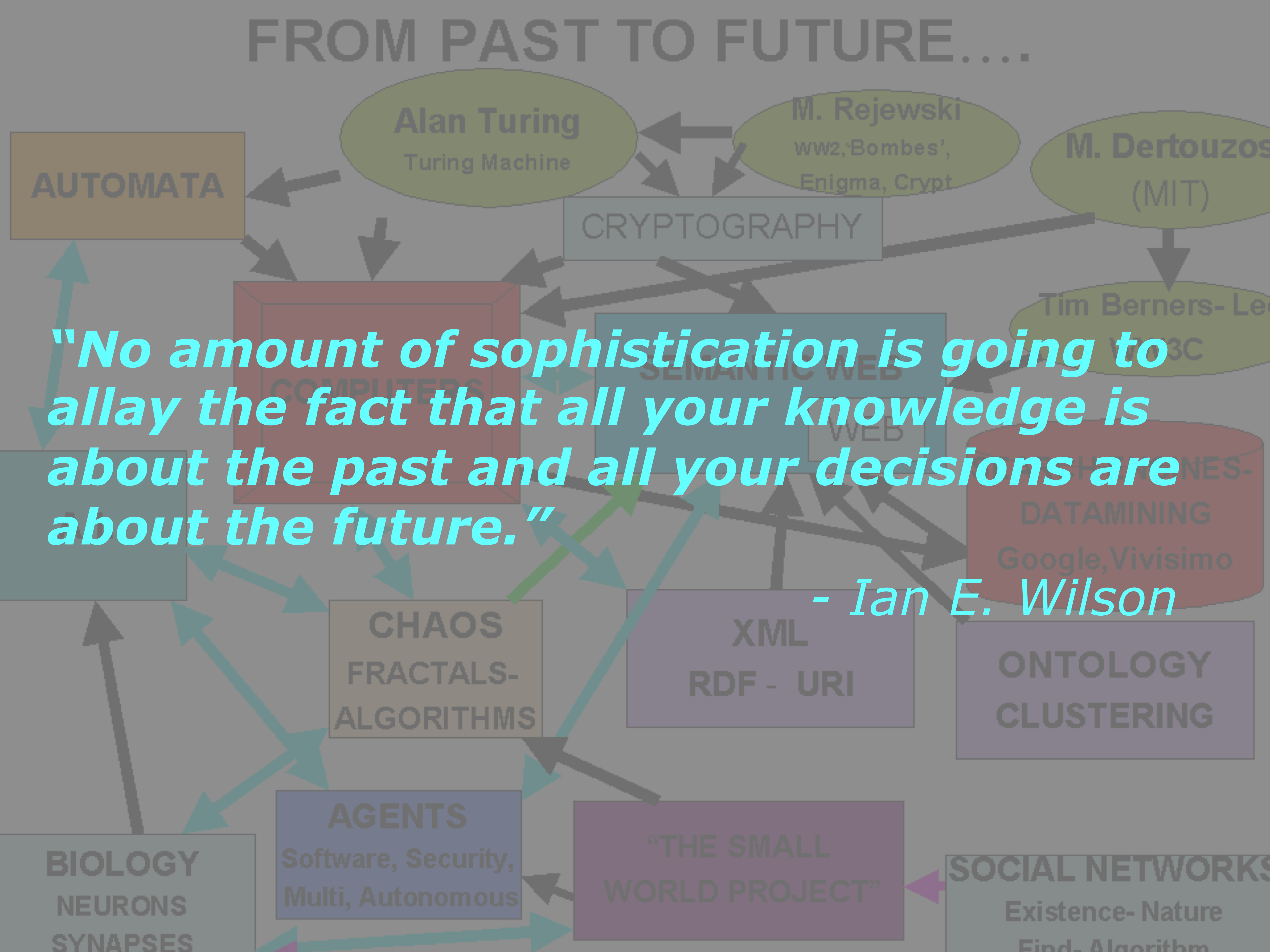
Scoring Customers for personalized offers:
Statistical Classification



"No amount of sophistication is going to allay the fact that all your knowledge is about the past and all your decisions are about the future."

- Ian E. Wilson

Classification Methods (Introduction)

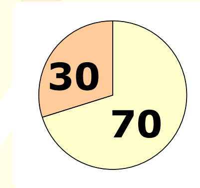
Classification is an important technique in statistical pattern recognition

- **Assigning entities or objects based on a set of unique *features* is the primary task of any classification effort.**
- **Classification models with a set of *training* data**
- ***Training* data is a set of records pertaining to an entity**
- **Each record corresponds to an individual**
- **Features are attributes of the individual like gender, marital status, location, income, credit history etc.**
- **Each record in *training data* carries an additional attribute called class label appropriate to classification problem.**
- **If the purpose of the model is to evaluate the credit risk then labels could be "High Risk", "Low Risk"**

Classification Methods (Introduction)

Classification dataset

Tom	M	New York	60000
Dick	M	Boston	90000
Mary	F	Dallas	45000
…	…	…	…



Training **dataset**

Tom	M	New York	60000	High Risk
Dick	M	Boston	90000	Low Risk
Mary	F	Dallas	45000	Low Risk
…	…	…	…	…

Class label

Validation **dataset**

Tim	M	Chicago	57000
Pam	F	Atlanta	80000
…	…	…	…

Statistical Classification

- **Statistical classification is a methodology that utilizes probabilistic ideas and concepts to derive *decision rules* also known as *classifiers* such that some optimality criterion is met.**
- **Consider the task of classifying dogs and cats into two groups. Since the features that distinguish adult dogs and cats are so unique and distinct it is a relatively simple exercise.**

But life is never that simple!

- **More often than not features that distinguish objects are not so distinct and unique. Features may overlap making perfect classification impossible.**
- **In the scenario where features belonging to different classes are not 100% unique, classification schemes are prone to error. Two types of *error* are likely**
  - **Assign an observation that truly belongs to class A to class B,**
  - **Assign an observation that truly belongs to Class B to Class A.**
- **The aim is to minimize this miss-alignment**

Assigned class

Actual class	A	B
A	✓	X
B	X	✓

Type errors

Probability based classifier methods
Fisher discriminant analysis

- **Principal among probability based methods is Fisher discriminant analysis. In this method**
  - **A *plug-in* decision rule is derived**
  - **Begin with the *training* data**
  - **Feature vector is assumed to follow a probability distribution**
  - **The *training* data is used to estimate unknown parameters**
  - **The estimates are plugged into decision rule to determine class membership**
  - **The *training* data is known as the supervisor**

Probability based classifier methods
k-nearest neighbor method

k-nearest neighbor classifier; a heuristic based non-parametric method

- **The *training* samples are described by n dimensional attributes**
- **Each sample represents a point in n dimensional space**
- **The classifier searches for k closest *training* samples**
- **The closeness is Euclidean distance in n-dimensional space**

Rich neighborhood

Income 100000

Income? ~100000

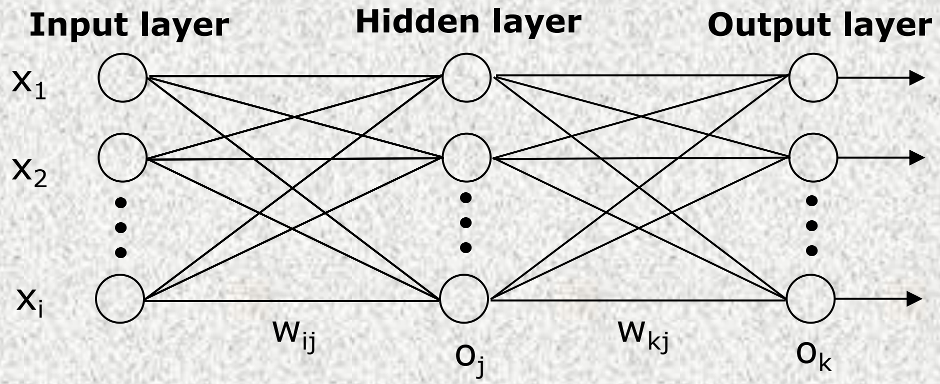
Poor neighborhood

Income 1000

Income? ~1000

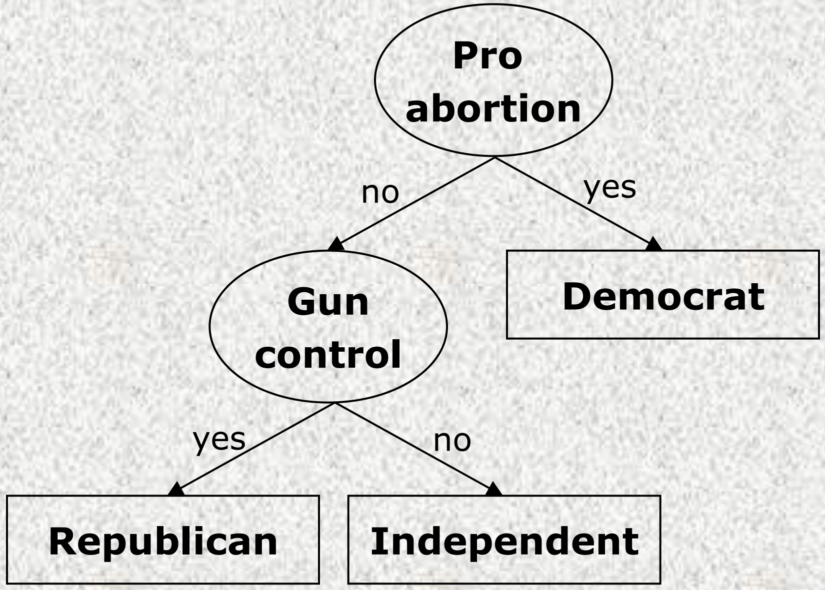
Probability based classifier methods
Artificial neural networks

- Artificial neural network; a non-parametric, non-linear modeling technique
  - By definition a neural network is a massively distributed parallel processor that acquires knowledge by optimizing the inter-neuron synaptic strengths by a learning process.
  - The computational elements in a neural network are known as nodes
  - The inputs are fed into a layer of units called input layer
  - The weighted output of these units are in turn fed into another layer called hidden layer (there can be more than one hidden layer)
  - The hidden layer’s weighted outputs are input to output layer
  - The output layer emits the network’s prediction for a given sample.



Probability based classifier methods
Decision trees

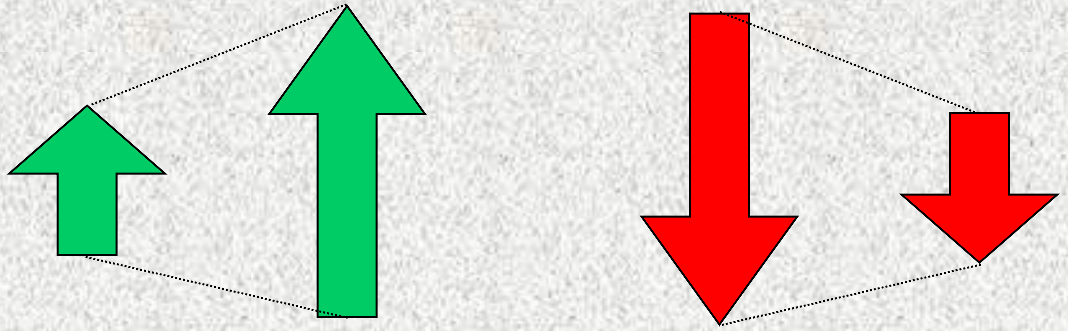
- **Decision trees**
  - **A decision tree is a flow-chart like tree structure**
  - **Each node denotes a test on an attribute or feature**
  - **Each branch represents an outcome of the test**
  - **Tree leaves represent classes or class labels**



Classification methods

- Common thread that runs across all these methods is a procedure that assigns an entity to a class based on a similarity measure with respect to some sort of an optimality criterion or a rule. The criterion may be maximize probability of a given class given an entity, minimize the distance between an entity and a given class

  (The class may be summarized in terms of its numerical characteristics.)



Raining cats and dogs

1/2

Let us revisit the classification problem of adult cats and dogs.

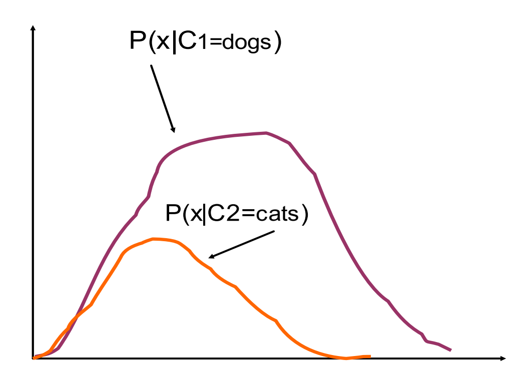
- Suppose that we do not get to see the *canines* and *felines*. However we have data on the following features and species.
- The features include
  - length of tongue, tail length, skull size, number of claws.

  So the four features together for each animal constitutes a four dimensional vector.

- The class variable is
  - "felines" or "canines."
- In this setting assigning animals to classes is not straight forward
- Since the features vary from animal to animal within a class, we may describe the data mathematically by a probability distribution.

Cats and dogs

- Figure shows distributions of measurements related to tail length of canines and felines. Since the data is separated by class labels prior probabilities, mixing proportions, component densities can be easily computed.



- Consider the two-category problem. Let the classes be denoted by $\boldsymbol{C_i}$, *(i=1,2)* and prior probabilities of the classes, and . Given an observation, the posterior probabilities of the two classes and are

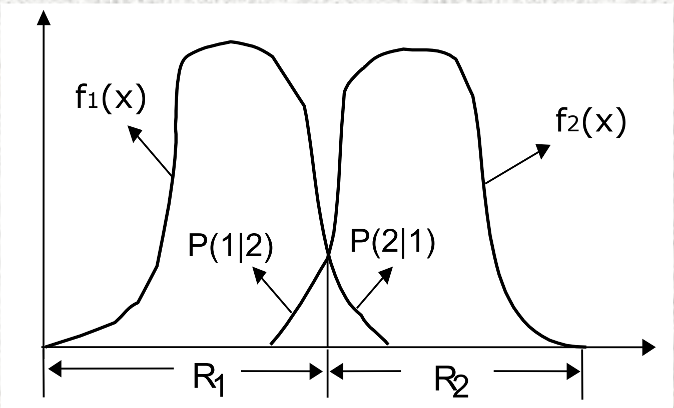
$$\pi(C_1 \mid x) = \frac{\pi(x \mid C_1)\pi}{\sum_{i=1}^{2} f(x \mid C_1)\pi + f(x \mid C_1)(1-\pi)}$$ and $$\pi(C_2 \mid x) = \frac{f(x \mid C_2)(1-\pi)}{\sum_{i=1}^{2} f(x \mid C_1)\pi + f(x \mid C_1)(1-\pi)}$$

- If $\pi(C_1 \mid x) > \pi(C_2 \mid x)$ then it is reasonable to suspect that x belongs to class C_1 or otherwise.

Fisher discriminant analysis

- **Statistical discriminant analysis computes a discriminant function for classifying observations into one of k categories on the basis of a set of quantitative variables.**
- **Discriminant analysis partitions a p-dimensional vector space into regions R_j (j=1,2,…,k), where R_j is the sub space containing all p-dimensional vectors x such that P(Ci | x) is a maximum.**
- **Under the assumption of multivariate normality, one can easily compute,** $P(C_j \mid x)$ **or some monotonic function** $g\{P(C_j \mid x)\}$ **.**
- **The posterior probability density and its variants serve as instruments or decision rules for classification.**

$P_{(error)}$= P(x falls in R_1 and indeed $x \in C_2$) +
P(x falls in R_2 and indeed $x \in C_1$)



Fisher discriminant analysis

- **Let $x_1, x_2, \ldots\ldots, x_n$ be a random sample from a normal population. In particular given an $x_i = x$, we have;**

$$f_1(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2\sigma^2}(x-\mu_1)^2}$$ **and** $$f_2(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}$$

and the ratio $$\frac{f_1(x)}{f_2(x)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2\sigma^2}(x-\mu_1)^2}}{\frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}}$$

By simplifying we get

$$x \geq \frac{\bar{x}_1 + \bar{x}_2}{2}$$

Linear Discriminant Function

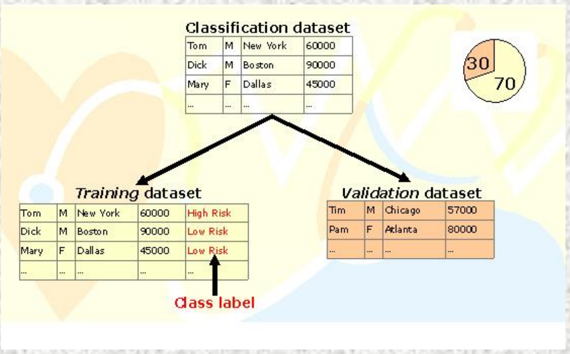
- Consider the Multivariate Gaussian density function. This function is a multivariate extension of the univariate normal density.
- The two-class problem we saw earlier can be extended to k-class problem.
- The data structure for k-class problem is

	class1						class2										classk			
X111	X121	X131	.	.	X1p1	X112	X122	X132	.	.	X1p2				X11k	X12k	X13k	.	.	X1pk
X211	X221	X231	.	.	X2p1	X212	X222	X232	.	.	X2p2				X21k	X22k	X23k	.	.	X2pk
X311	X321	X331	.	.	X3p1	X312	X322	X332	.	.	X3p2	.	.	.	X31k	X32k	X33k	.	.	X3pk
.	.	.	.	.	.	.	.	.	.	.	.				.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.				.	.	.	.	.	.
Xn11	Xn21	Xn21	.	.	Xnp1	Xn12	Xn22	Xn22	.	.	Xnp2				Xn1k	Xn2k	Xn2k	.	.	Xnpk

Discriminant analysis as a classifier

- In the application of discriminant analysis as a classifier, the labeled data is split into two sets; training data set and validation data set.
- Typically, 60-70% of the samples are utilized for training and the reminder for validation.
- Performance of the classifier is evaluated by computing the apparent error rate (AER).

Actual Class \ ***Predicted Class***	C_1	C_2
C_1	n_{11}	n_{12}
C_2	n_{21}	n_{22}



- Mathematically apparent error rate is simply

$$AER = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

Cross Validation

- The cross validation methodology is known as k-fold cross validation
- In this techniques labeled data set is divided into k sets of equal size
- The classifier is trained ***k*** times, such that one of the sets is held out as a validation data set at each training step
- The performance of the classifier is simply the average apparent error rate across the ***k*** training data sets.

Example

- The iris data published by Dr. Ronald A. Fisher in 1936 is widely used to benchmark performance of classifiers.
- Iris data consists of a total of 150 observations on sepal length, sepal width, petal length, and petal width measured in millimeters.
- Fifty iris specimens from each of three species, iris setosa, iris versicolor, and iris virginica are taken.
- The data set is given in Table

Example

150 observations on the four attributes of three species of flowers

50	33	14	2	1	64	28	56	22	3	65	28	46	15	2	67	31	56	24	3	47	32	13	2	1	46	31	15	2	1
63	28	51	15	3	46	34	14	3	1	69	31	51	23	3	62	22	45	15	2	74	28	61	19	3	59	30	42	15	2
59	32	48	18	2	46	36	10	2	1	61	30	46	14	2	60	27	51	16	2	56	28	49	20	3	60	22	40	10	2
65	30	52	20	3	56	25	39	11	2	65	30	55	18	3	58	27	51	19	3	49	31	15	1	1	67	31	47	15	2
68	32	59	23	3	51	33	17	5	1	57	28	45	13	2	62	34	54	23	3	56	30	41	13	2	63	25	49	15	2
77	38	67	22	3	63	33	47	16	2	67	33	57	25	3	76	30	66	21	3	51	25	30	11	2	57	28	41	13	2
49	25	45	17	3	55	35	13	2	1	67	30	52	23	3	70	32	47	14	2	54	39	13	4	1	51	35	14	3	1
64	32	45	15	2	61	28	40	13	2	48	31	16	2	1	59	30	51	18	3	61	29	47	14	2	56	29	36	13	2
55	24	38	11	2	63	25	50	19	3	64	32	53	23	3	52	34	14	2	1	68	30	55	21	3	55	25	40	13	2
49	36	14	1	1	54	30	45	15	2	79	38	64	20	3	44	32	13	2	1	45	23	13	3	1	57	25	50	20	3
67	33	57	21	3	50	35	16	6	1	58	26	40	12	2	44	30	13	2	1	55	23	40	13	2	66	30	44	14	2
77	28	67	20	3	63	27	49	18	3	47	32	16	2	1	55	26	44	12	2	51	37	15	4	1	52	35	15	2	1
50	23	33	10	2	72	32	60	18	3	48	30	14	3	1	51	38	16	2	1	63	33	60	25	3	53	37	15	2	1
61	30	49	18	3	48	34	19	2	1	50	30	16	2	1	50	32	12	2	1	69	32	57	23	3	62	29	43	13	2
61	26	56	14	3	64	28	56	21	3	43	30	11	1	1	58	40	12	2	1	51	34	15	2	1	50	35	13	3	1
51	38	19	4	1	67	31	44	14	2	62	28	48	18	3	49	30	14	2	1	73	29	63	18	3	67	25	58	18	3
51	35	14	2	1	56	30	45	15	2	58	27	41	10	2	50	34	16	4	1	63	23	44	13	2	54	37	15	2	1
46	32	14	2	1	60	29	45	15	2	57	26	35	10	2	57	44	15	4	1	61	28	47	12	2	64	29	43	13	2
50	36	14	2	1	77	30	61	23	3	63	34	56	24	3	58	27	51	19	3	65	30	58	22	3	69	31	54	21	3
57	29	42	13	2	72	30	58	16	3	54	34	15	4	1	52	41	15	1	1	72	36	61	25	3	65	32	51	20	3
71	30	59	21	3	64	31	55	18	3	60	30	48	18	3	63	29	56	18	3	69	31	49	15	2	64	27	53	19	3
49	24	33	10	2	56	27	42	13	2	57	30	42	12	2	55	42	14	2	1	48	34	16	2	1	48	30	14	1	1
49	31	15	2	1	77	26	69	23	3	60	22	50	15	3	54	39	17	4	1	57	38	17	3	1	51	38	15	3	1
66	29	46	13	2	52	27	39	14	2	60	34	45	16	2	50	34	15	2	1	68	28	48	14	2	54	34	17	2	1
44	29	14	2	1	50	20	35	10	2	55	24	37	10	2	58	27	39	12	2	58	28	51	24	3	67	30	50	17	2

Example

- We will use the **PROC DISCRIM** procedure in SAS
- The data set consisting of 50 observations is set aside for a validation
- Fragments of output from the program is reproduced in tables below

Class Level Information					
spec_no	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	26	26.0000	0.262626	0.333333
2	_2	40	40.0000	0.404040	0.333333
3	_3	33	33.0000	0.333333	0.333333
Summary of data in the three classes. Note that spec_no is species number. 1,2, and 3 map to setosa, versicula and virginica respectively. Since there are 50 observations in each species type, the prior probability of classes is 0.33 each.					

Example

Generalized Squared Distance Function

$$D^2_j(X) = (X-\overline{X}_j)' \, COV^{-1} \, (X-\overline{X}_j)$$

Number of Observations and Percent Classified into spec_no				
From spec no	1	2	3	Total
1	26 100.00	0 0.00	0 0.00	26 100.00
2	0 0.00	39 97.50	1 2.50	40 100.00
3	0 0.00	1 3.03	32 96.97	33 100.00
Total	26 26.26	40 40.40	33 33.33	99 100.00
Priors	0.33333	0.33333	0.33333	

Output shows classification of observations from one species class into another species class. Notice that none of the samples belonging to species setosa is misclassified. However, 1 out 40 from versicula, and 1 out 33 from virginica are misclassified. These results produced using the training data.

Example

Error Count Estimates for spec_no				
	1	2	3	Total
Rate	0.0000	0.0250	0.0303	0.0184
Priors	0.3333	0.3333	0.3333	
Table is simply a compendium of error rates/misclassification rates and prior probabilities				

Example

Number of Observations and Percent Classified into spec_no				
From spec_no	1	2	3	Total
1	24 100.00	0 0.00	0 0.00	24 100.00
2	0 0.00	10 100.00	0 0.00	10 100.00
3	0 0.00	0 0.00	17 100.00	17 100.00
Total	24 47.06	10 19.61	17 33.33	51 100.00
Priors	0.33333	0.33333	0.33333	

Output shows classification of observations from one species class into another species class. Notice that none of the samples belonging to any of the three species is misclassified. These results produced using the validation data set consisting of 51 observations.

Example

Error Count Estimates for spec_no				
	1	2	3	Total
Rate	0.0000	0.0000	0.0000	0.0000
Priors	0.3333	0.3333	0.3333	
This Table is simply a compendium of error rates/misclassification rates and prior probabilities				


THANK YOU