SLIDE-1

Data Mining Supervised:

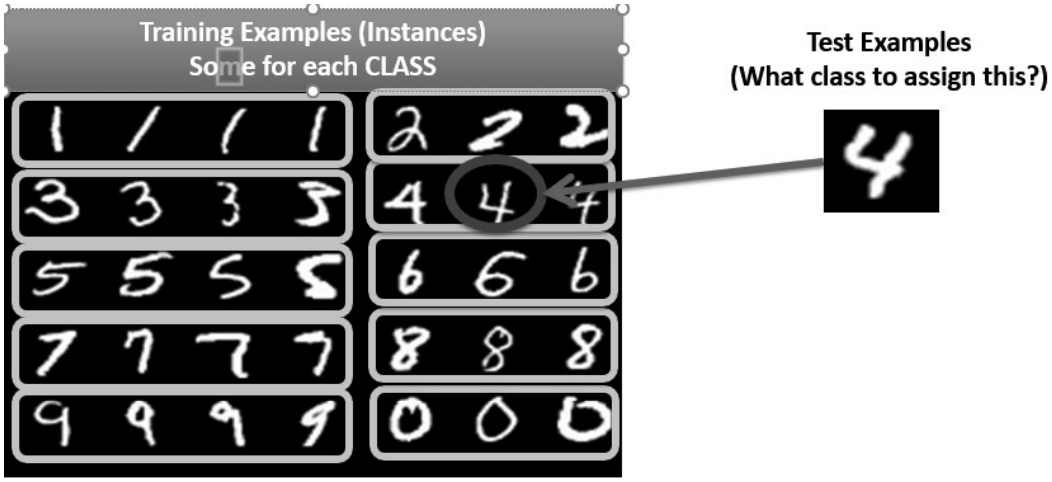
- **KNN**
- **Naïve Bayes**
- **Decision trees**
- **Random forest**
- **Neural networks**
- **Support vector machines**

SLIDE-2

k-Nearest Neighbor Classifiers

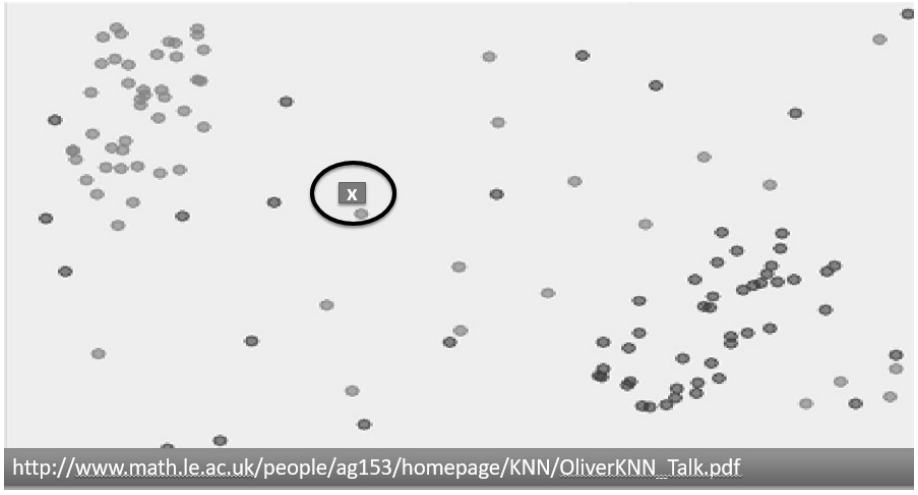
SLIDE-3

1-Nearest Neighbor Classifier



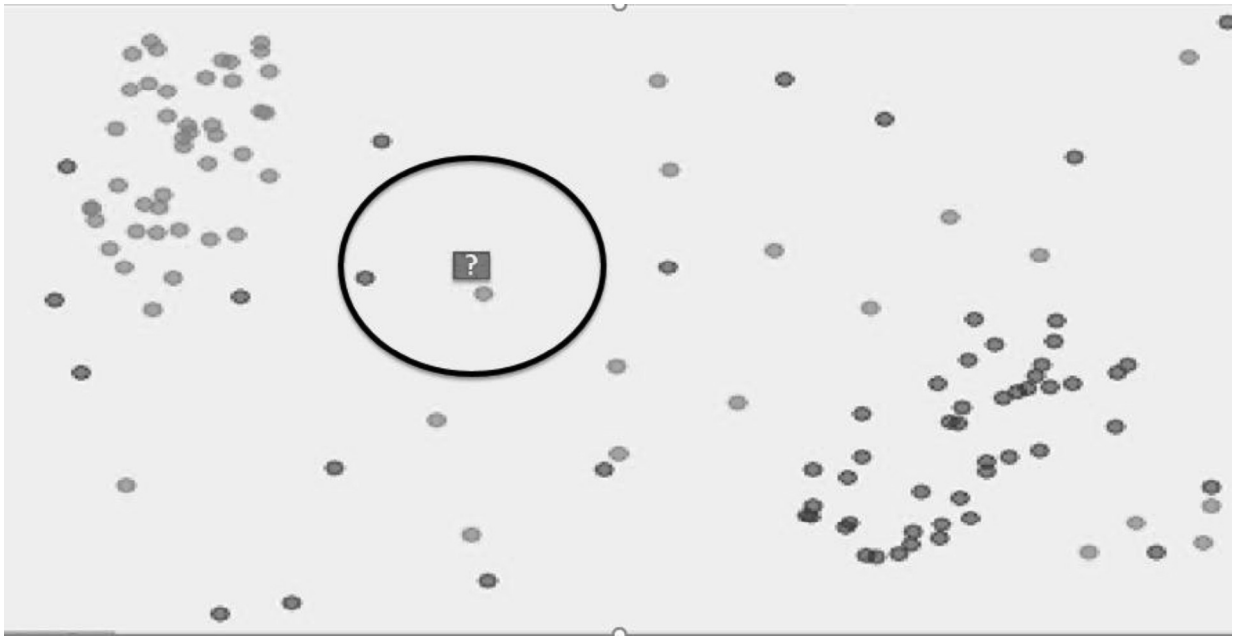
SLIDE-4

1-Nearest Neighbor



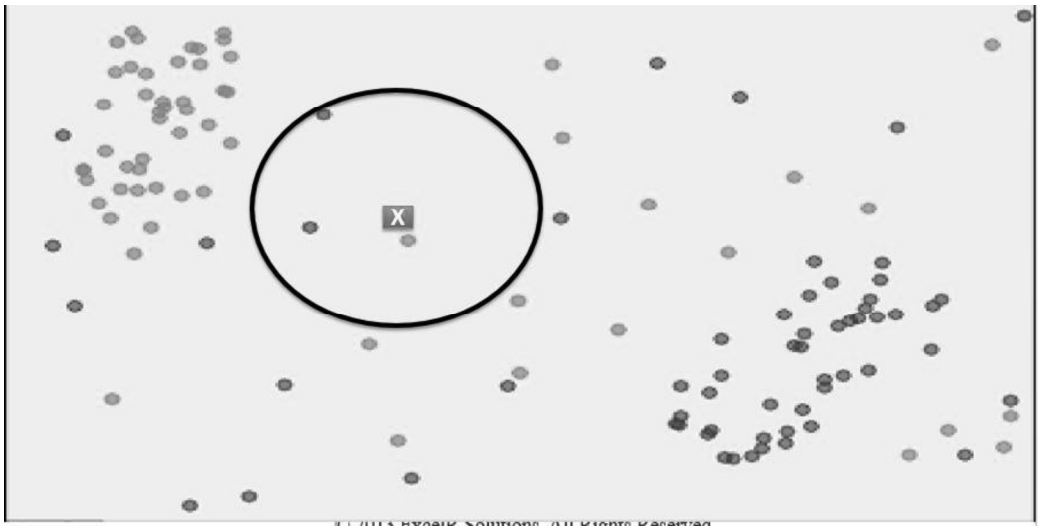
SLIDE-5

2-Nearest Neighbor



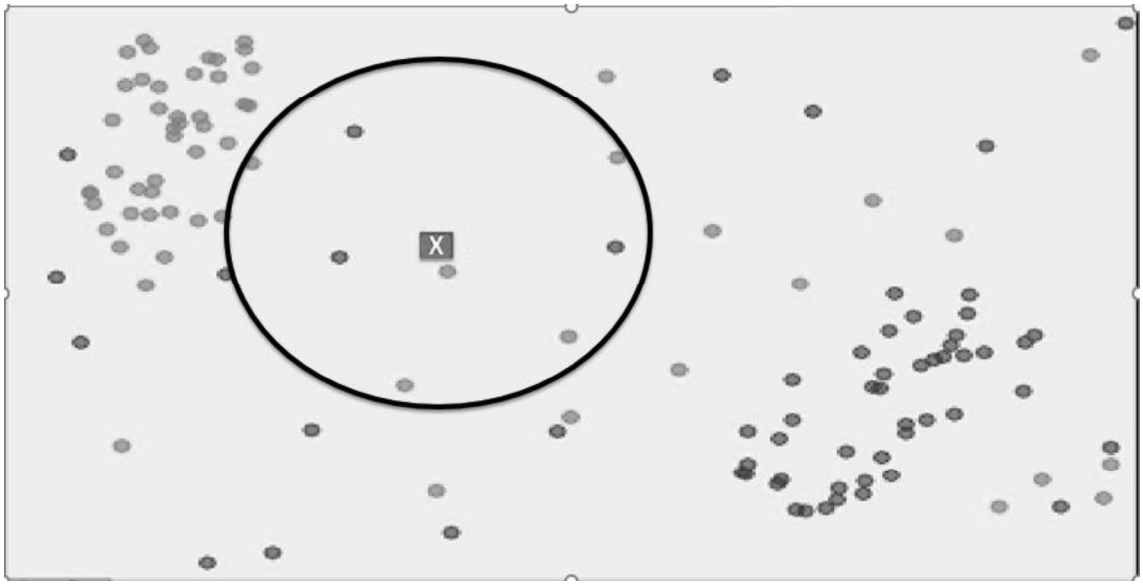
SLIDE-6

3-Nearest Neighbor



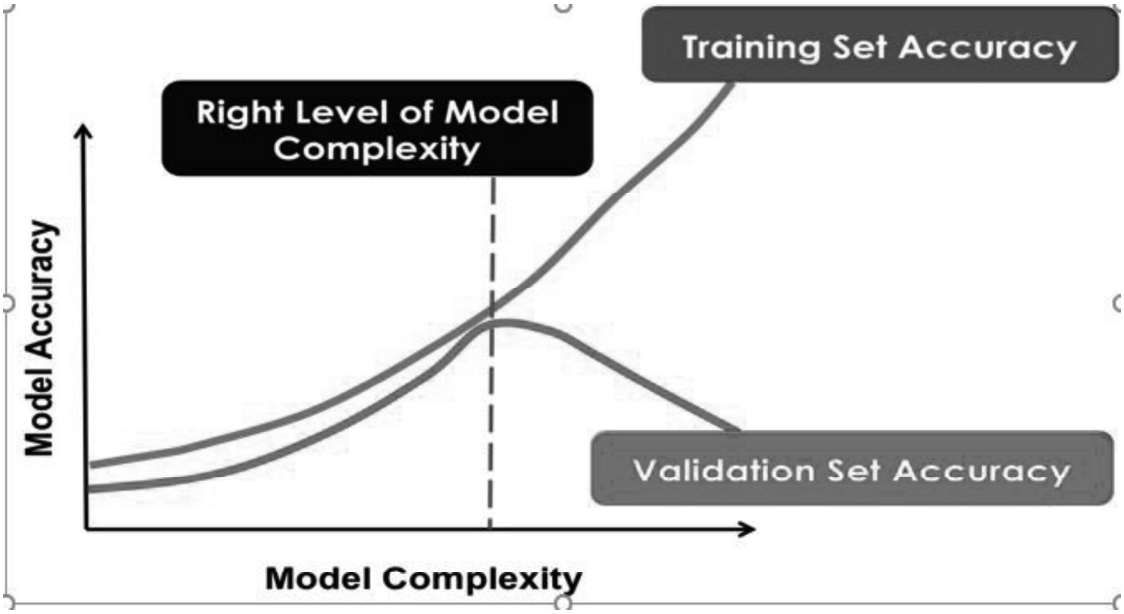
SLIDE-7

8-Nearest Neighbor



SLIDE-8

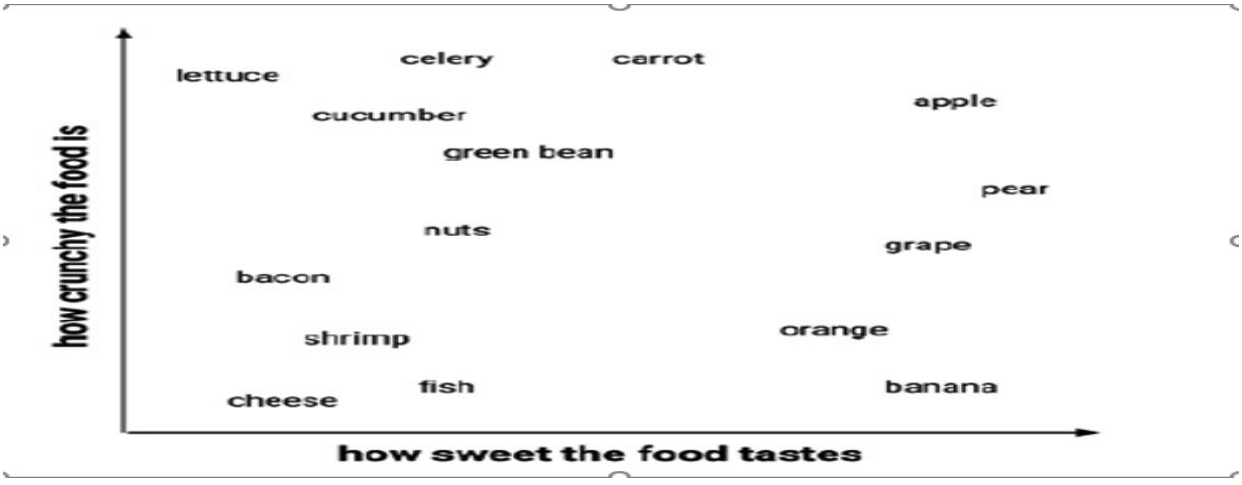
Controlling COMPLEXITY in k-NN



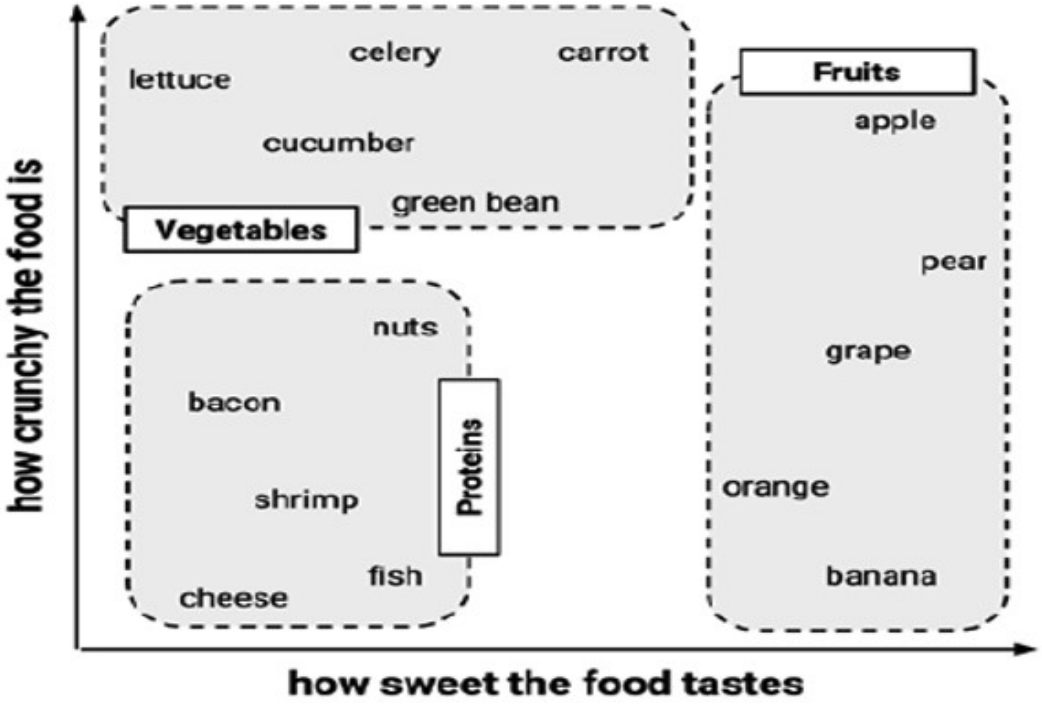
SLIDE-9

Ingredient	Sweetness	Crunchiness	Food type
apple	10	9	fruit
Bacon	1	4	protein
banana	10	1	fruit
carrot	7	10	vegetable
celery	3	10	vegetable
cheese	1	1	protein

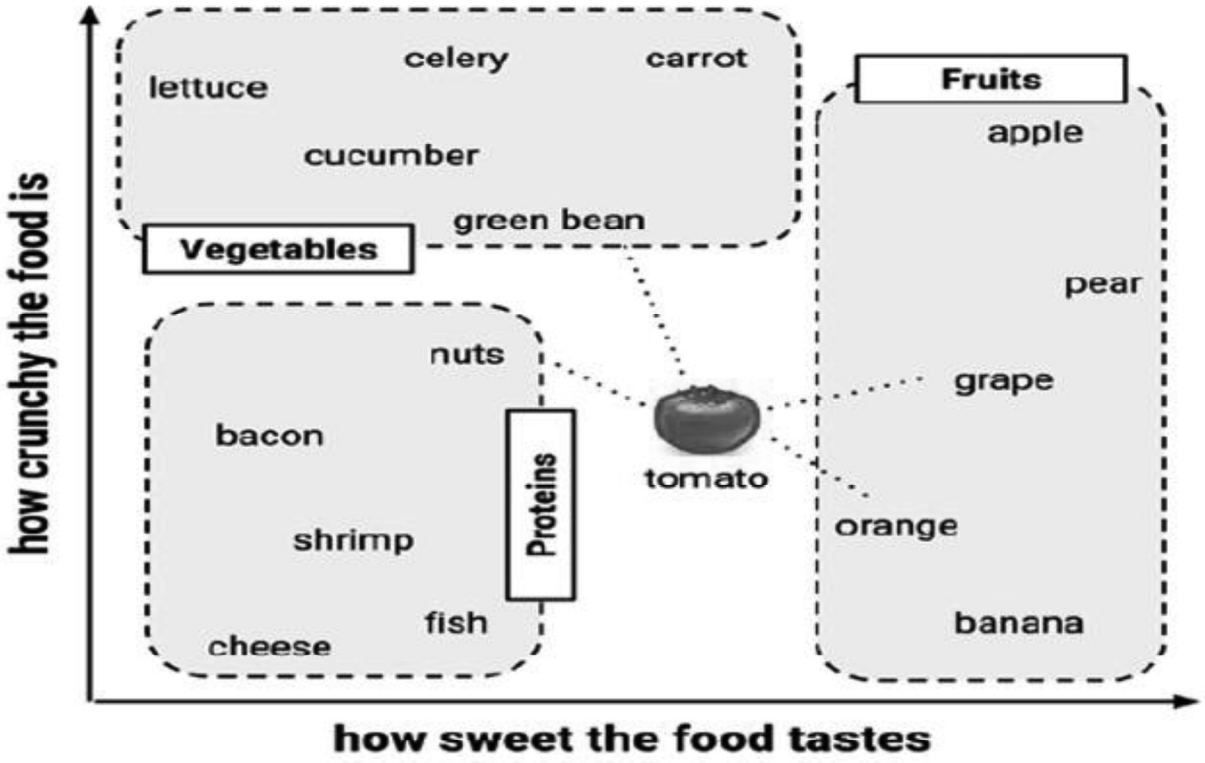
SLIDE-10



SLIDE-11



SLIDE-12



SLIDE-13

Measuring similarity with distance

$$\text{dist}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

Which Class Tomoto belongs to given the feature values:

Tomato (*sweetness = 6*, *crunchiness = 4*),

SLIDE-14

Bayesian Classifiers

SLIDE-15

Understanding probability

The probability of an event is estimated from the observed data by dividing the number of trials in which the event occurred by the total number of trials

For instance, if it rained 3 out of 10 days with similar conditions as today, the probability of rain today can be estimated as 3 / 10 = 0.30 or 30 percent.

Similarly, if 10 out of 50 prior email messages were spam, then the probability of any incoming message being spam can be estimated as 10 / 50 = 0.20 or 20 percent.

For example, given the value P(spam) = 0.20, we can calculate P(ham) = 1 – 0.20 = 0.80

Note: The probability of all the possible outcomes of a trial must always sum to 1

SLIDE-16

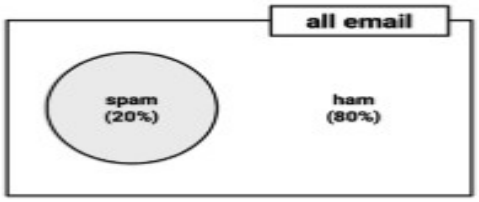
Understanding probability cont..

For example, given the value P(spam) = 0.20, we can calculate P(ham) = 1 – 0.20 = 0.80

Because an event cannot simultaneously happen and not happen, an event is always mutually exclusive and exhaustive with its complement

The complement of event A is typically denoted Ac or A'.

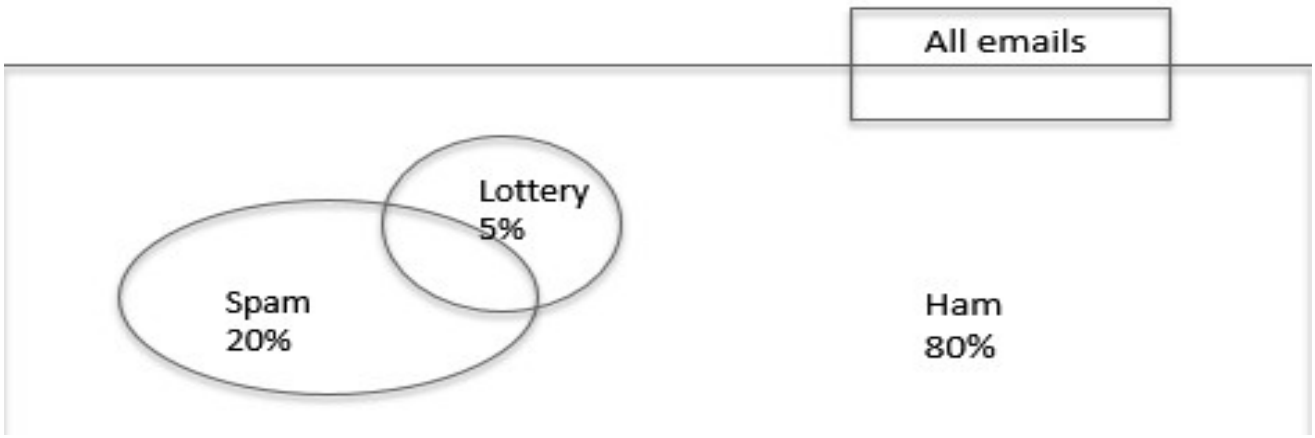
Additionally, the shorthand notation P(¬A) can used to denote the probability of event A not occurring, as in P(¬spam) = 0.80. This notation is equivalent to P(Ac).



SLIDE-17

Understanding joint probability

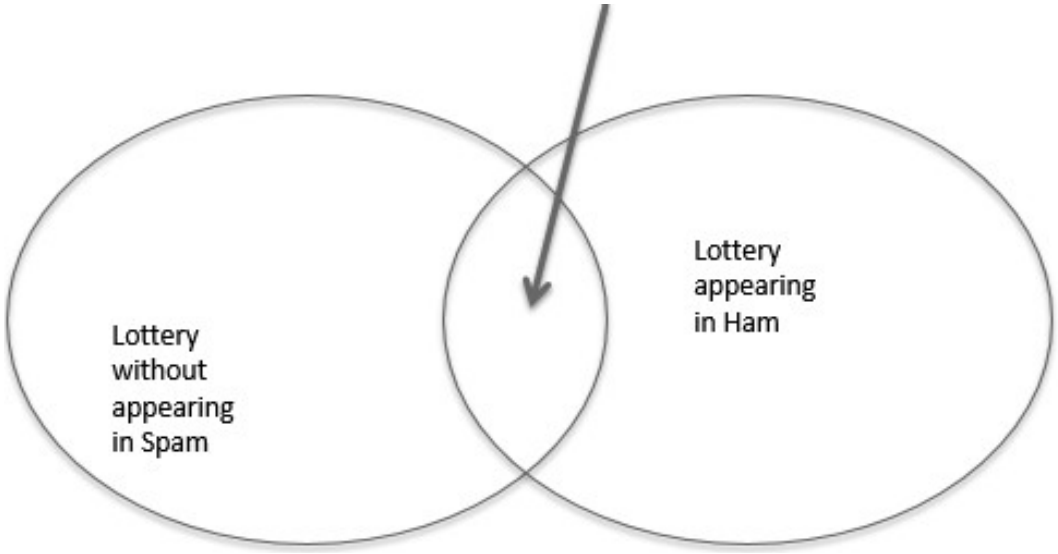
Often, we are interested in monitoring several nonmutually exclusive events for the same trial



SLIDE-18

Understanding joint probability

Lottery appearing in spam



Estimate the probability that both P(spam) and P(Spam) occur, which can be written as P(spam ∩ Lottery). the notation A ∩ B refers to the event in which both A and B occur.

SLIDE-19

Calculating P(spam ∩ Lottery) depends on the joint probability of the two events or how the probability of one event is related to the probability of the other.

If the two events are totally unrelated, they are called independent events

If P(spam) and P(Lottery) were independent, we could easily calculate P(spam ∩ Lottery), the probability of both events happening at the same time.

Because 20 percent of all the messages are spam, and 5 percent of all the e-mails contain the word Lottery, we could assume that 1 percent of all messages are spam with the term Lottery.

More generally, for independent events A and B, the probability of both happening can be expressed as P(A ∩ B) = P(A) * P(B).

0.05 * 0.20 = 0.01

SLIDE-20

Bayes Rule

$P(A|B) = P(A \cap B)/P(B) = P(B/A)P(A)/P(B)$

- Bayes Rule: The most important Equation in ML!!

P(CLASS/DATA) = P(CLASS) P(DATA/CLASS)/P(DATA)

LHS → POSTERIOR PROBABILITY (Probability of class AFTER seeing the data)

RHS → DATA PRIOR (MARGINAL)

SLIDE-21

Naïve Bayes Classifier

SLIDE-22

Conditional Independence

$P(Fever,BodyAche|Viral) = P(Fever|Viral) * P(BodyAche|Viral)$

Viral Infection

- Fever
- Body Ache

  $P(Fever,BodyAche) \neq P(Fever)P(BodyAche)$

- Simple Independence between two variables:

  $P(X1,X2) = P(X1)\ P(X2)$

- Class Conditional Independence assumption:

  $P(X1,X2/C) = P(X1/C)P(X2/C)$

SLIDE-23

Naïve Bayes Classifier

Conditional Independence among variables given Classes!

$$P(C|X_1,X_2,...,X_D) = \frac{P(C)P(X_1,X_2,...,X_D|C)}{\sum_{C'} P(X_1,X_2,...,X_D|C')} = \frac{P(C)\prod_{d=1}^{D} P(X_d|C)}{\sum_{C'}\prod_{d=1}^{D} P(X_d|C')}$$

- Simplifying assumption
- Baseline model especially when large number of features
- Taking log and ignoring denominator:

$$\log\left(P(C|X_1,X_2,...,X_D)\right) \propto \log(P(C)) + \sum_{d=1}^{D} \log\left(P(X_d|C)\right)$$

SLIDE-24

Naïve Bayes Classifier for Categorical Valued Variables

Let's Naïve Bayes!

$$\log\left(P(C|X_1,X_2,...,X_D)\right) \propto \log(P(C)) + \sum_{d=1}^{D} \log\left(P(X_d|C)\right)$$

Class Prior Parameters:

$P(\text{Like} = \text{Y}) = ???$

$P(\text{Like} = \text{N}) = ???$

Class Conditional Likelihoods

$P(\text{Color} = \text{Red} | \text{Like} = \text{Y}) = ????$

$P(\text{Color} = \text{Red} | \text{Like} = \text{N}) = ????$

...

$P(\text{Shape} = \text{Triangle} | \text{Like} = \text{N}) = ????$

#EXMPLS	COLOR	SHAPE	LIKE
20	Red	Square	Y
10	Red	Circle	Y
10	Red	Triangle	N
10	Green	Square	N
5	Green	Circle	Y
5	Green	Triangle	N
10	Blue	Square	N
10	Blue	Circle	N
20	Blue	Triangle	Y

SLIDE-26

Parameter Estimation

$$\log\left(P\left(C|X_1, X_2, ..., X_D\right)\right) \propto \log\left(P(C)\right) + \sum_{d=1}^{D} \log\left(P\left(X_d|C\right)\right)$$

- **What / How many Parameters?**

P(C)

$$= \frac{N(c)}{N} \approx \frac{N(c)+\lambda}{N(c)+\lambda \times |\mathbf{C}|}$$

- **Class Priors:**
- **Conditional Probabilities:**

$$X_d = \left\{ v_1^{(d)}, v_2^{(d)}, \ldots, v_{M_d}^{(d)} \right\}$$

$$P\left(v_m^{(d)} | c\right) = \frac{N\left(v_m^{(d)}, c\right)}{N(c)} \cup \frac{N\left(v_m^{(d)}, c\right)+\lambda}{N(c)+M_d\lambda}$$

SLIDE-27

Naïve Bayes Classifier for Text Classifier

Text Classification Example

- Doc1 = {buy two shirts get one shirt half off}
- Doc2 = {get a free watch. send your contact details now}
- Doc3 = {your flight to chennai is delayed by two hours}
- Doc4 = {you have three tweets from @sachin}

Four Class Problem:

- Spam, P(promo|doc1) =0.84
- Promotions, P(spam|doc2) = 0.94
- Social, P(main|doc3) = 0.75
- Main P(social|doc4) = 0.91

SLIDE-29

Bag-of-Words Representation

- Structured (e.g. Multivariate) data – fixed number of features
- Unstructured (e.g. Text) data
  - Arbitrary length documents,
  - High dimensional feature space (many words in vocabulary),
  - Sparse (small fraction of vocabulary words present in a doc.)
- Bag-of-Words Representation:
  - Ignore Sequential order of words
  - Represent as a Weighted-Set – Term Frequency of each term
- RawDoc = {buy two shirts get one shirt half off}

- Stemming = {buy two shirt get one shirt half off}
- BoW's = {buy:1, two:1, shirt:2, get:1, one:1, half:1, off:1}

SLIDE-30

Naïve Bayes Classifier with BoW

BoW = {buty:1, two:1, shirt:2, get:1, one:1, half:1, off:1}

- Make an "independence assumption" about words | class

P(doc1|promo) = P(buy:1,two:1,shirt:2,get:1,one:1,half:1,off:1|promo)

=P(buy:1|promo)*P(two:1|promo)*P(shirt:2|promo)*P(get:1|promo)*P(one:1|promo)*P(free:1|promo)

= $P(buy|promo)^1$ * $P(two|promo)^1$ * $P(shirt|promo)^2$ *$P(get|promo)^1$ * $P(one|promo)^1$ * $P(free|promo)^1$

SLIDE-31

Naïve Bayes Text Classifiers

- **Log Likelihood of document given class.**

$$doc = \left\{ tf\left(w_m\right) \right\}_{m=1}^{M}$$

$$tf\left(w_m\right) = \text{Number of times word } w_m \text{ occurs in doc}$$

$$\mathrm{P}\left(doc|class\right) = \mathrm{P}\left(w_1|class\right)^{tf(w_1)} \mathrm{P}\left(w_2|class\right)^{tf(w_2)} ... \mathrm{P}\left(w_M|class\right)^{tf(w_M)}$$

- **Parameters in Naïve Bayes Text classifiers:**

$$\mathrm{P}\left(w_m|c\right) = \text{Probability that word } w_m \text{ occurs in documents of class } c$$

$$\mathrm{P}\left(shirt|promo\right), \mathrm{P}\left(free|spam\right), \mathrm{P}\left(buy|spam\right), \mathrm{P}\left(buy|promo\right), ...$$

$$\text{Number of parameters} = ??$$

SLIDE-32

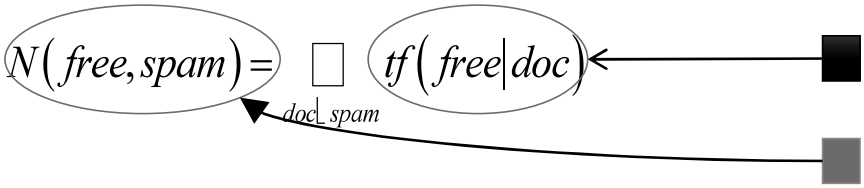
Naïve Bayes Parameters

- **Likelihood of a word given class. For each word, each class.**
- **Estimating these parameters from data:**

$$\mathrm{P}\left(w_m|c\right) = \text{Probability that word } w_m \text{ occurs in documents of class } c$$

- **Estimating these parameters from data:**

$$N(w_m, c) = \text{Number of times word } w_m \text{ occurs in documents of class } c$$



$$N(free, promo) = \sum_{doc \lfloor promo} tf(free|doc)$$

SLIDE-33

- Likelihood of a word given class. For each word, each class.

$$P(w_m|c) = \text{Probability that word } w_m \text{ occurs in documents of class } c$$

- Normalize these counts to probabilities.

$$P(free|spam) = \frac{N(free, spam)}{\sum_{m'=1}^{M} N(w_{m'}, spam)} \approx \frac{1 + N(free, spam)}{M + \sum_{m'=1}^{M} N(w_{m'}, spam)}$$

- Smoothing is done to make sure we don't get zero products

SLIDE-34

Bayesian Classifier Multi-variate real-valued data

SLIDE-35

Bayes Rule

Class Prior **Data Likelihood given Class**

$$P(Class|Data) = \frac{P(Class)P(Data|Class)}{P(Data)}$$

Posterior Probability Data Prior (Marginal)

(Probability of class AFTER seeing the data)

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})}$$

$$\mathbf{x} \in R^D$$

SLIDE-36

Simple Bayesian Classifier

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})}$$

P(x/c) →Each Class Conditional Probability is assumed to be a Uni-Modal (Single Cloud) (NORMAL) Distribution.

$$P(\mathbf{x}|c) = \mathrm{N}(\mathbf{x}|\mathbf{m}_c, \Sigma_c) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_c)^T \Sigma_c^{-1}(\mathbf{x}-\mathbf{m}_c)\right)$$

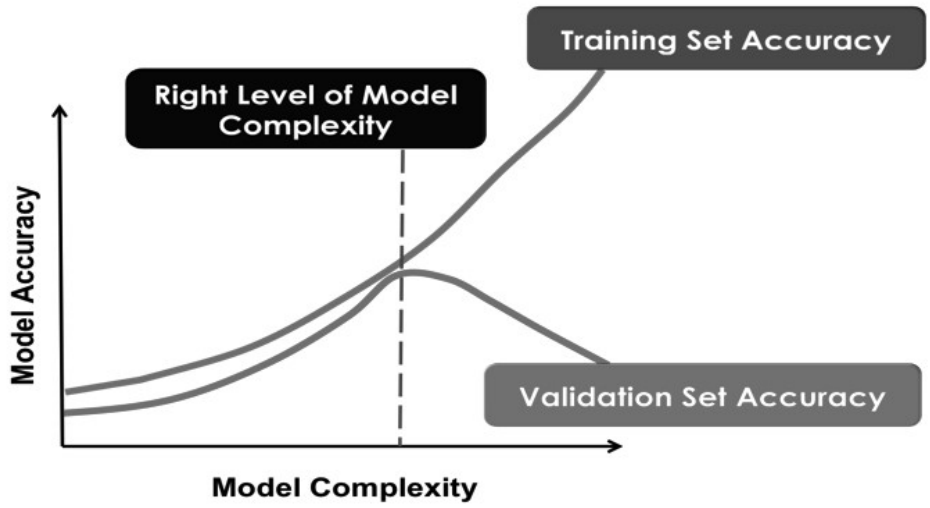
$$\text{Sum:} \int_{\mathbf{x} \in R^D} P(\mathbf{x}|c)\, d\mathbf{x} = 1$$

$$\text{Mean:} \int_{\mathbf{x} \in R^D} \mathbf{x} P(\mathbf{x}|c)\, d\mathbf{x} = \mathbf{m}_c$$

$$\text{Co-Variance:} \int_{\mathbf{x} \in R^D} (\mathbf{x} - \mathbf{m}_c)(\mathbf{x} - \mathbf{m}_c)^T P(\mathbf{x}|c)\, d\mathbf{x} = \Sigma_c$$

SLIDE-37

Controlling COMPLEXITY



SLIDE-38

Spherical-SAME Covariance/Class

$$P(\mathbf{x}|c) = \frac{1}{Z(\mathbf{m}_c, \Sigma_c)} \exp\left(-\frac{1}{2} \Delta(\mathbf{x}, \mathbf{m}_c|\Sigma_c)^2 \right)$$

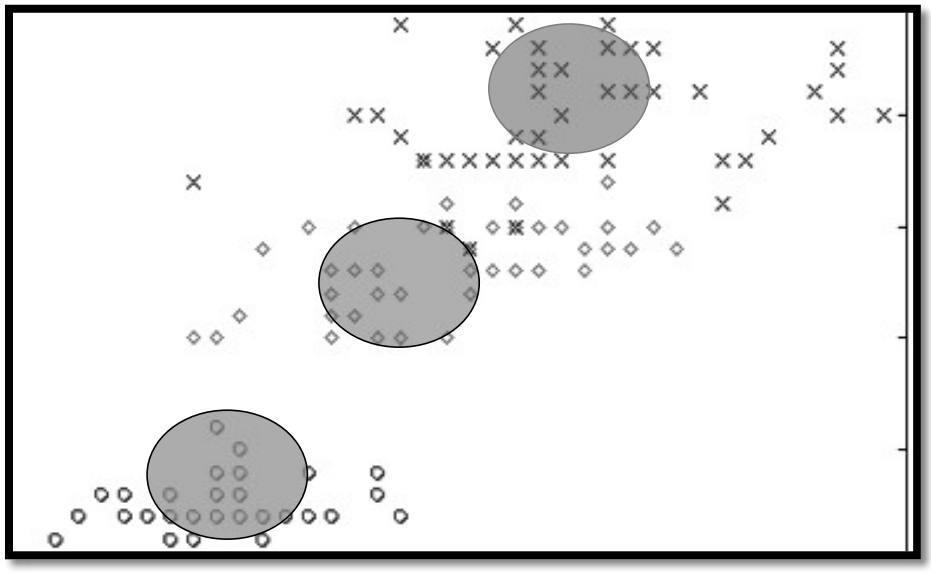
$$= \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{m}_c)^T \Sigma_c^{-1} (\mathbf{x} - \mathbf{m}_c) \right)$$

$$\Sigma_c = \begin{bmatrix} \sigma & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 \\ 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & \sigma \end{bmatrix}$$

$$= \frac{1}{\left(\sqrt{2\pi}\right)^D \sigma} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_d - m_{c,d}}{\sigma} \right)^2 \right)$$

Euclidian Distance

$$\Delta\left(\mathbf{x}, \mathbf{m}_c \middle| \Sigma_c\right) = \sqrt{\sum_{d=1}^{D} \left(\frac{x_d - m_{c,d}}{\sigma} \right)^2}$$



SLIDE-39

Spherical-**DIFFERENT** Covariance/Class

$$P(\mathbf{x}|c) = \frac{1}{Z(\mathbf{m}_c, \Sigma_c)} \exp\left(-\frac{1}{2} \Delta(\mathbf{x}, \mathbf{m}_c|\Sigma_c)^2 \right)$$

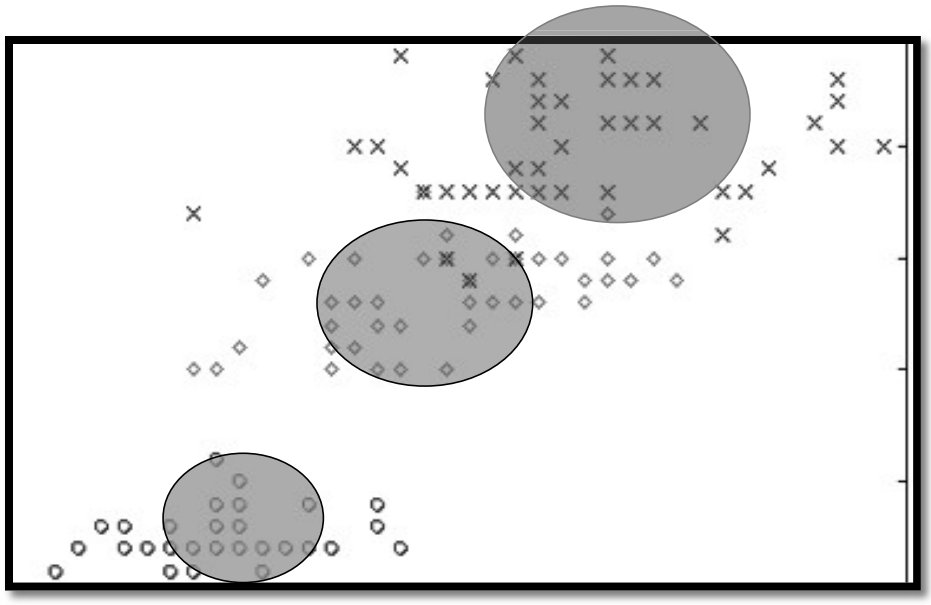
$$= \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{m}_c)^T \Sigma_c^{-1} (\mathbf{x} - \mathbf{m}_c) \right)$$

$$\Sigma_c = \begin{bmatrix} \sigma_c & 0 & 0 & 0 \\ 0 & \sigma_c & 0 & 0 \\ 0 & 0 & \sigma_c & 0 \\ 0 & 0 & 0 & \sigma_c \end{bmatrix}$$

$$= \frac{1}{\left(\sqrt{2\pi}\right)^D \sigma} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_d - m_{c,d}}{\sigma} \right)^2 \right)$$

Class Scaled Euclidian Distance

$$\Delta(\mathbf{x}, \mathbf{m}_c|\Sigma_c) = \sqrt{\sum_{d=1}^{D} \left(\frac{x_d - m_{c,d}}{\sigma_c} \right)^2}$$



SLIDE-40

Independent-Covariance/Class

$$P(\mathbf{x}|c) = \frac{1}{Z(\mathbf{m}_c, \Sigma_c)} \exp\left(-\frac{1}{2} \Delta(\mathbf{x}, \mathbf{m}_c|\Sigma_c)^2 \right)$$

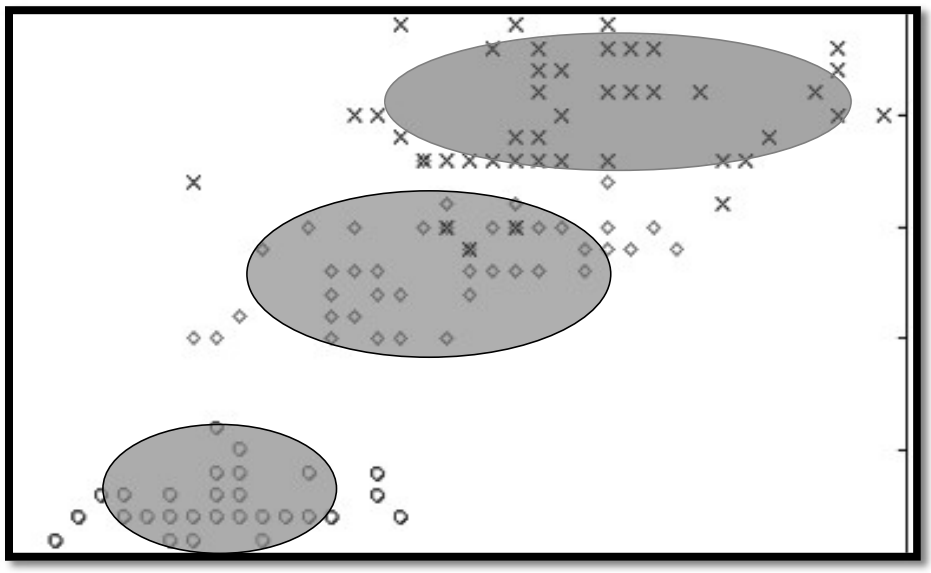
$$= \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mathbf{m}_c)^T \Sigma_c^{-1} (\mathbf{x} - \mathbf{m}_c) \right)$$

$$\Sigma_c = \begin{bmatrix} \sigma_{c,1} & 0 & 0 & 0 \\ 0 & \sigma_{c,2} & 0 & 0 \\ 0 & 0 & \sigma_{c,3} & 0 \\ 0 & 0 & 0 & \sigma_{c,4} \end{bmatrix}$$

$$= \frac{1}{\left(\sqrt{2\pi}\right)^D \sigma} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_d - m_{c,d}}{\sigma} \right)^2 \right)$$

Class/Dim. Scaled Eucledian Distance

$$\Delta\left(\mathbf{x},\mathbf{m}_c\middle|\Sigma_c\right)=\sqrt{\sum_{d=1}^{D}\left(\frac{x_d-m_{c,d}}{\sigma_{c,d}}\right)^2}$$



SLIDE-41

FULL-Covariance/Class

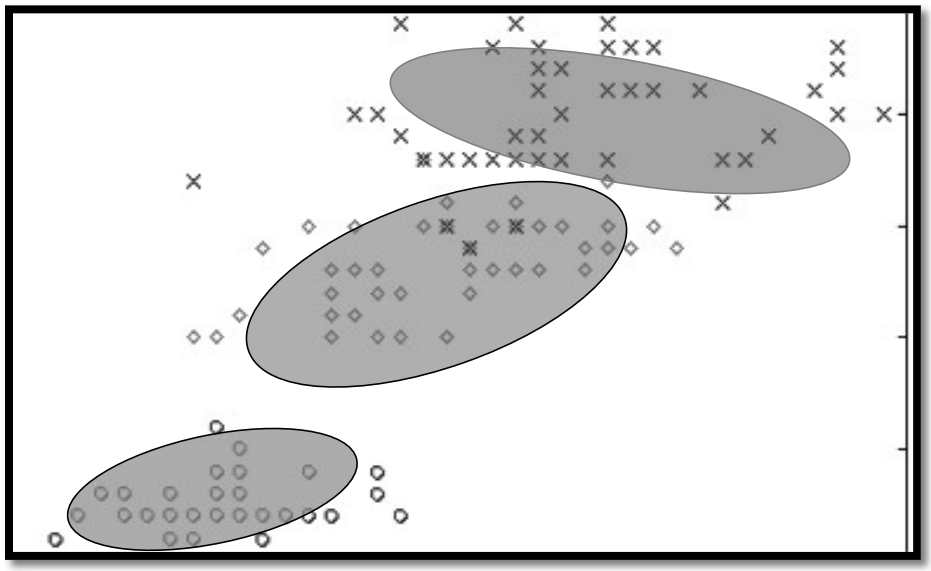
$$P\left(\mathbf{x}|c\right)=\frac{1}{Z\left(\mathbf{m}_c,\Sigma_c\right)}\exp\left(-\frac{1}{2}\Delta\left(\mathbf{x},\mathbf{m}_c\middle|\Sigma_c\right)^2\right)$$

$$=\frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_c\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{m}_c\right)^T\Sigma_c^{-1}\left(\mathbf{x}-\mathbf{m}_c\right)\right)$$

$$\Sigma_c = \begin{bmatrix} \sigma_{1,1}^{(c)} & \sigma_{1,2}^{(c)} & \sigma_{1,3}^{(c)} & \sigma_{1,4}^{(c)} \\ \sigma_{2,1}^{(c)} & \sigma_{2,2}^{(c)} & \sigma_{2,3}^{(c)} & \sigma_{2,4}^{(c)} \\ \sigma_{3,1}^{(c)} & \sigma_{3,2}^{(c)} & \sigma_{3,3}^{(c)} & \sigma_{3,4}^{(c)} \\ \sigma_{4,1}^{(c)} & \sigma_{4,2}^{(c)} & \sigma_{4,3}^{(c)} & \sigma_{4,4}^{(c)} \end{bmatrix}$$

Mahalanobis Distance

$$\Delta(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \sum^{-1} (\mathbf{x}-\mathbf{y})}$$



SLIDE-42

Summary

- Classifier = Partitions Input Space into “Pure” regions
- Rule Based Classifiers:
  - Version Space

  - Decision Trees
- Descriptive Classification:
  - Bayesian Classifiers
- Discriminative Classification:
  - Perceptron
  - Logistic Regression