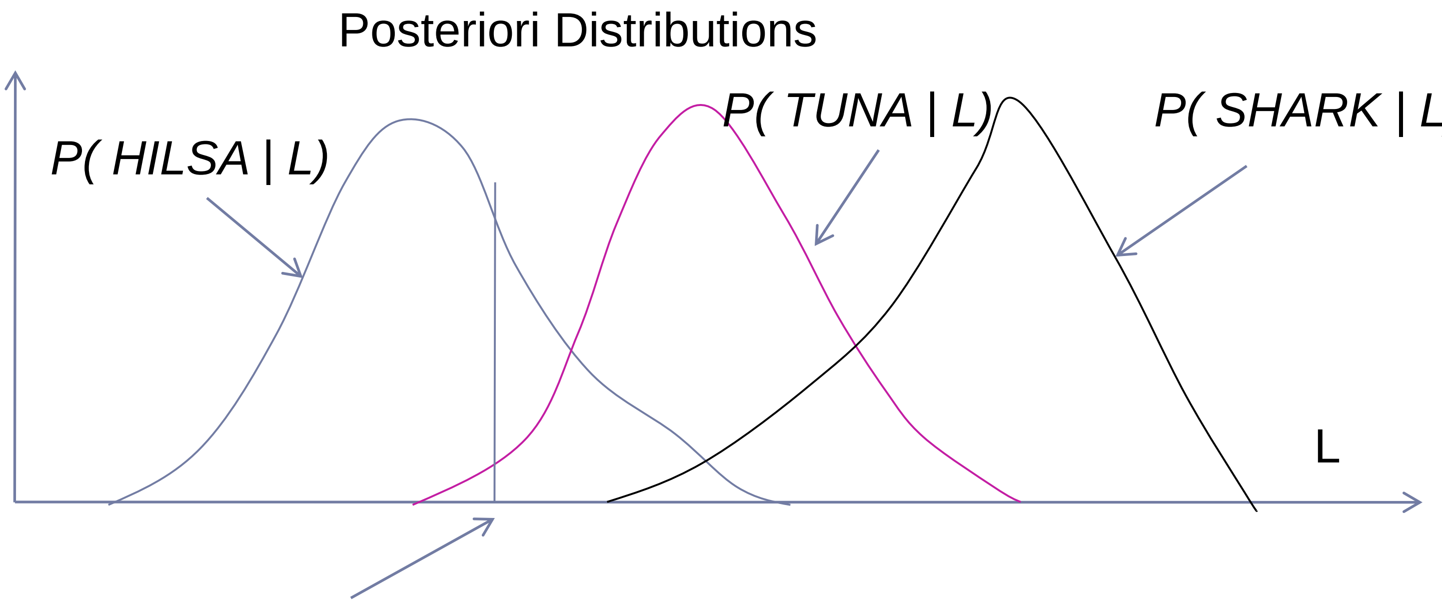
MAP Representation of Bayes Classifier



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 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 $P(A_1, A_2, \ldots, A_n \mid C)$?

Example of 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

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

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 $P(C_j) \prod P(A_i | C_j)$ 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

Naïve Bayes Classifier

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

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

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

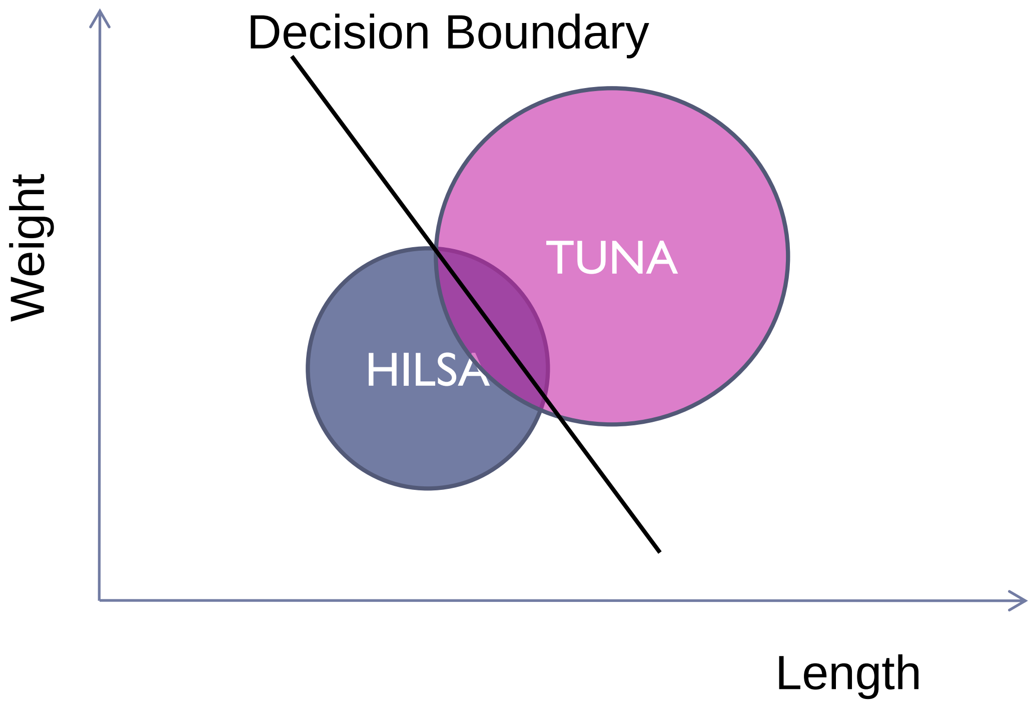
$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Multivariate Gaussian Bayes Classifier

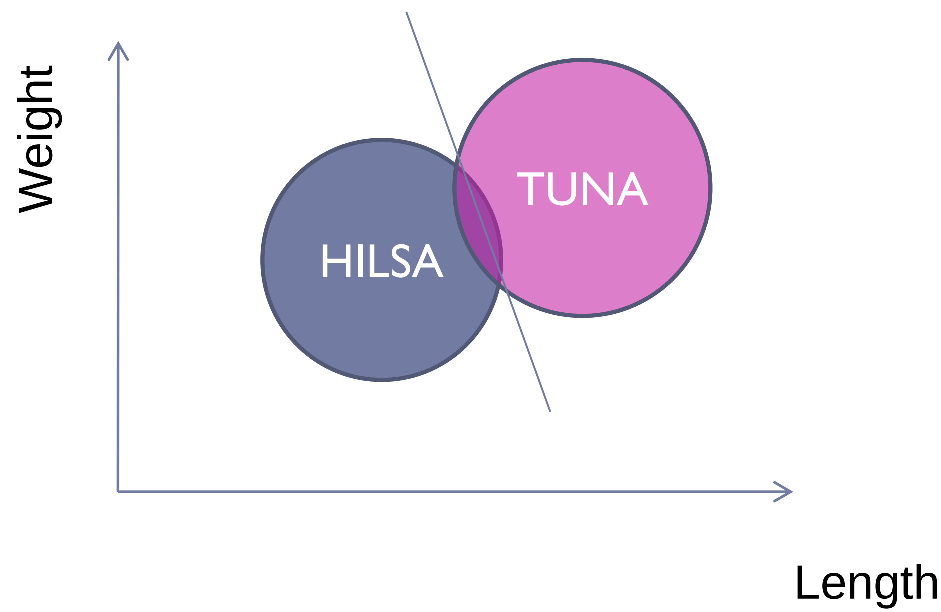


•Feature or Attribute Space

•Class Seperability

Decision Boundary: Normal Distribution

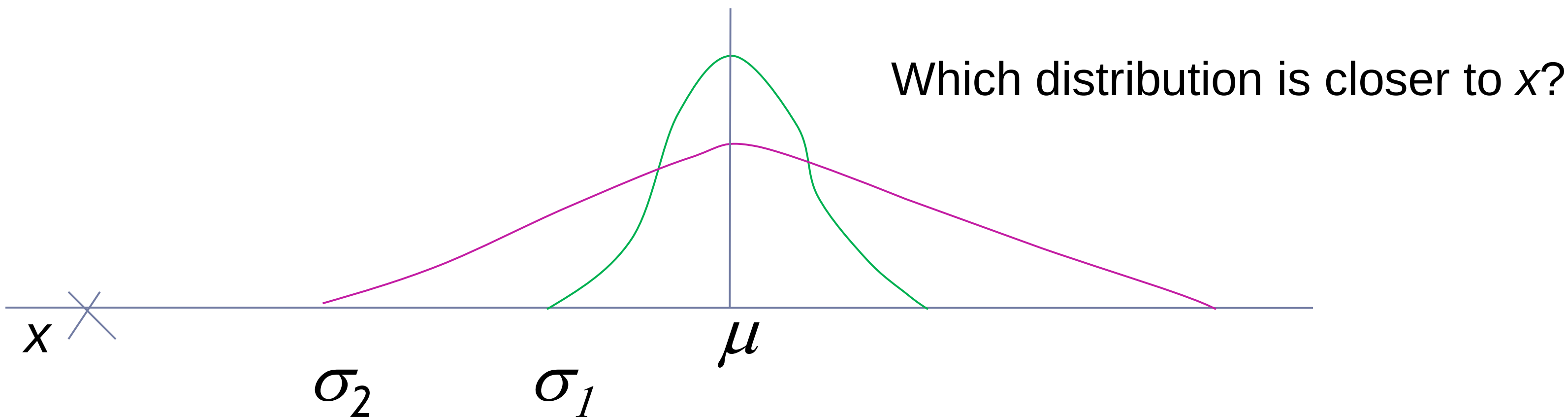
- Two spherical classes having different means, but same variance (diagonal covariance matrix with same variances)



Decision Boundary: Perpendicular bisector of the mean vectors

Distances

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

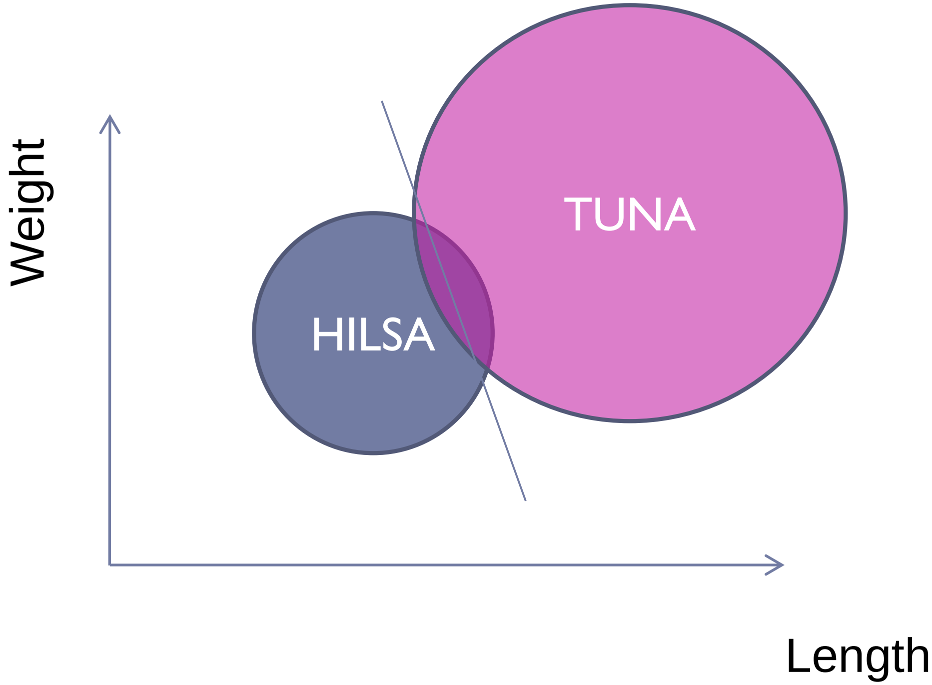


$$d_M = \frac{(x-\mu)^2}{\sigma}, d_M = (X-\mu)\Sigma^{-1}(X-\mu)^T$$

- Between two distributions: Kullback-Liebler Divergence

Decision Boundary: Normal Distribution

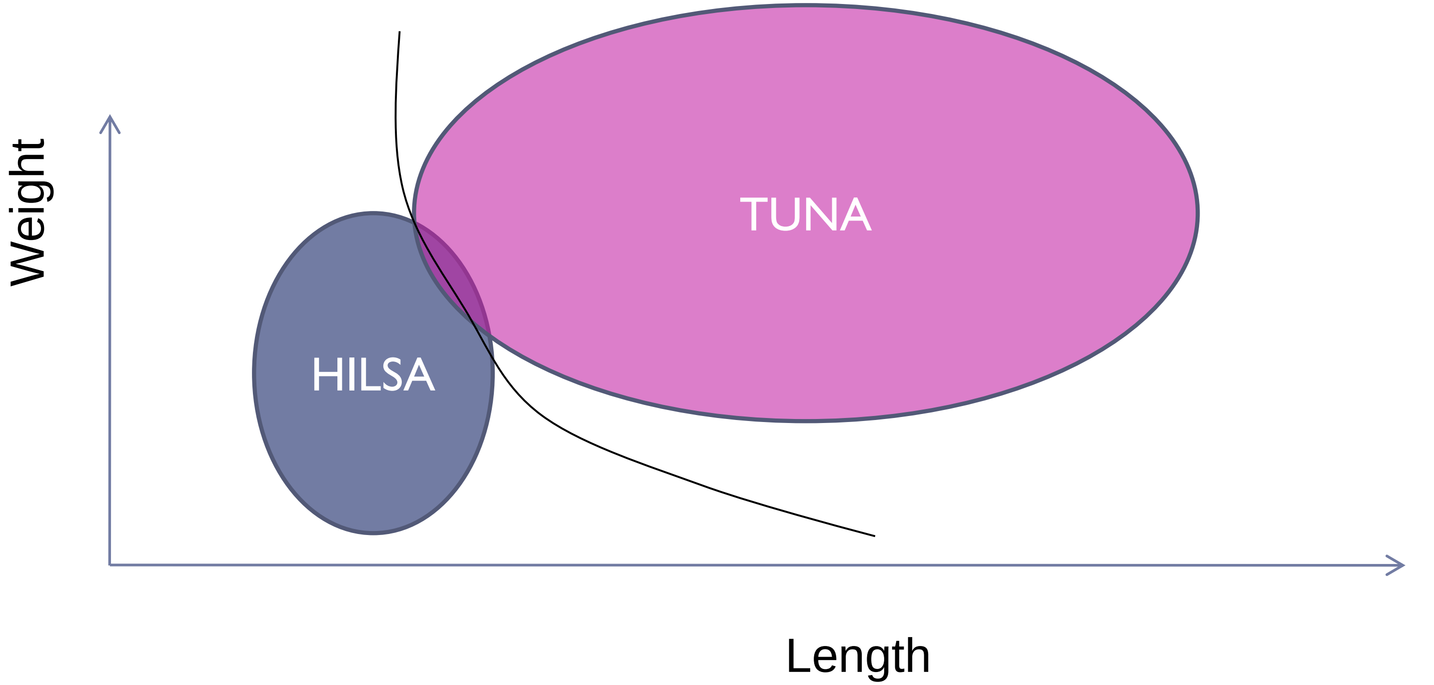
- Two spherical classes having different means and variances (diagonal covariance matrix with different variances)



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)



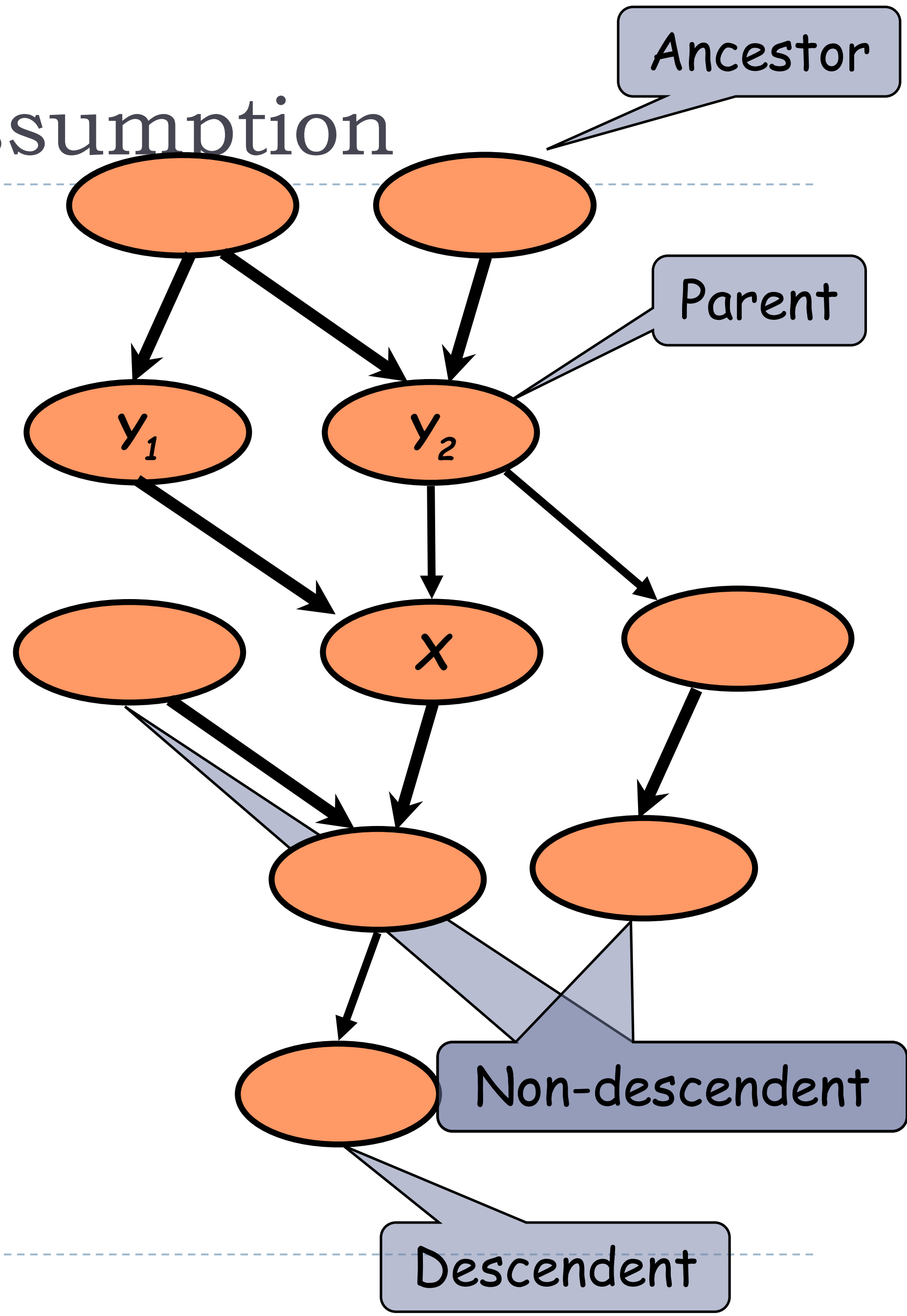
Class Boundary: Parabolic

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

BayesNet: Markov Assumption

- We now make this independence assumption more precise for **directed acyclic graphs** (DAGs)
- Each random variable X, is independent of its non-descendents, given its parents Pa(X)
- Formally,
  I (X, NonDesc(X) | Pa(X))



Why Evaluate ML Models?

- Is the model good enough for use?
- What is the best hyper-parameter value?
- How do we compare various models?

ML Evaluation Measures

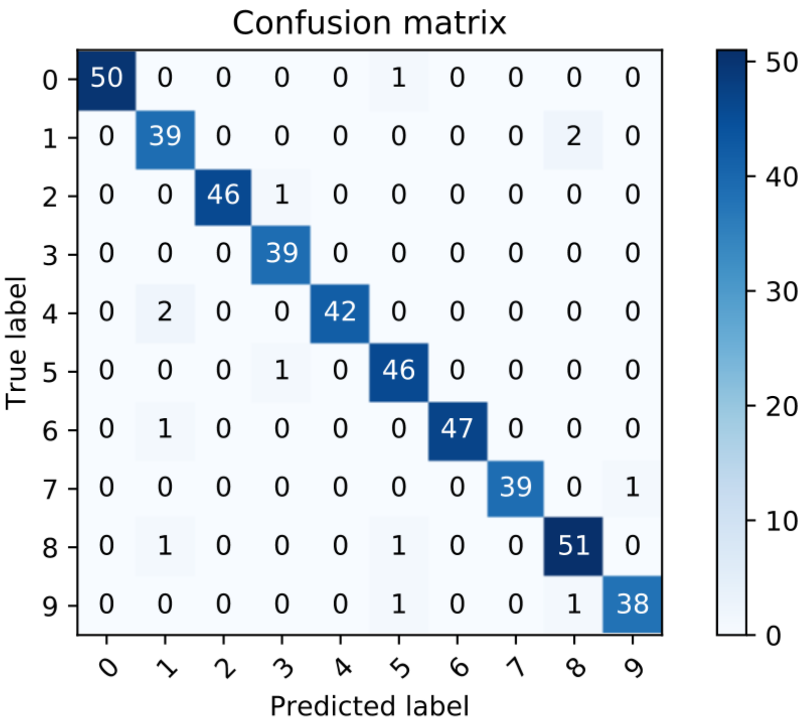
CSEDU-ML

- <u>Classification</u>
  - Confusion matrix
  - Precision, Recall, F-Score
  - AUROC
- Regression
  - Mean squared error, RMSE
  - Mean absolute error
- Unsupervised clustering
  - Silhouette coefficient
  - Davis-Bouldin index

- Application Independent Measures
- Application Dependent Measures

Classifier Evaluation: Confusion Matrix

Digit Recognition



Spam Filter!

	Predicted	
Actual	Inbox	Spam
Inbox	25	0
Spam	5	60

Robot 1

	Predicted	
Actual	Inbox	Spam
Inbox	20	5
Spam	0	65

Robot 2

COVID Test!

	Predicted	
Actual	Negative	Positive
Negative	25	0
Positive	5	60

Robot 1

	Predicted	
Actual	Negative	Positive
Negative	20	5
Positive	0	65

Robot 2

Only Accuracy not Enough!

- Unequal cost of decision
  - *Medical diagnosis*
  - *Spam Filtering*

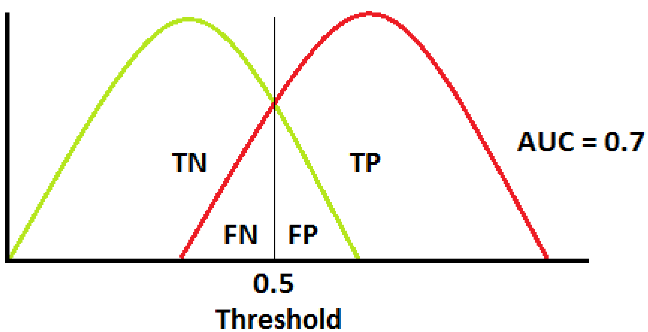
- Unbalanced Classes
  - *Medical diagnosis: 95 % healthy, 5% disease.*
  - *e-Commerce: 99 % do not buy, 1 % buy*

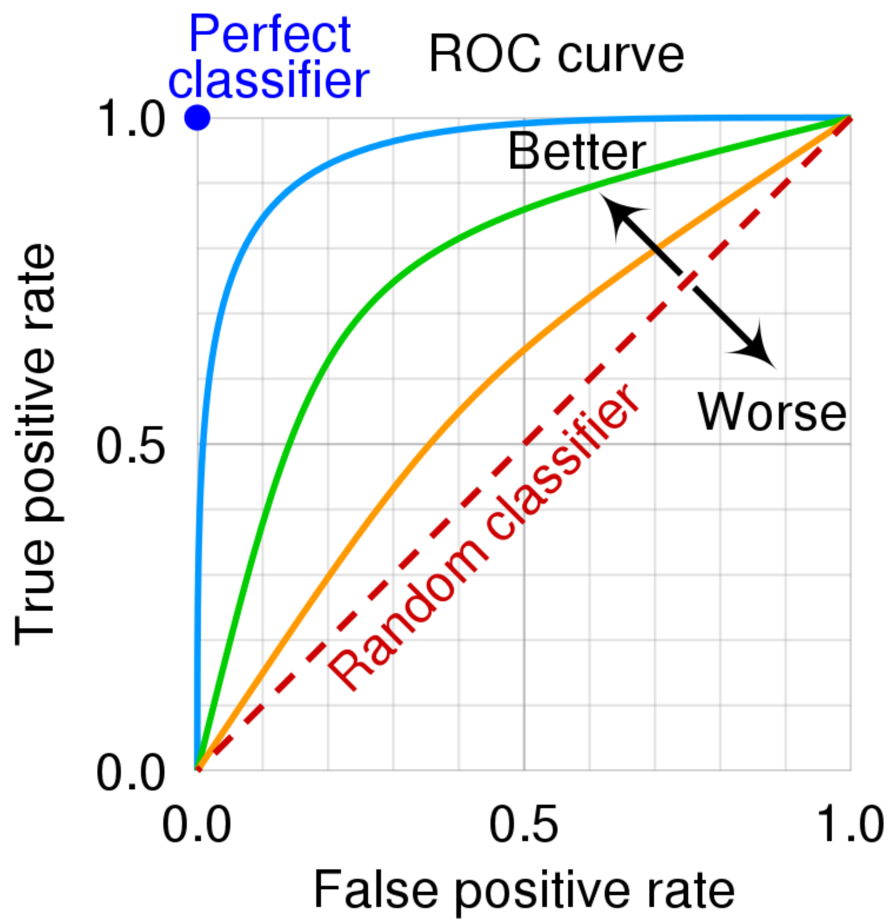
Multiple Scores

actual examples		predicted: negative	predicted: positive
	negative	*a* TN - True Negative correct rejections	*b* FP - False Positive false alarms type I error
	positive	*c* FN - False Negative misses, type II error overlooked danger	*d* TP - True Positive hits

- Accuracy $= (a+d)/(a+b+c+d) = (TN + TP)/total$
- **True positive rate**, recall, sensitivity $= d/(c+d) = TP/\textit{actual positive}$
- Specificity, true negative rate $= a/(a+b) = TN/\textit{actual negative}$
- Precision, predicted positive value $= d/(b+d) = TP/\textit{predicted positive}$
- **False positive rate**, false alarm $= b/(a+b) = FP/\textit{actual negative} = 1 - \textit{specificity}$
- False negative rate $= c/(c+d) = FN/\textit{actual positive}$

ROC Curve


TN
TP
FN
FP
AUC = 0.7
0.5
Threshold


Perfect classifier
ROC curve
1.0
0.5
0.0
Better
Worse
Random classifier
True positive rate
0.0
0.5
1.0
False positive rate

Estimation of Generalization Performance

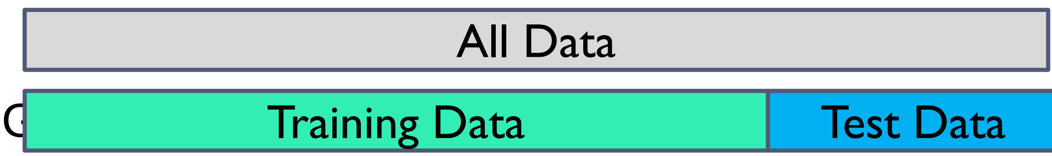
- A classifier should perform well on <u>unseen</u> examples drawn from the underlying data distribution
  - Underlying distribution unknown

- We only have a sample from the data distribution!

- How to estimate true generalization error?
  - Robust estimation using the sample

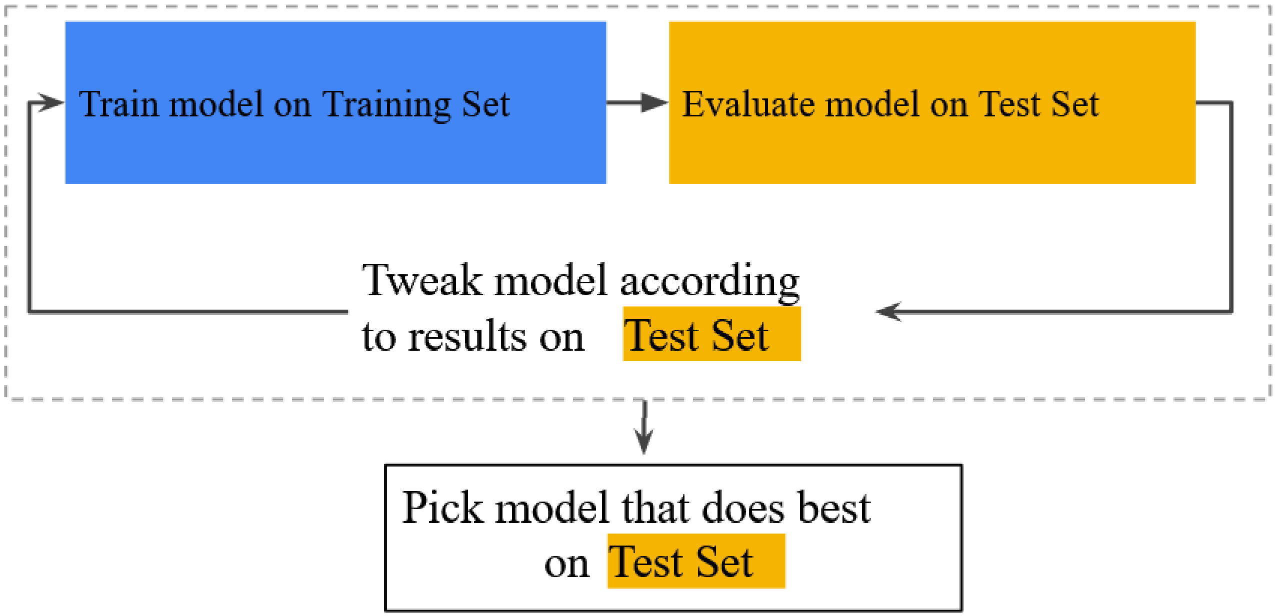
Hold-Out Set

- Randomly partition data into Train Set and Test Set



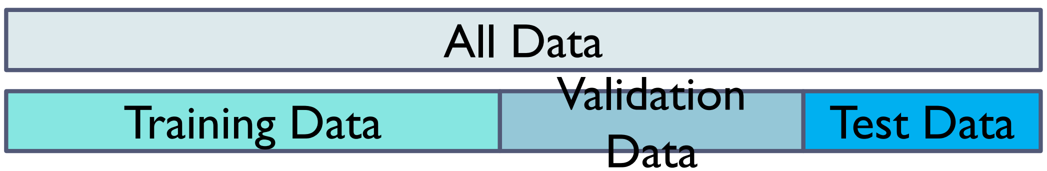
- G ator of good performance on the new data in general

Using the Test Set



Validation Set

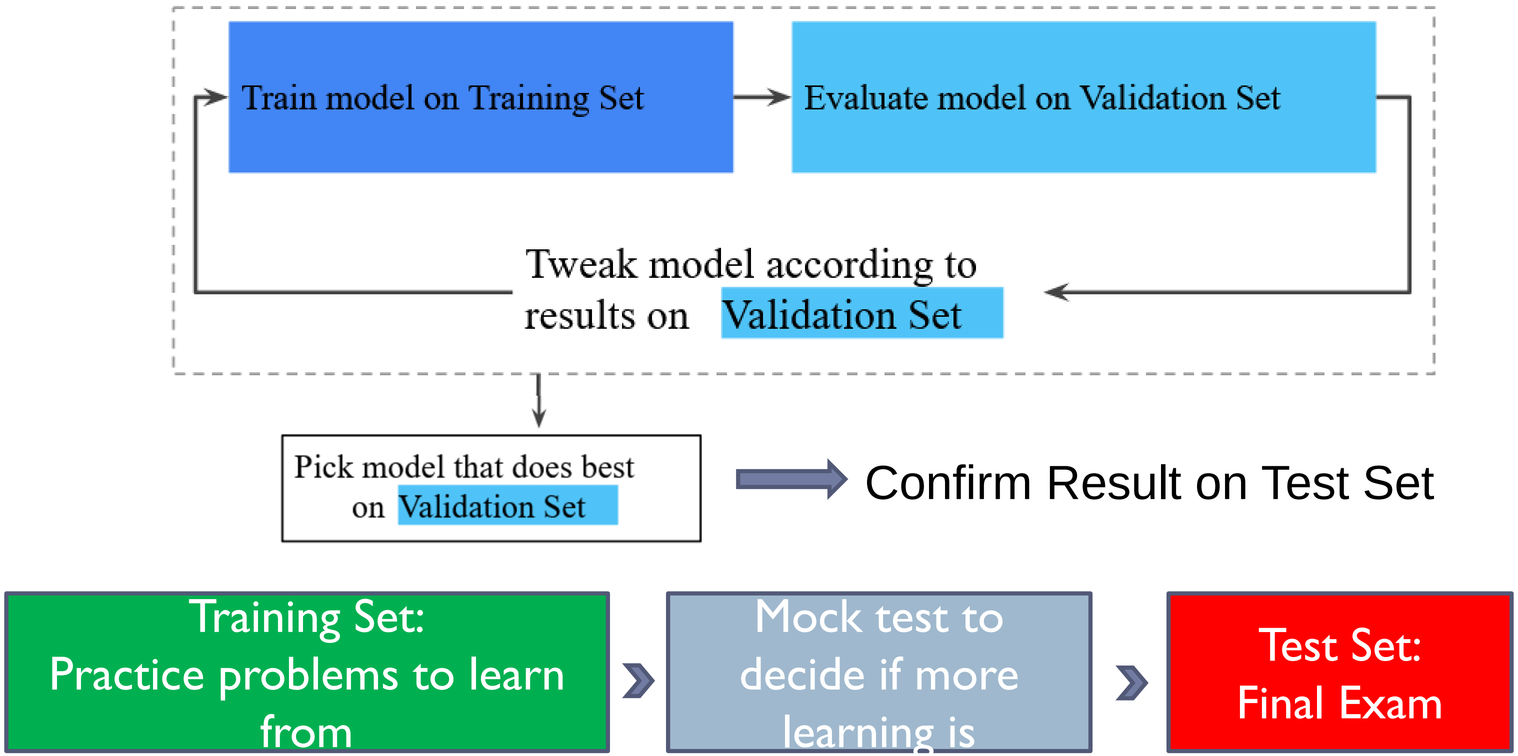
- Randomly partition data into Train, Validation, and Test Sets



- Motivation: One should never use test data during training.

Use of Validation Set

CSEDU-ML

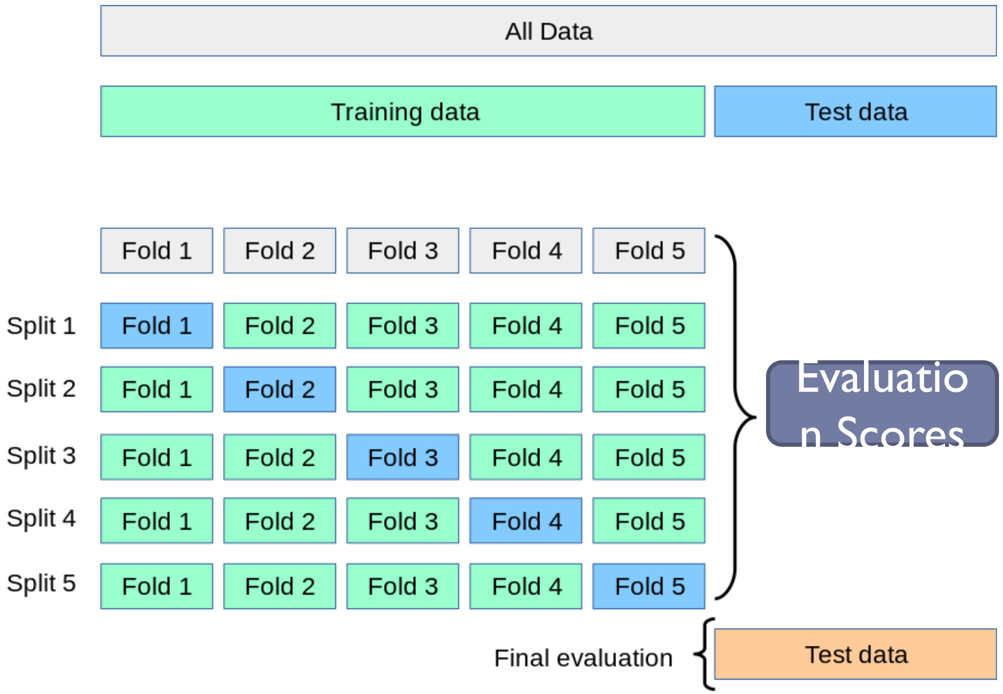


Bootstrap Estimates

Sampling with replacements
Data points may repeat

Average of scores over each bootstrap sample

K-Fold Cross Validation

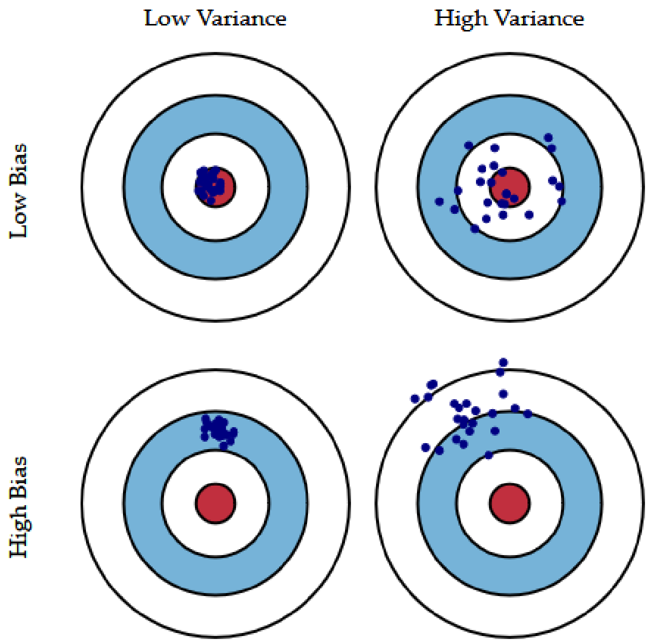


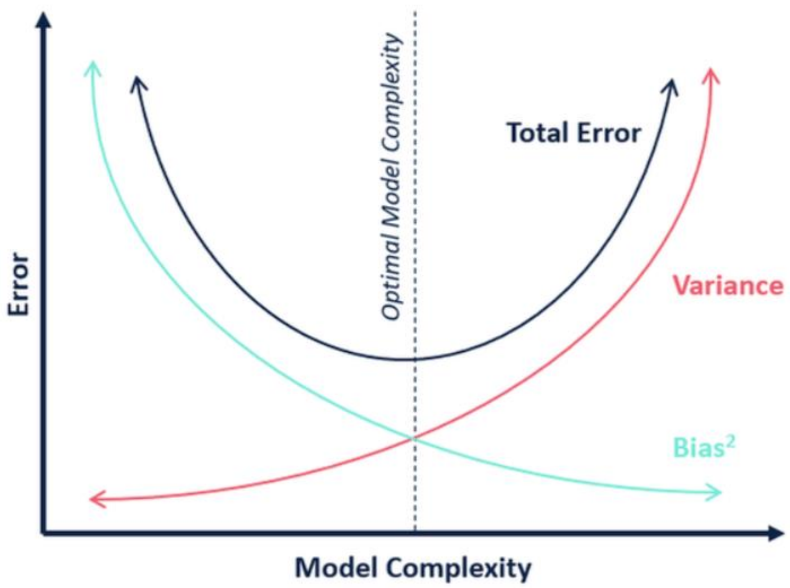
K = 5, 10

Leave-one-out: K = N,
N: size of data set

Error = Bias + Variance

CSEDU-ML

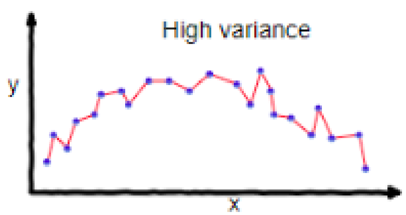




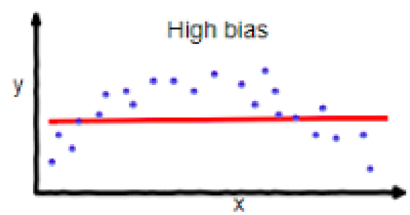
Reducing Bias-Variance Errors

CSEDU-ML

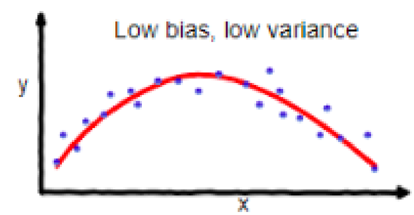
- Bias
  - Choose a more sophisticated model
- Variance
  - Regularization



overfitting



underfitting



Good balance