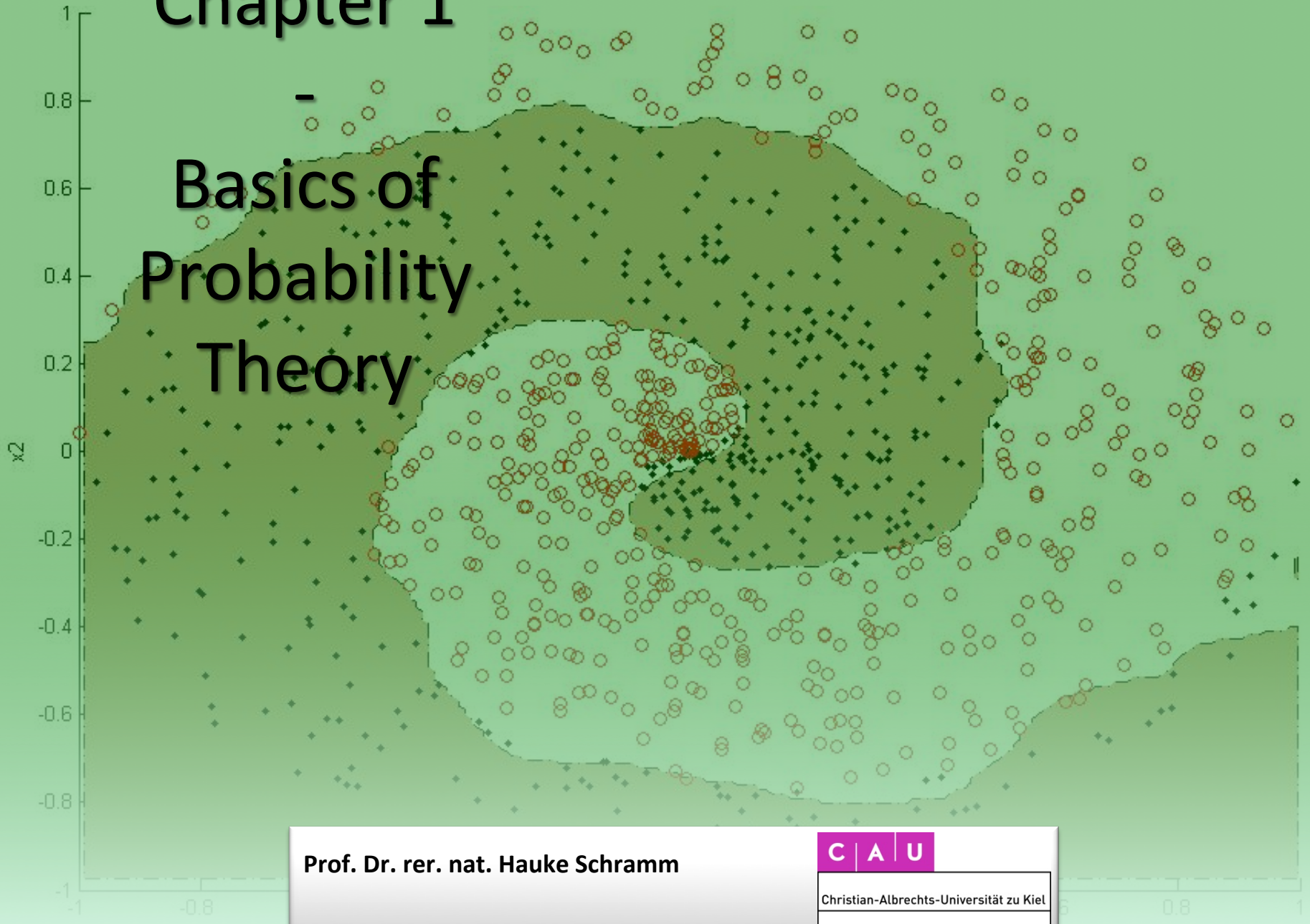
Chapter 1
-
Basics of Probability Theory



Prof. Dr. rer. nat. Hauke Schramm

C A U
Christian-Albrechts-Universität zu Kiel
Institut für Informatik

Topics of the lecture

Introduction

1. Fundamentals of probability theory

   Probability distribution, expectation, covariance matrix, marginal, multivariate Gaussian, ...

2. Bayesian decision theory

   Bayes theorem, prior-/a-priori-/a-posteriori-probabilities, Bayes risk, decision boundaries, ...

3. Parameter estimation

   Maximum-likelihood, Bayes learning, Distances in $\mathbb{R}^D$

4. Non-parametric techniques

   Density estimation, Parzen windows, nearest neighbor classification

Outline of Chapter 1

1. Basics of probability theory
    1. Discrete random variables
    2. Expected values
    3. Pairs of discrete random variables
    4. Marginal distributions
    5. Statistical independence
    6. Expected values of functions of two variables
    7. Covariance
    8. Conditional probability
    9. Vector random variables
    10. Continuous random variables

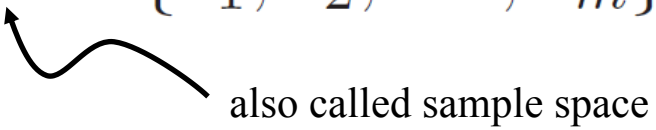
Outline of Chapter 1

1. Basics of probability theory (continuation)
   12. Normal distributions
   13. Mahalanobis distance
   14. Multivariate normal densities
   15. Mixture densities
   16. Bayes rule

1. Basics of probability theory

1.1 Discrete random variables

The discrete random variable x can assume a finite number m of different values in the set: $X = \{v_1, v_2, \ldots, v_m\}$



Probability that x takes the value v_i is denoted by:

$$p_i = Pr(x = v_i), \; i = 1, \ldots, m$$

Probabilities p_i must satisfy the following conditions:

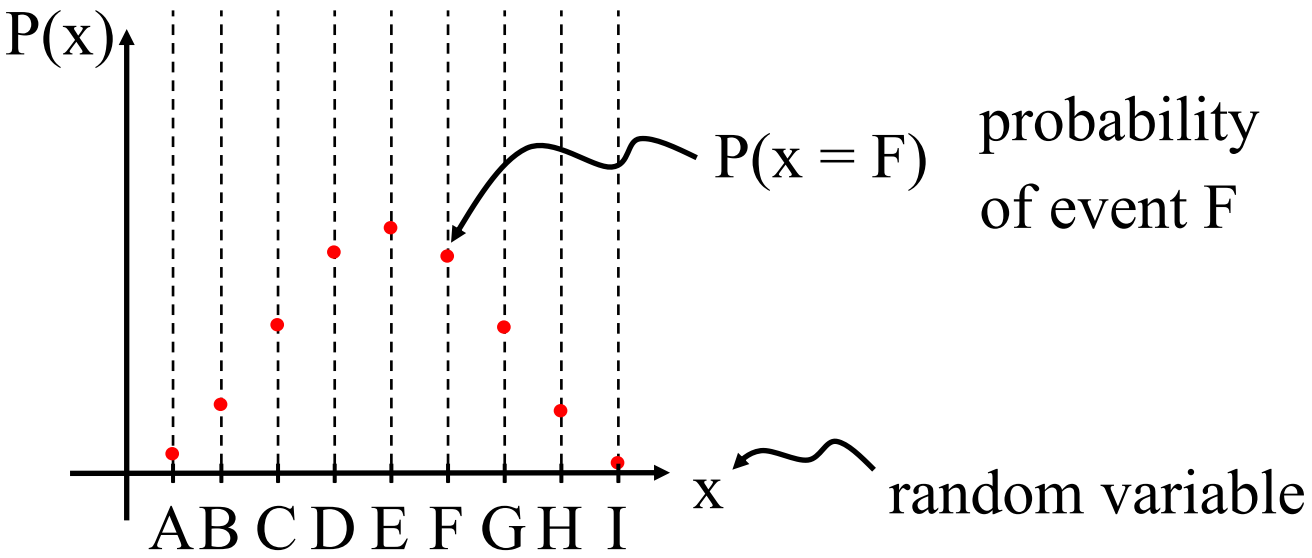
$$p_i \geq 0 \qquad \sum_{i=1}^{m} p_i = 1$$

1.1 Discrete random variables

For convenience set of probabilities $\{p_1, p_2, ..., p_m\}$ is expressed as

probability mass function P(x) which must satisfy the conditions:

$$P(x) \geq 0 \qquad \sum_{x \in X} P(x) = 1 \qquad \sum_{x \notin X} P(x) = 0$$



1.1 Discrete random variables

The set of all realizations that have a strictly positive probability of being observed is called the **support** S_x .

Example:

If the discrete random variable x has the probability mass function

$$P(x) = \begin{cases} 0.5 & \text{if } x = 0 \\ 0.5 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

the support is: $S_x = \{0, 1\}$

1.2 Expected values

The expected value (mean) of a random variable x is defined as:

$$E[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i\, p_i$$

Interpretation: μ is arithmetic average of values in a large random sample

1.2 Expected values

The expected value (mean) of a random variable x is defined as:

$$E[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i\, p_i$$

Example:

Exam results:

1	2	3	4	5
0	1	3	0	0

Usual computation of the mean result of an exam?

Correspondence to the equation above?

1.2 Expected values

The expected value (mean) of a random variable x is defined as:

$$E[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i\, p_i$$

Example:

Exam results:

1	2	3	4	5
0	1	3	0	0

μ = (2 + 3 + 3 + 3) / 4 = 1 · 0 + 2 · ¼ + 3 · ¾ + 4 · 0 + 5 · 0

μ = 1 · p(1) + 2 · p(2) + 3 · p(3) + 4 · p(4) + 5 · p(5)

1.2 Expected values

The expected value (mean) of a random variable x is defined as:

$$E[x] = \mu = \sum_{x \in X} x\, P(x) = \sum_{i=1}^{m} v_i\, p_i$$

Generalization to any function f(x):

$$E[f(x)] = \sum_{x \in X} f(x) P(x)$$

expected value of function f

1.2 Expected values

Variance and standard deviation:

$$Var[x] = \sigma^2 = E[(x-\mu)^2] = \sum_{x \in X} (x-\mu)^2 P(x)$$

Variance

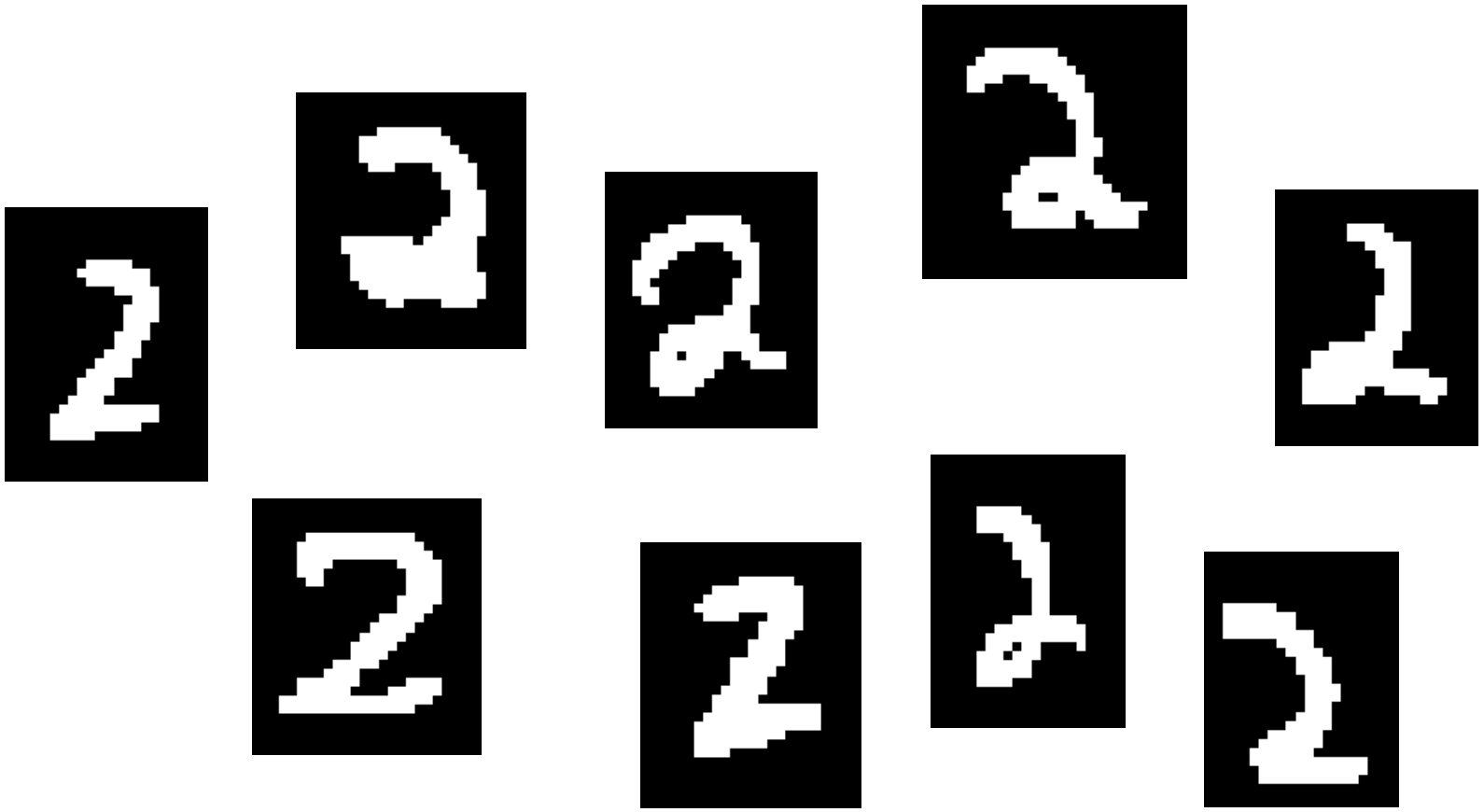
Standard deviation σ

Simple but effective measure of how far values of x are likely to depart from the mean value.

1.2 Expected values

Example for practical computation of mean and variance:

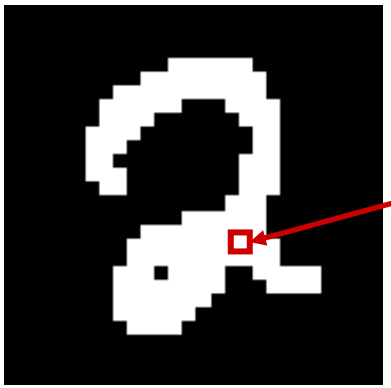
Consider a data set of 1000 handwritten digits of '2'



1.2 Expected values

Example for practical computation of mean and variance:

- Each digit image has 28 x 28 = 784 pixels
- Each pixel can be considered as a single random variable



Each pixel has gray values between 0 and 255

1.2 Expected values

Goal:

- Compute the mean for each pixel and show result as new image
- Do the same for variance

1.2 Expected values

Goal:

- Compute the mean for each pixel and show result as new image
- Do the same for variance

Example for pixel 140:

$$E[x_{140}] = \sum_{x=0}^{255} x \cdot P(x) = 0 \cdot P(0) + 1 \cdot P(1) + 2 \cdot P(2) + \ldots$$

Gray value 1

Probability of gray value 1 in pixel 140

1.2 Expected values

Goal:

- Compute the mean for each pixel and show result as new image
- Do the same for variance

Example for pixel 140:

$$E[x_{140}] = \sum_{x=0}^{255} x \cdot P(x) = 0 \cdot P(0) + 1 \cdot P(1) + 2 \cdot P(2) + \ldots$$

Gray value 1

Probability of gray value 1 in pixel 140

Practically:

For each pixel p:

Step 1: Sum gray values of pixel p in all 1000 images

Step 2: Divide by 1000

1.2 Expected values

Use the resulting mean values as new gray values:

mean matrix

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	3	4	5	4	4	4	4	2	1	1	0	0	0	0
0	0	0	1	3	6	12	21	29	38	43	43	41	32	24	16	9	4	1	0	0
0	0	0	5	12	23	40	61	81	102	118	129	128	114	89	60	36	19	8	2	0
0	0	2	11	25	45	71	100	126	148	164	173	172	165	145	109	73	42	17	6	1
0	2	6	19	37	60	87	113	135	149	158	164	167	169	161	139	100	63	29	11	2
0	3	11	24	43	65	86	105	116	118	116	118	127	141	149	141	115	75	39	15	3
1	4	11	22	38	58	72	83	83	78	72	74	89	112	136	141	120	82	44	18	3
1	3	8	17	29	42	52	57	55	47	42	45	65	99	133	143	122	85	45	19	3
1	2	5	10	17	26	31	33	30	26	23	34	60	98	136	142	120	80	42	17	3
0	1	3	6	10	14	18	18	16	15	21	37	71	111	140	137	111	70	36	14	3
0	1	2	4	7	9	11	12	13	17	31	56	93	125	140	128	96	59	28	11	4
1	1	2	4	7	11	15	19	24	36	57	90	117	137	137	117	84	48	23	9	3
1	1	4	9	15	23	33	45	59	80	106	129	141	142	130	104	69	38	19	9	5
1	3	10	21	35	50	68	87	111	134	153	160	156	143	124	95	62	37	22	13	9
2	8	21	42	62	81	105	128	150	170	181	176	162	143	122	95	71	50	33	25	18
4	15	36	63	86	110	131	148	163	176	177	171	160	146	126	106	89	70	56	43	31
6	23	49	80	108	128	142	154	166	171	170	165	157	146	135	125	112	97	80	62	44
8	27	58	91	118	142	157	170	175	172	164	154	144	138	135	134	126	116	95	73	49
7	28	58	92	124	154	173	182	178	169	153	138	125	118	121	123	121	108	89	66	40
6	23	48	78	113	141	162	168	160	145	126	108	96	91	92	91	87	74	59	42	26
3	12	28	48	74	101	119	122	112	98	82	70	60	53	50	49	45	38	30	21	12
1	4	10	18	29	43	49	52	49	42	33	26	19	16	16	17	16	14	11	8	4
0	0	1	3	3	5	5	6	7	6	5	4	3	3	2	2	3	2	2	1	0
0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

1.2 Expected values

Use the resulting mean values as new gray values:

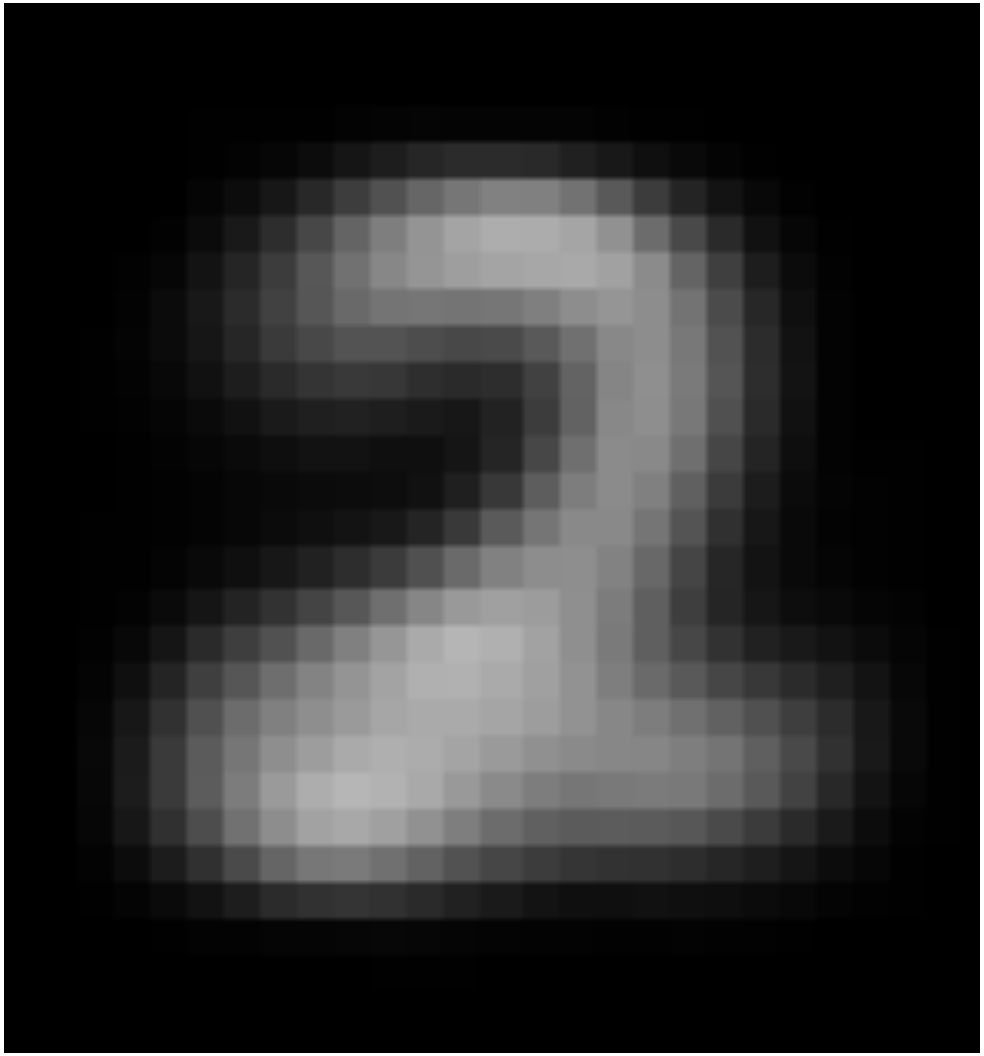


illustration as new image

1.2 Expected values

Variance image:

1.3 Pairs of discrete random variables

Let x and y be random variables which can take on values in:

$$X = \{v_1, v_2, ..., v_m\} \qquad Y = \{w_1, w_2, ..., w_n\}$$

Each pair of events (v_i, w_j) has a *joint probability*

$$p_{ij} = Pr(x = v_i, y = w_j)$$

with

$$p_{ij} \geq 0 \qquad \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} = 1$$

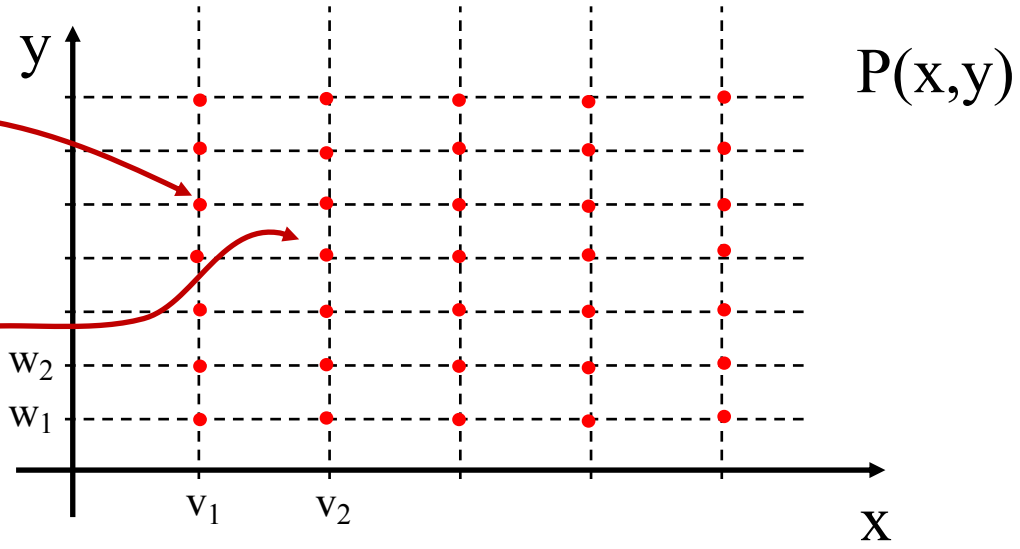
1.3 Pairs of discrete random variables

We can define a **joint probability mass function** P(x, y) for which

$$P(x,y) \geq 0 \qquad \sum_{x \in X} \sum_{y \in Y} P(x,y) = 1$$

Each point (v_i, w_i) is assigned a probability value P(x=v_i, y=w_i)

Note the discrete nature of the random variables.



1.3 Pairs of discrete random variables

Example:

Let x and y be random variables with two possible values (events):

$X = \{v_1, v_2\}$ and $Y = \{w_1, w_2\}$

v_1 : rain v_2 : no rain

w_1 : sunshine w_2 : no sunshine

Let also be P(x, y) the known joint probability mass function.

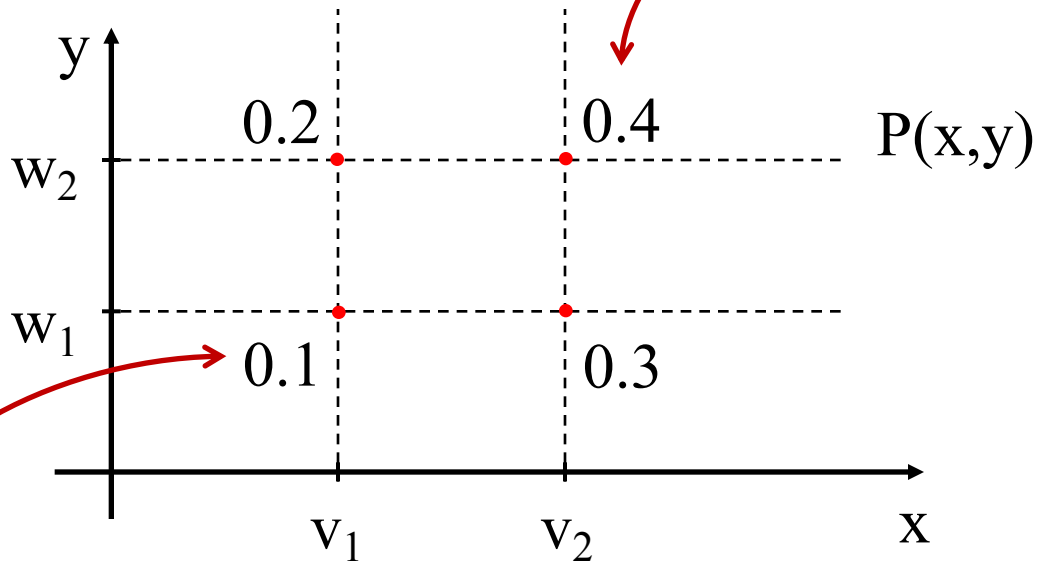
1.3 Pairs of discrete random variables

v_1 : rain v_2 : no rain

w_1 : sunshine w_2 : no sunshine

Illustration of P(x, y)

Combination of no rain and no sunshine is likely.



Combination of rain and sunshine is unlikely.

$$P(x, y) \geq 0$$

$$\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

1.4 Marginal distributions

Let P(x, y) be a known joint probability mass function

The **marginal distribution** for x is given by summing over all y values:

$$P_x(x) = \sum_{y \in Y} P(x, y)$$

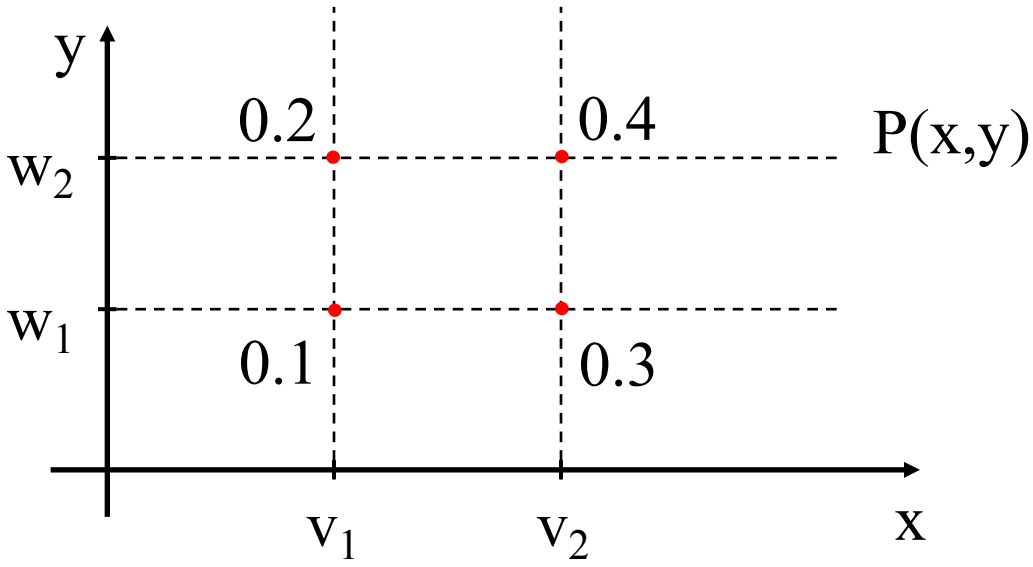
Remark: Subscript is used to emphasize that $P_x()$ has a different functional form than $P_y()$.

The marginal distribution for y is given by summing over all x values:

$$P_y(y) = \sum_{x \in X} P(x, y)$$

1.4 Marginal distributions

Example: Marginal distribution of x

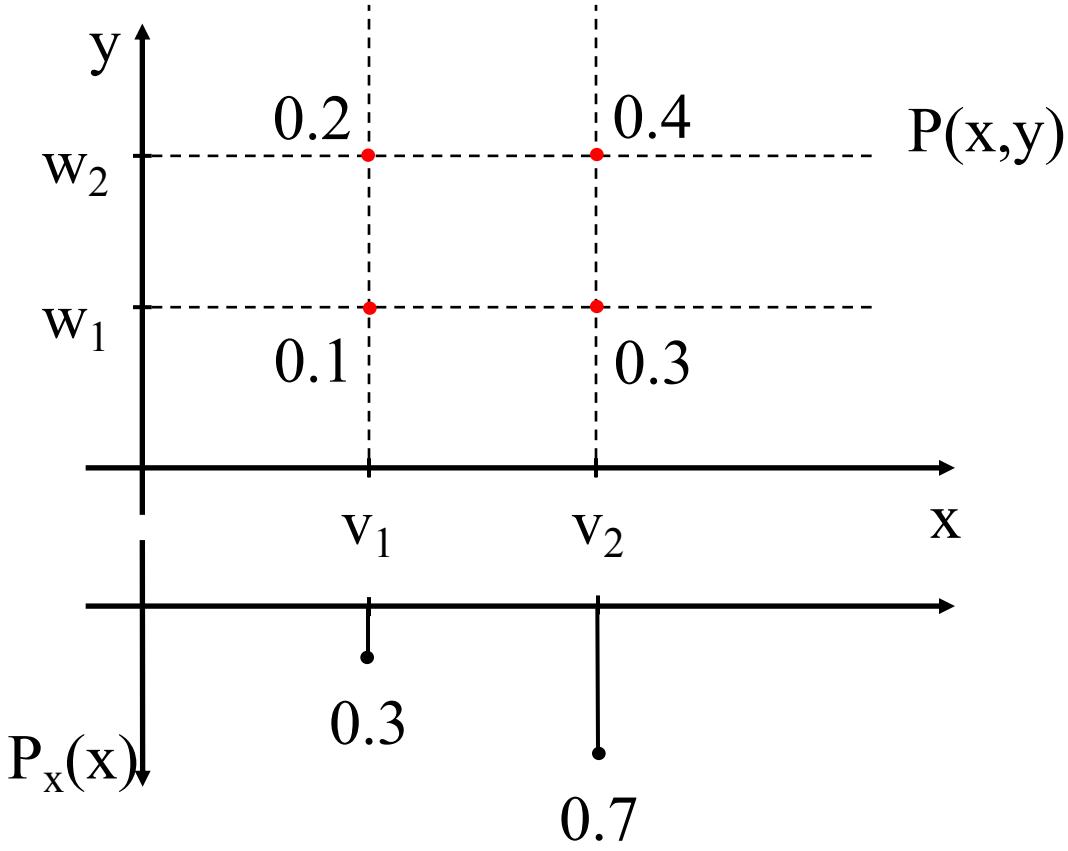


$$P_x(v_1) = \sum_{\text{all events in Y}} P(v_1,y) = P(v_1,w_1) + P(v_1,w_2) = 0.1 + 0.2$$

$$P_x(v_2) = \sum_{\text{all events in Y}} P(v_2,y) = P(v_2,w_1) + P(v_2,w_2) = 0.3 + 0.4$$

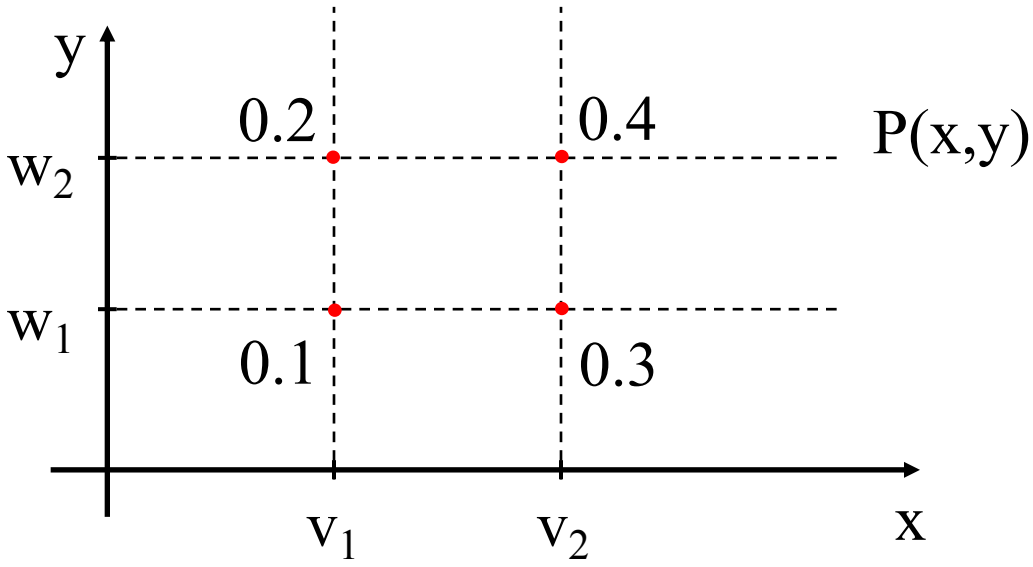
1.4 Marginal distributions

Example: Marginal distribution of x



1.4 Marginal distributions

Example: Marginal distribution of y

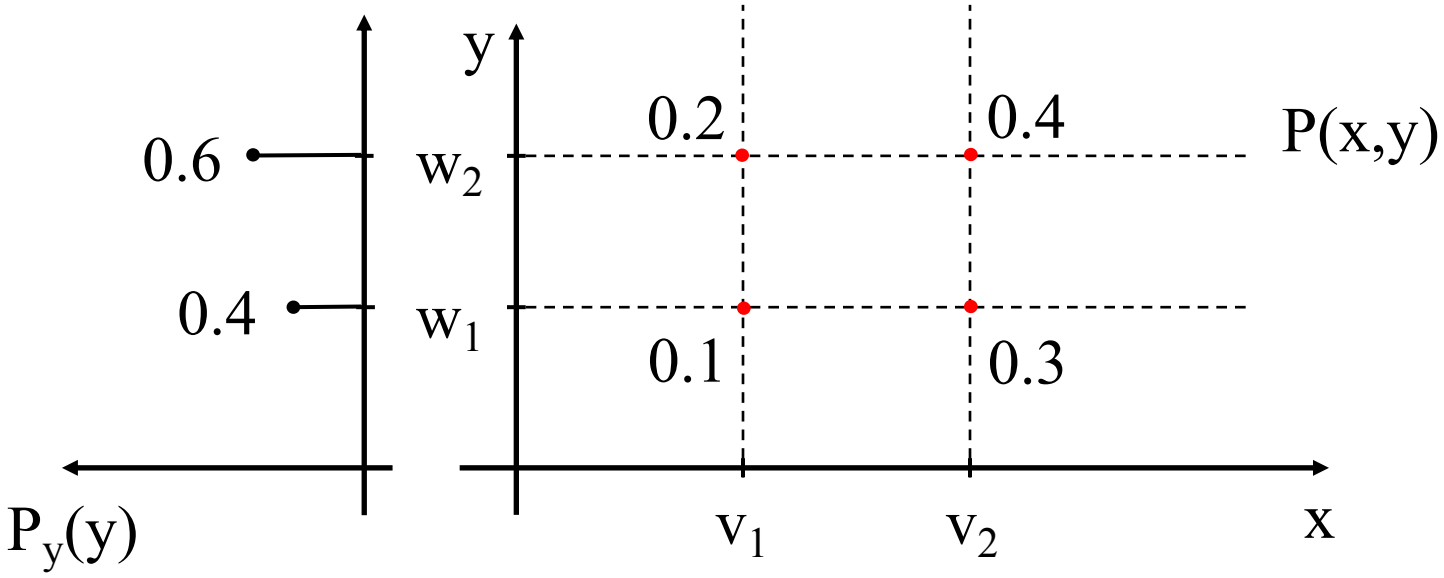


$$P_y(w_1) = \sum_{\text{all events in X}} P(x, w_1) = P(v_1, w_1) + P(v_2, w_1) = 0.1 + 0.3$$

$$P_y(w_2) = \sum_{\text{all events in X}} P(x, w_2) = P(v_1, w_2) + P(v_2, w_2) = 0.2 + 0.4$$

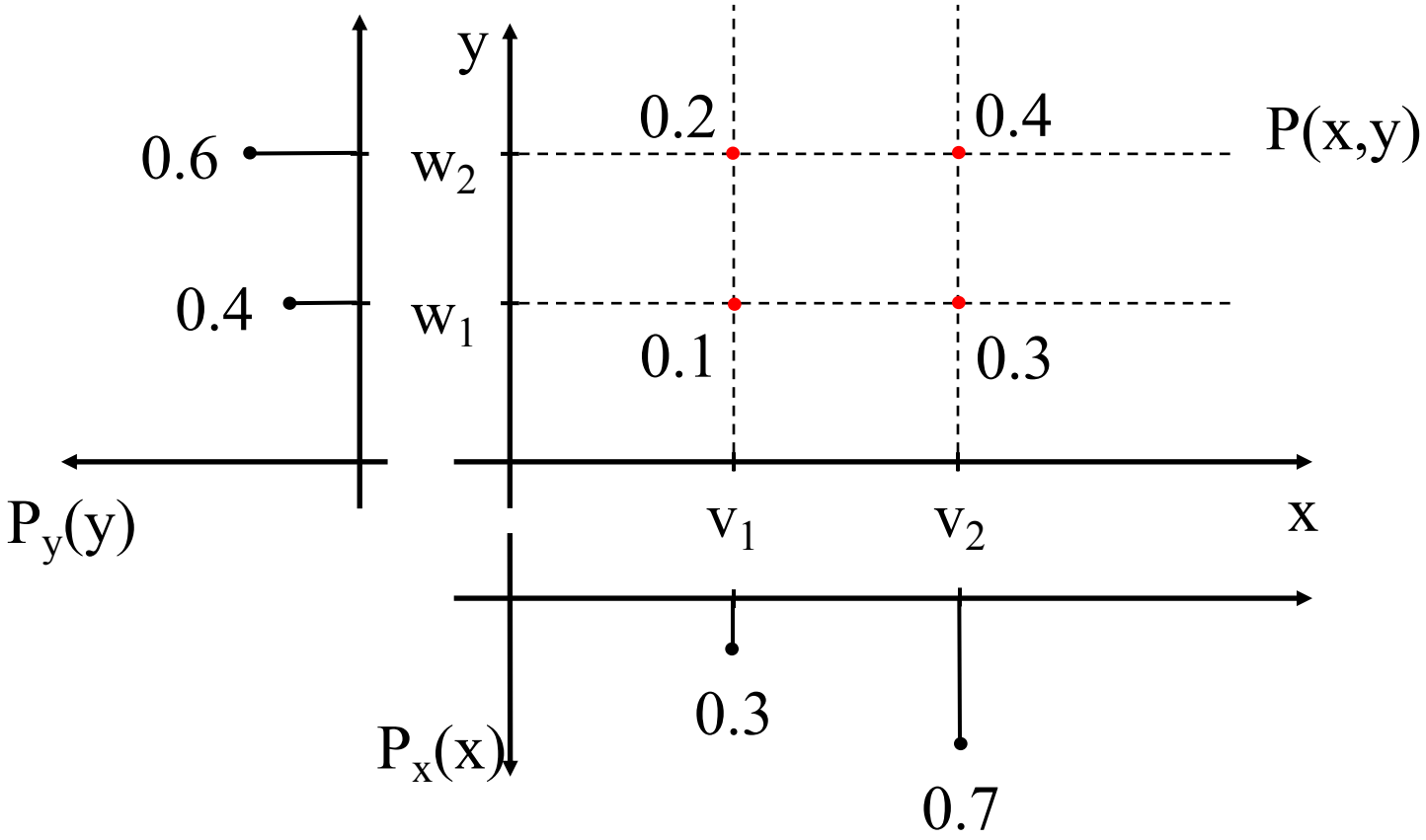
1.4 Marginal distributions

Example: Marginal distribution of y



1.4 Marginal distributions

Example: Marginal distribution of x and y



1.5 Statistical independence

Variables x and y are said to be **statistically independent** if and only if:

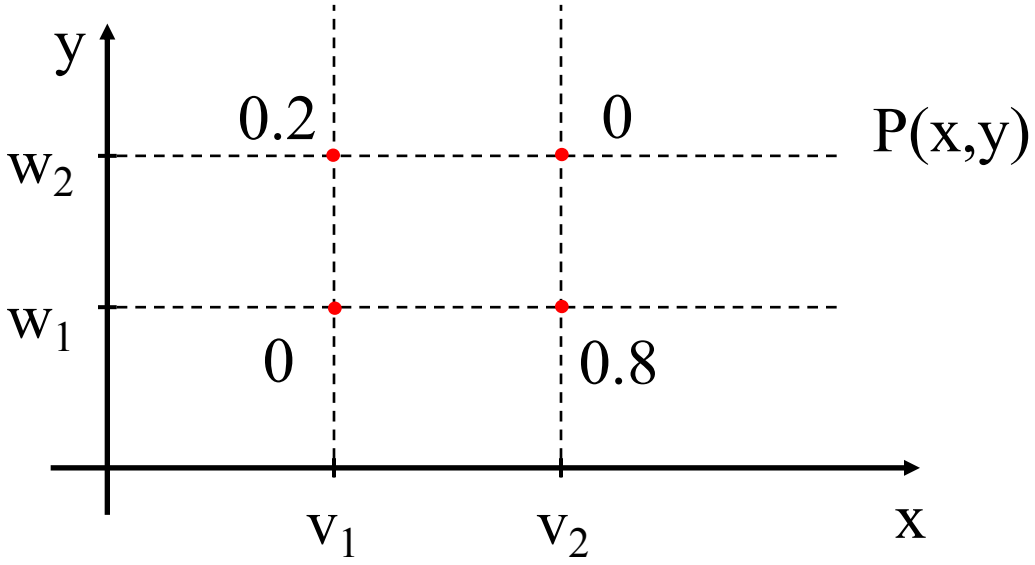
$$P(x, y) = P_x(x)\ P_y(y)$$

Interpretation:

Knowing the value of x does not give us additional knowledge about the possible values of y and vice versa.

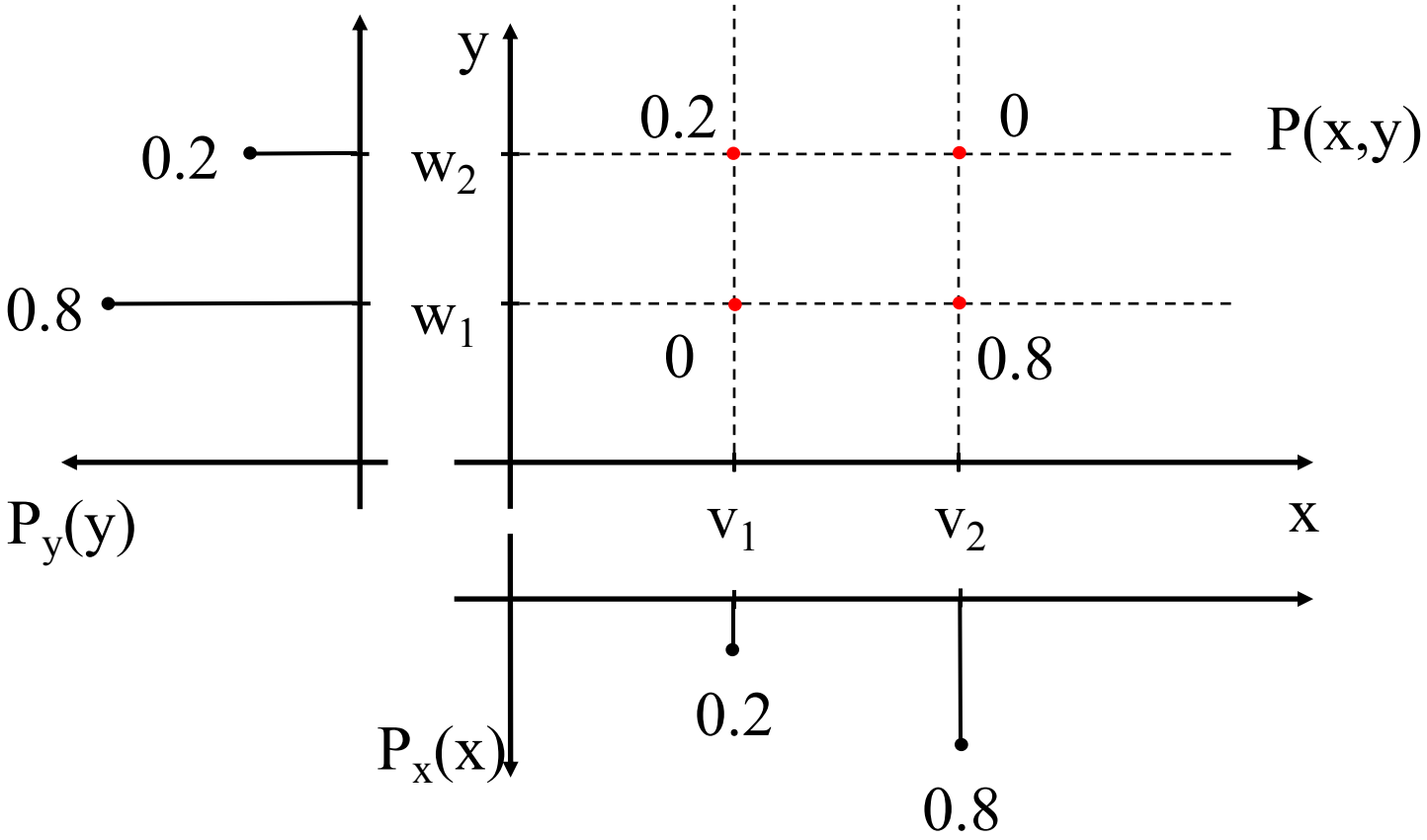
1.5 Statistical independence

Question: Are x and y statistically independent in this example?



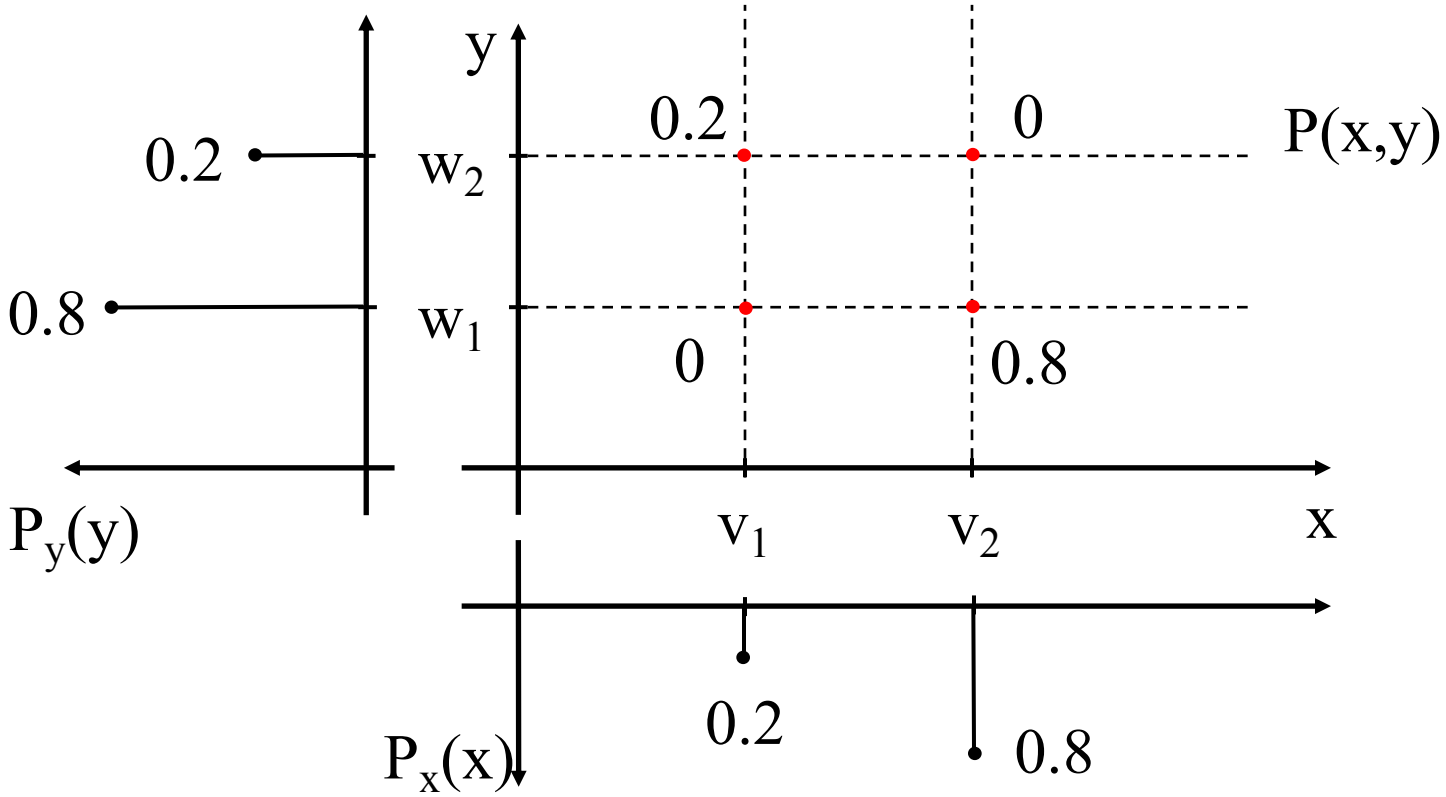
1.5 Statistical independence

Marginal distributions



1.5 Statistical independence

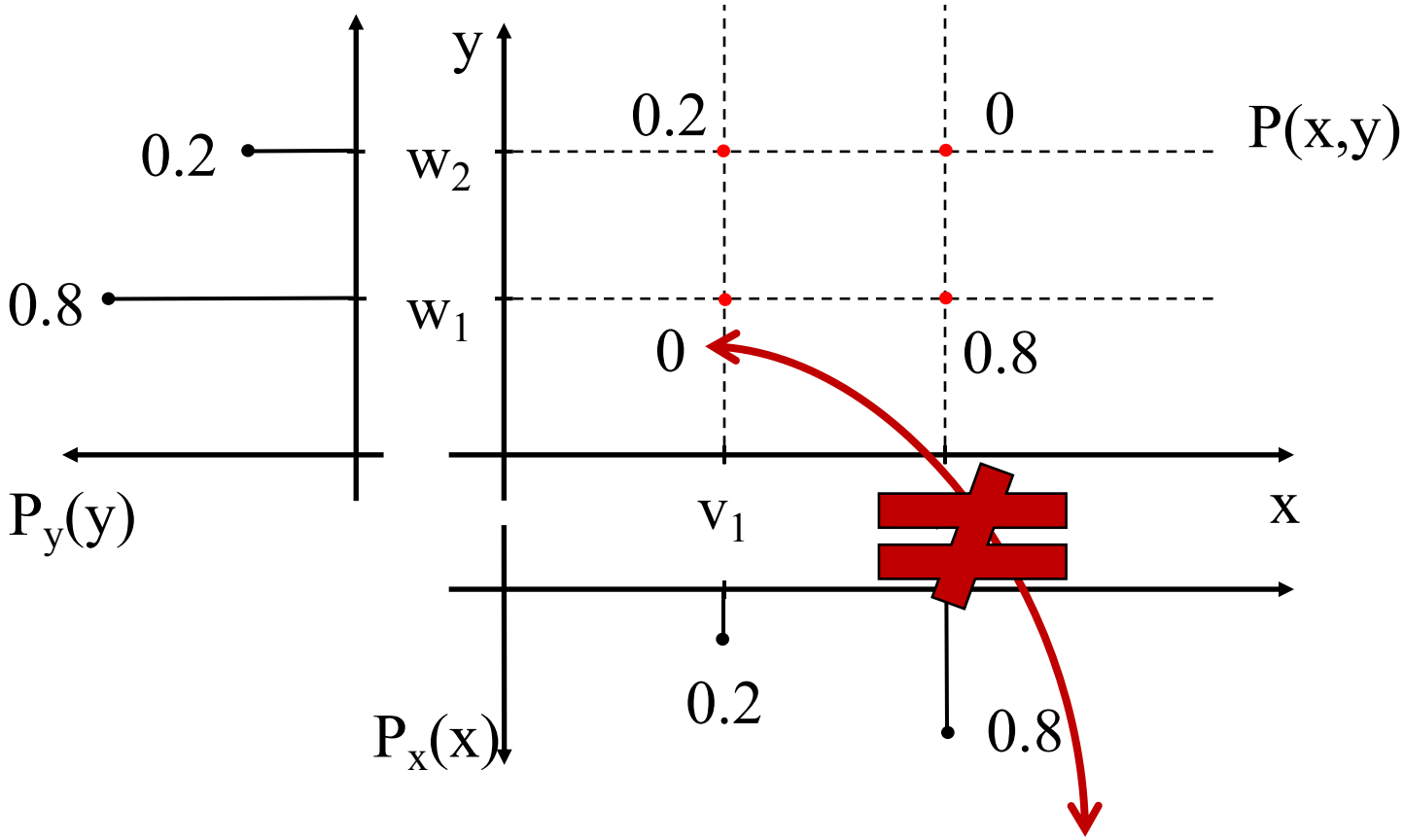
In case of statistical independence: $P(x,y) = P_x(x)\ P_y(y)$



$$P(v_1, w_1) = P_x(v_1) \cdot P_y(w_1) = 0.2 \cdot 0.8 = 0.16$$

1.5 Statistical independence

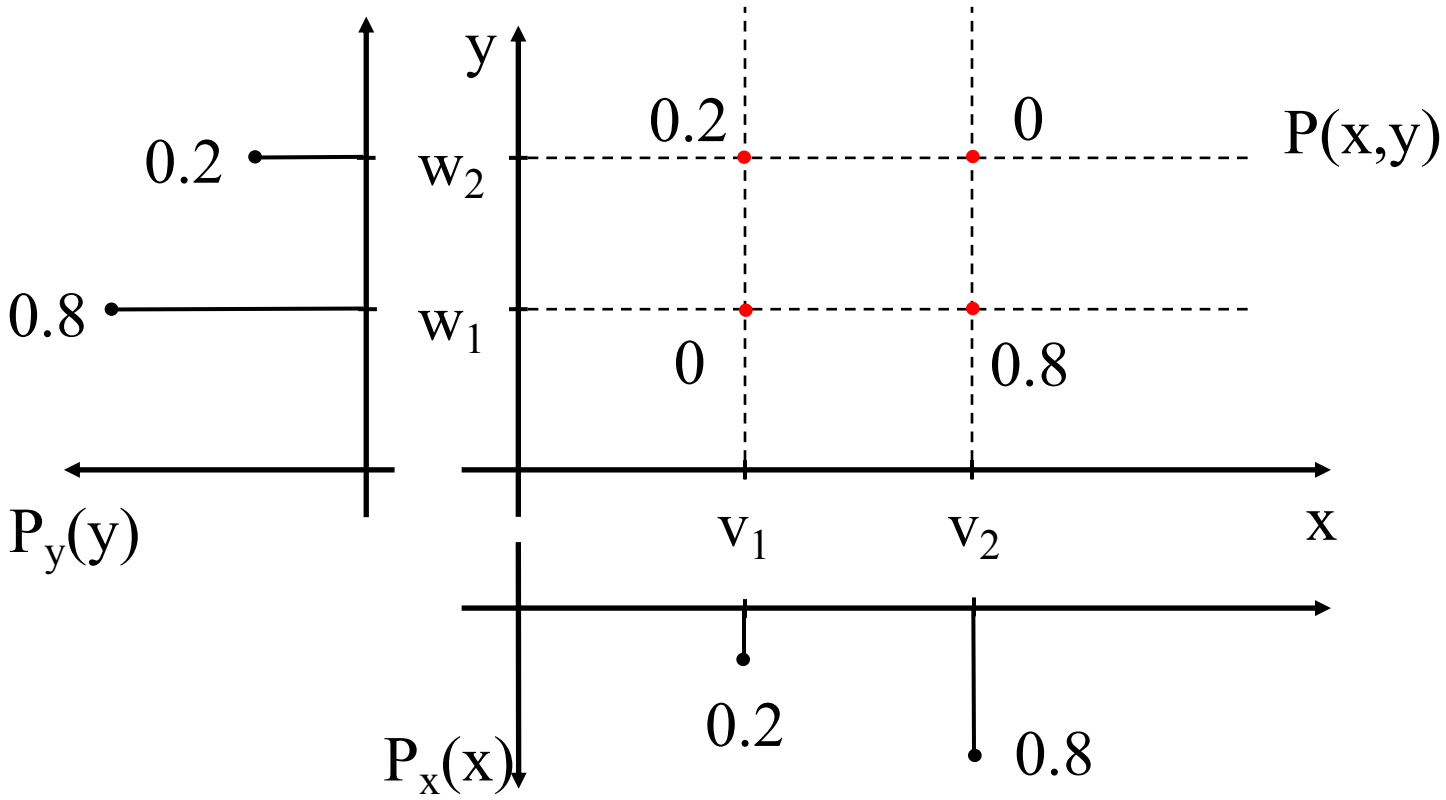
In case of statistical independence: $P(x,y) = P_x(x)\ P_y(y)$



$$P(v_1, w_1) = P_x(v_1) \cdot P_y(w_1) = 0.2 \cdot 0.8 = 0.16$$

1.5 Statistical independence

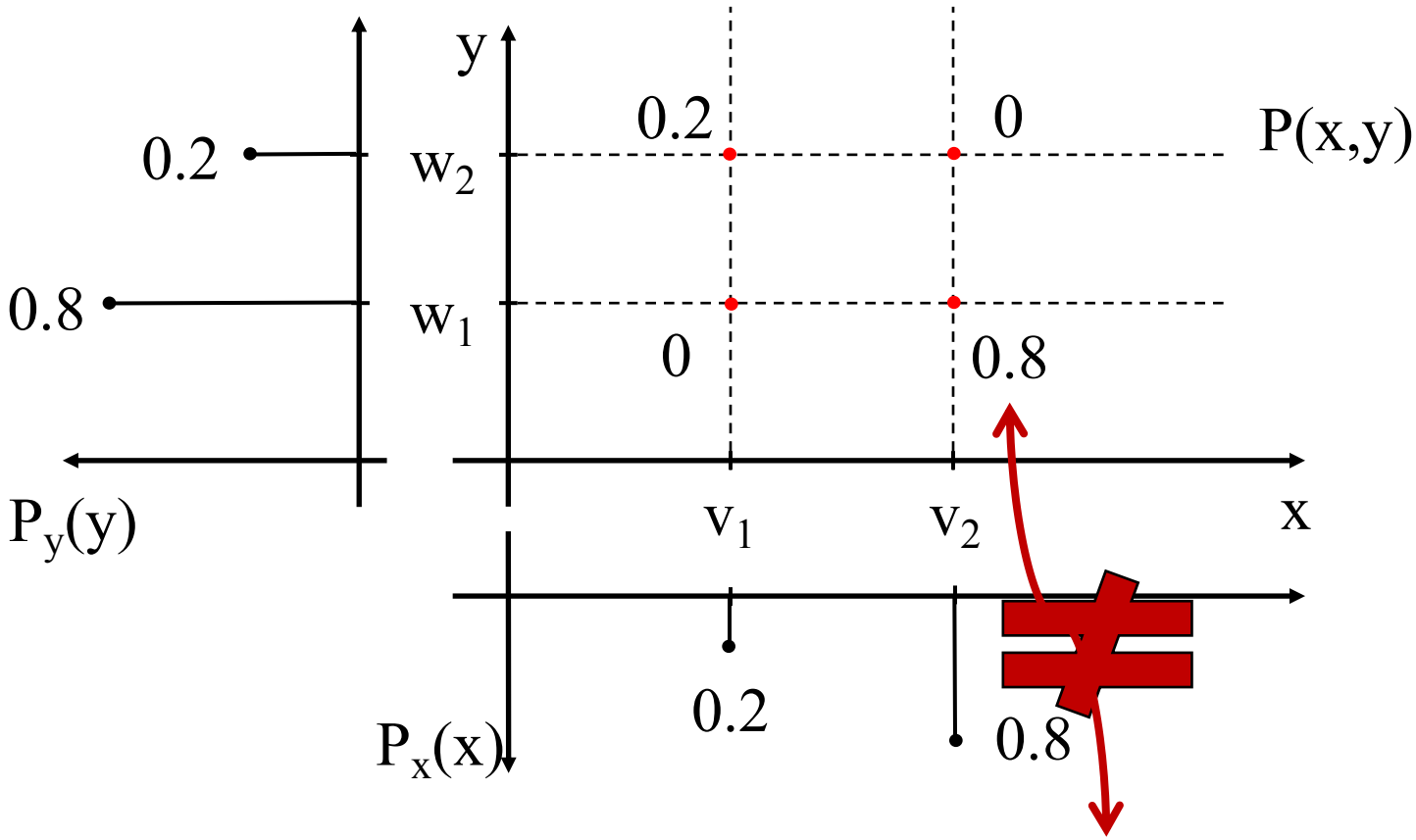
In case of statistical independence: $P(x,y) = P_x(x)\ P_y(y)$



$$P(v_2, w_1) = P_x(v_2) \cdot P_y(w_1) = 0.8 \cdot 0.8 = 0.64$$

1.5 Statistical independence

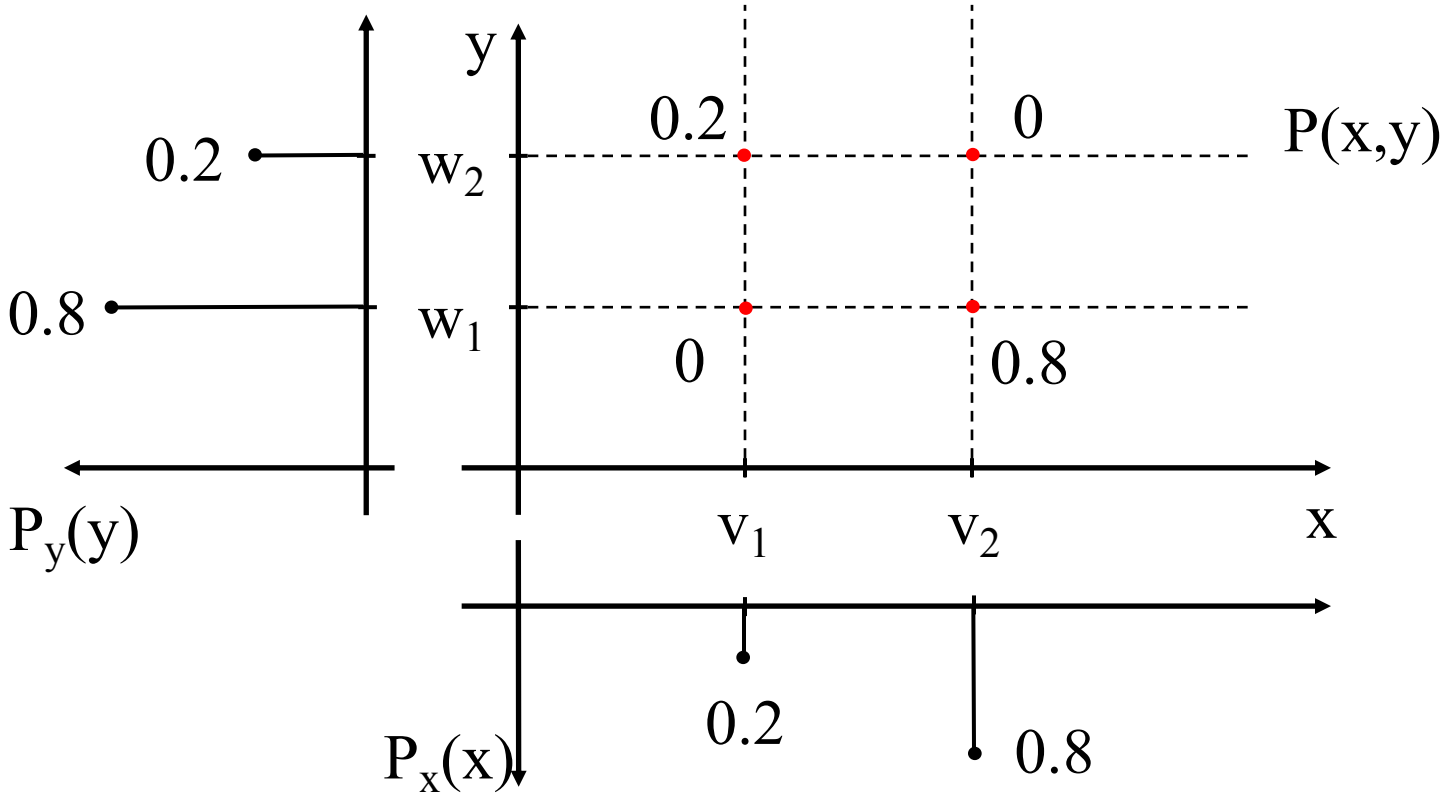
In case of statistical independence: $P(x,y) = P_x(x)\ P_y(y)$



$$P(v_2, w_1) = P_x(v_2) \cdot P_y(w_1) = 0.8 \cdot 0.8 = 0.64$$

1.5 Statistical independence

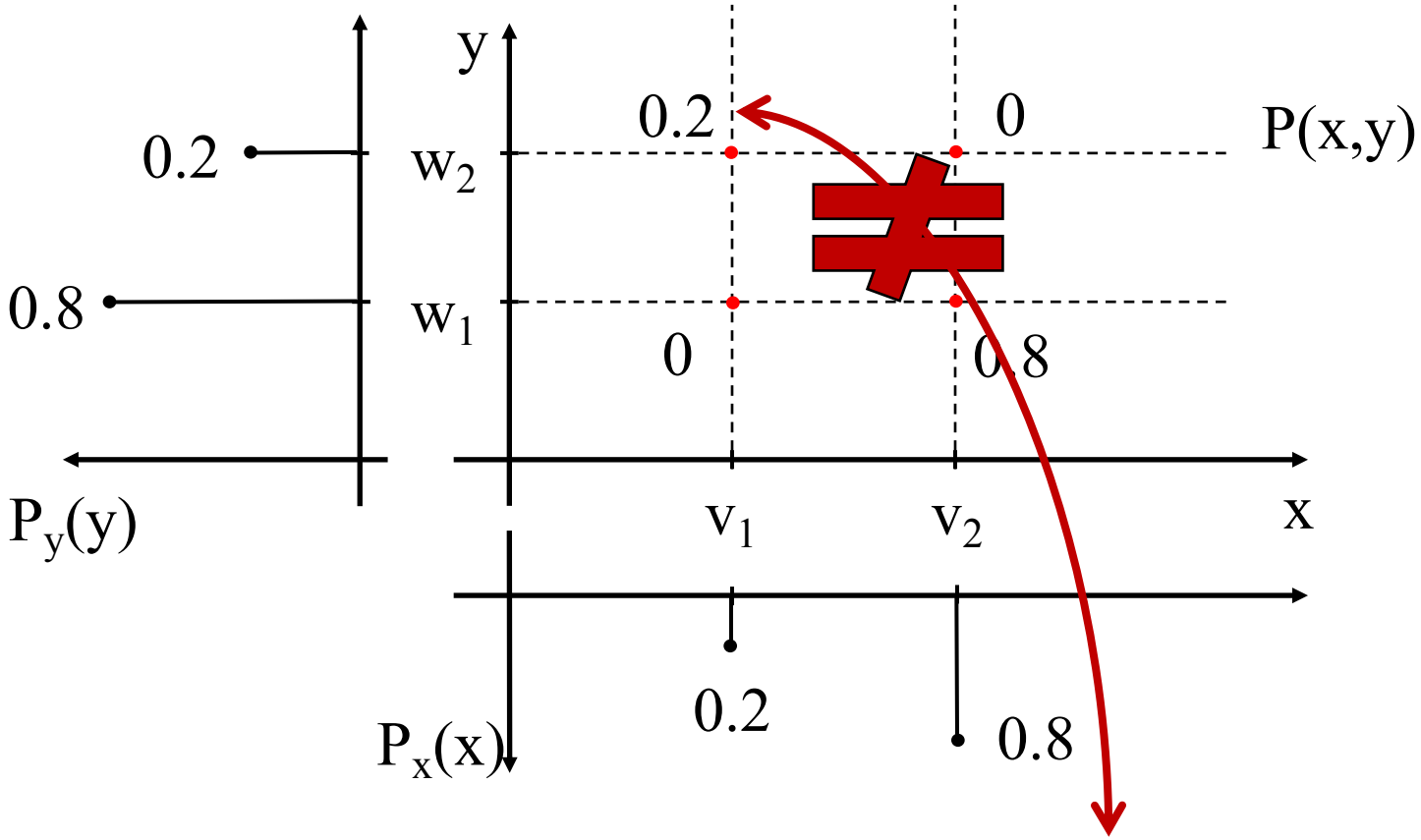
In case of statistical independence: $P(x, y) = P_x(x)\ P_y(y)$



$$P(v_1, w_2) = P_x(v_1) \cdot P_y(w_2) = 0.2 \cdot 0.2 = 0.04$$

1.5 Statistical independence

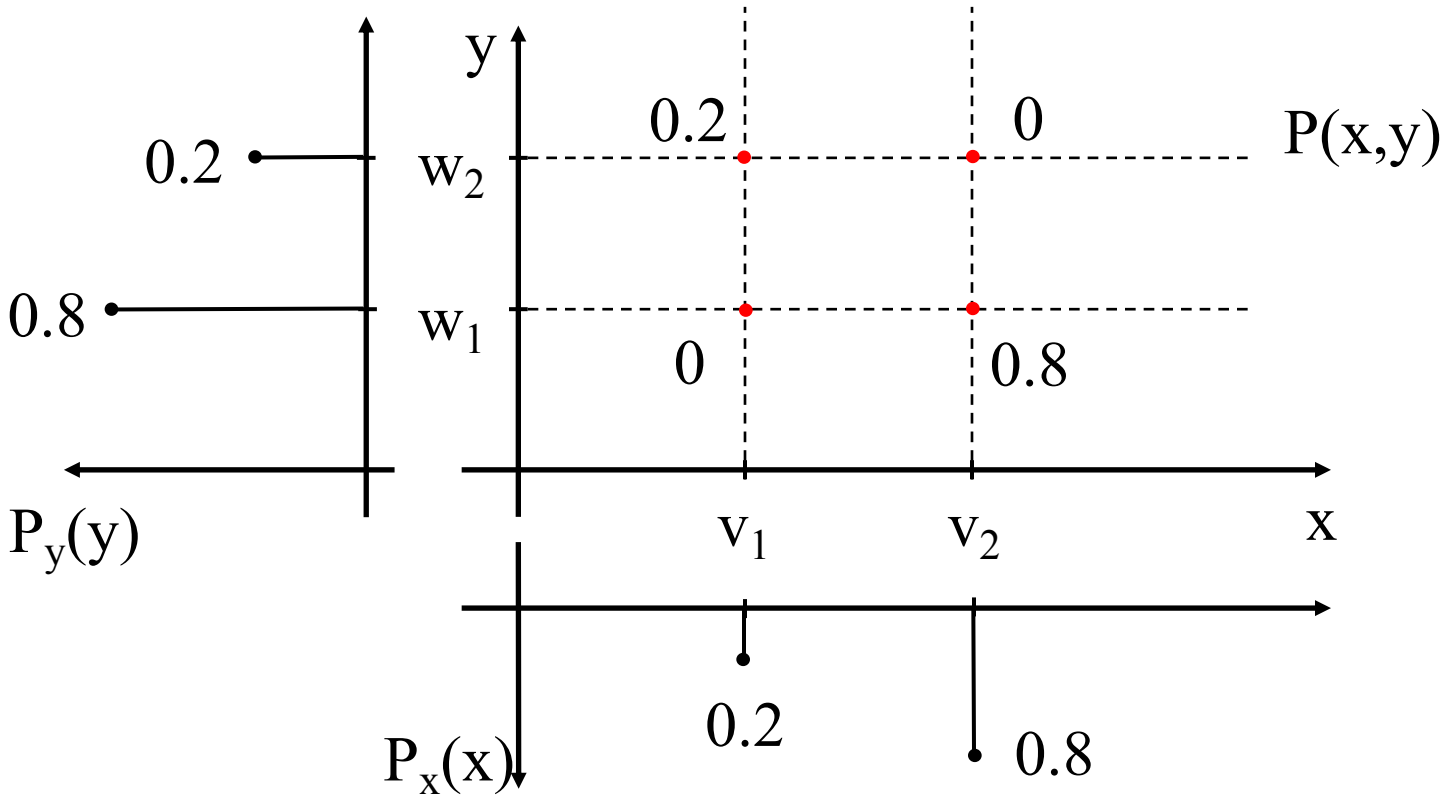
In case of statistical independence: $P(x, y) = P_x(x)\ P_y(y)$



$$P(v_1, w_2) = P_x(v_1) \cdot P_y(w_2) = 0.2 \cdot 0.2 = 0.04$$

1.5 Statistical independence

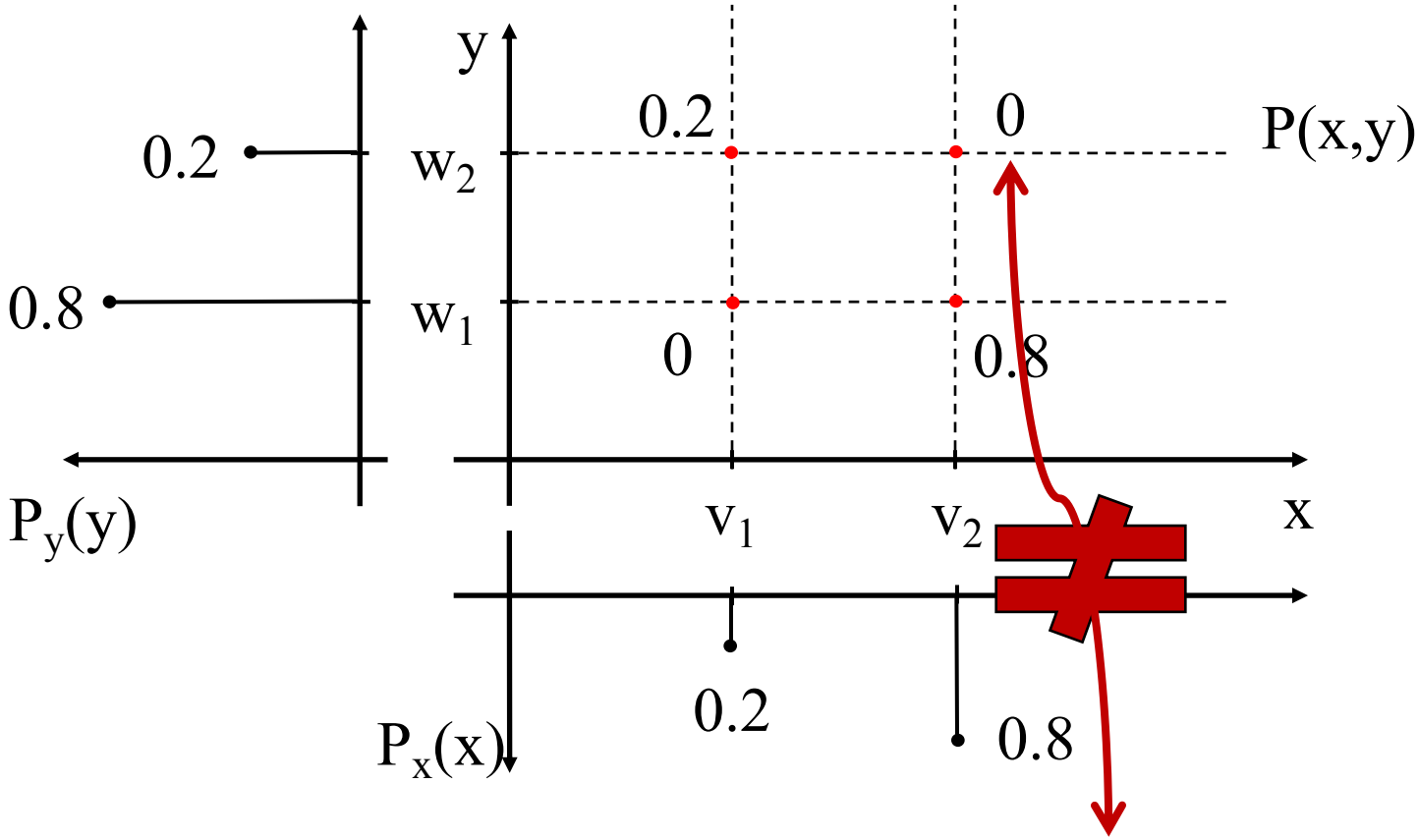
In case of statistical independence: $P(x,y) = P_x(x)\ P_y(y)$



$$P(v_2, w_2) = P_x(v_2) \cdot P_y(w_2) = 0.8 \cdot 0.2 = 0.16$$

1.5 Statistical independence

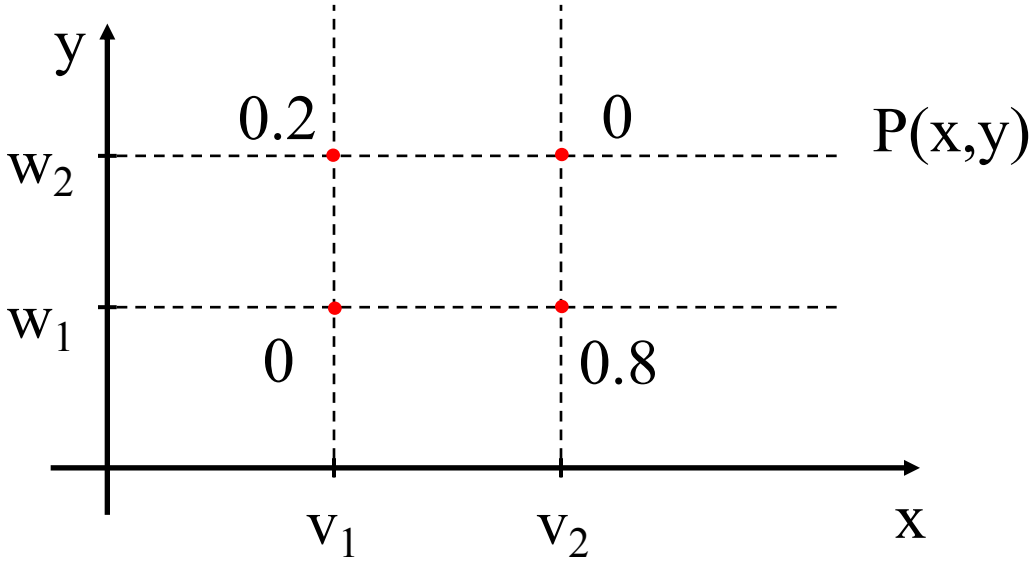
In case of statistical independence: $P(x,y) = P_x(x)\ P_y(y)$



$$P(v_2, w_2) = P_x(v_2) \cdot P_y(w_2) = 0.8 \cdot 0.2 = 0.16$$

1.5 Statistical independence

In case of statistical independence: $P(x, y) = P_x(x)\ P_y(y)$

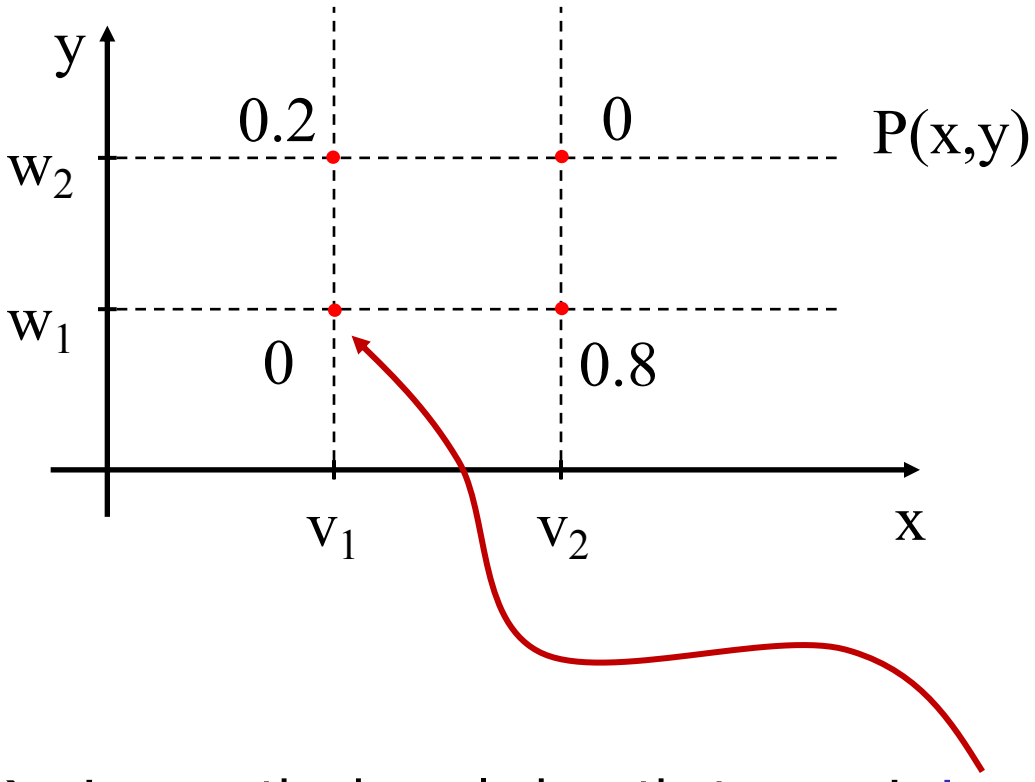


→ x and y are statistically dependent

Knowing the value of x lets us get a better estimate of the values of y and vice versa.

1.5 Statistical independence

In case of statistical independence: $P(x, y) = P_x(x)\ P_y(y)$



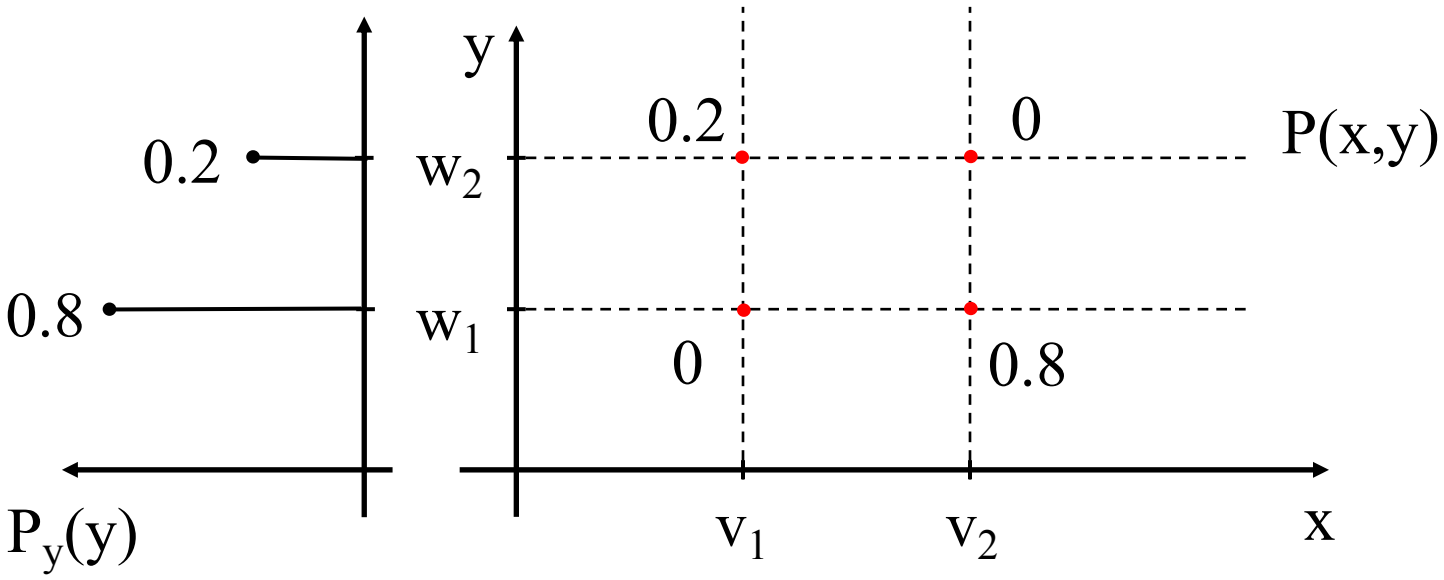
Example:

Let x = v_1 → gives us the knowledge, that y=w_1 is impossible!

Let x = v_2 → gives us the knowledge, that y=w_2 is impossible!

1.5 Statistical independence

In case of statistical independence: $P(x, y) = P_x(x)\ P_y(y)$



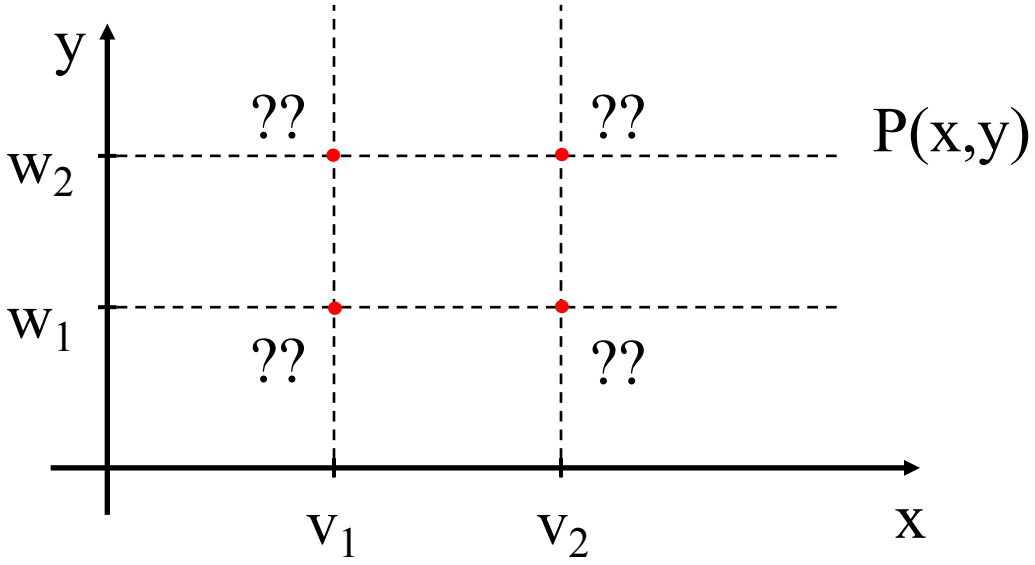
Example:

Without knowledge about the value of x we must use the marginal distribution $P_y(y)$

1.5 Statistical independence

$$P(x, y) = P_x(x)\ P_y(y)$$

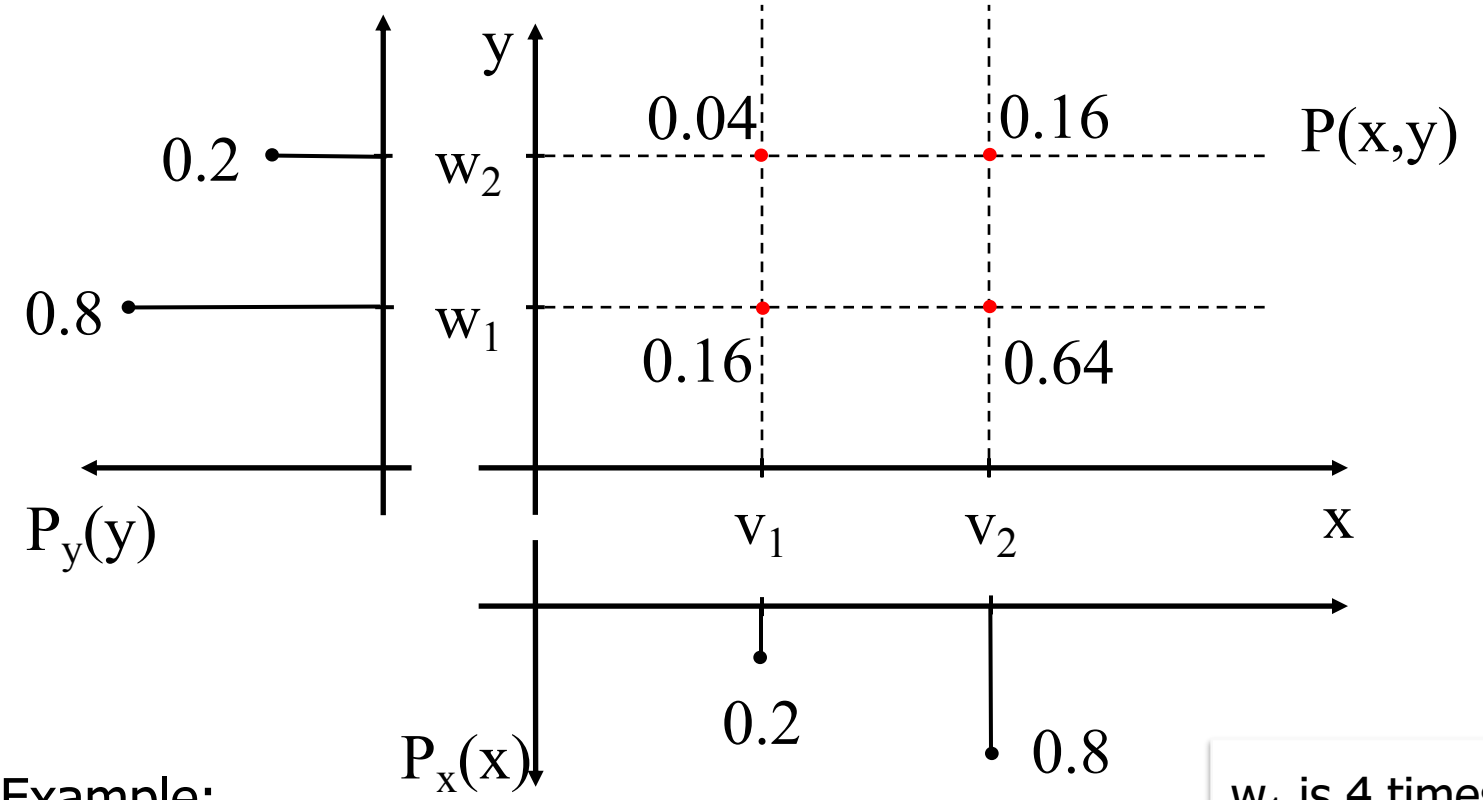
How could a distribution look like in case of statistical independence?



1.5 Statistical independence

$$P(x, y) = P_x(x)\, P_y(y)$$

How could a distribution look like in case of statistical independence?



Example:

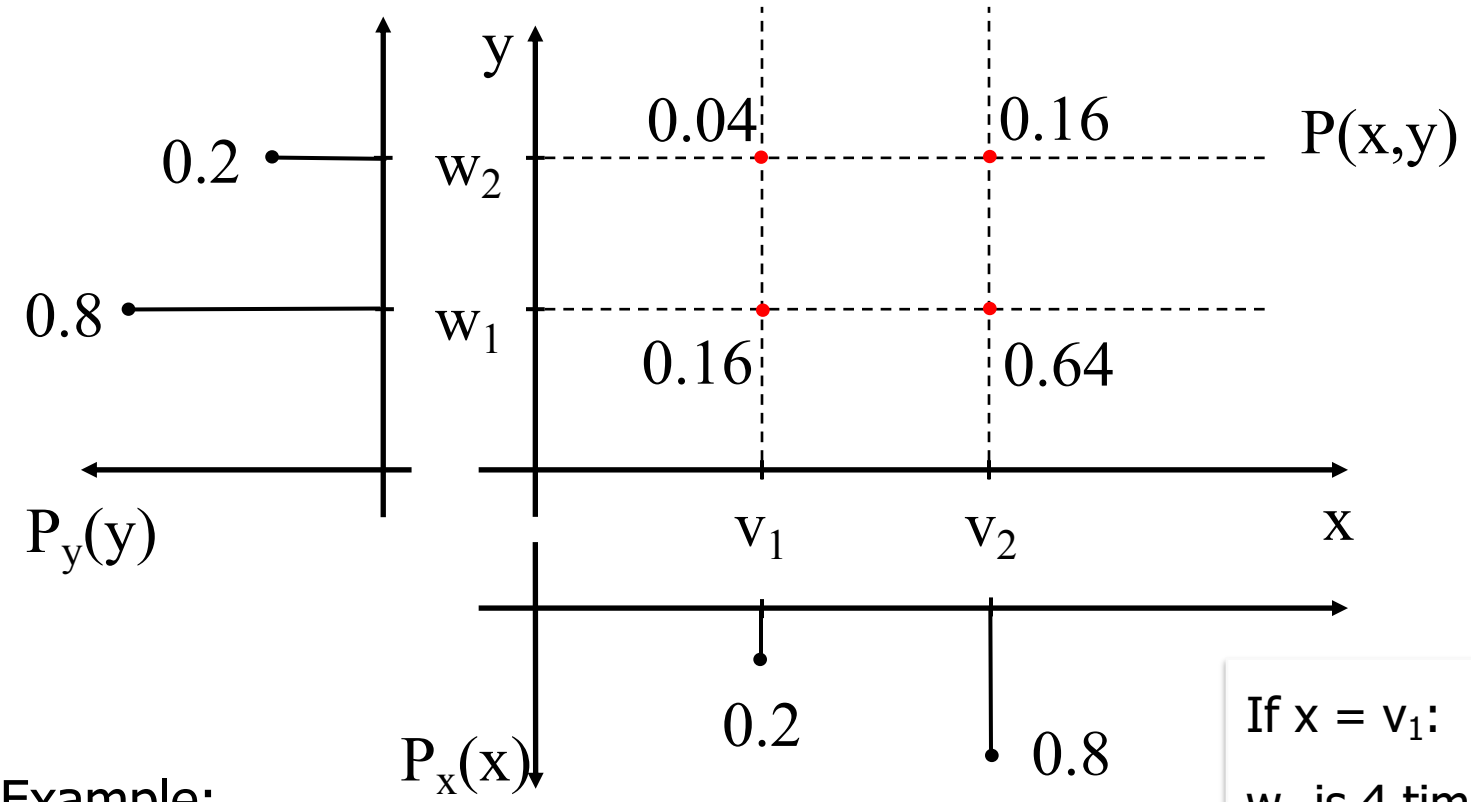
w_1 is 4 times as likely as w_2

Without knowledge of x value we know that: $P(w_1) = 4 \cdot P(w_2)$

1.5 Statistical independence

$$P(x, y) = P_x(x)\, P_y(y)$$

How could a distribution look like in case of statistical independence?



If $x = v_1$:

w_1 is 4 times as likely as w_2

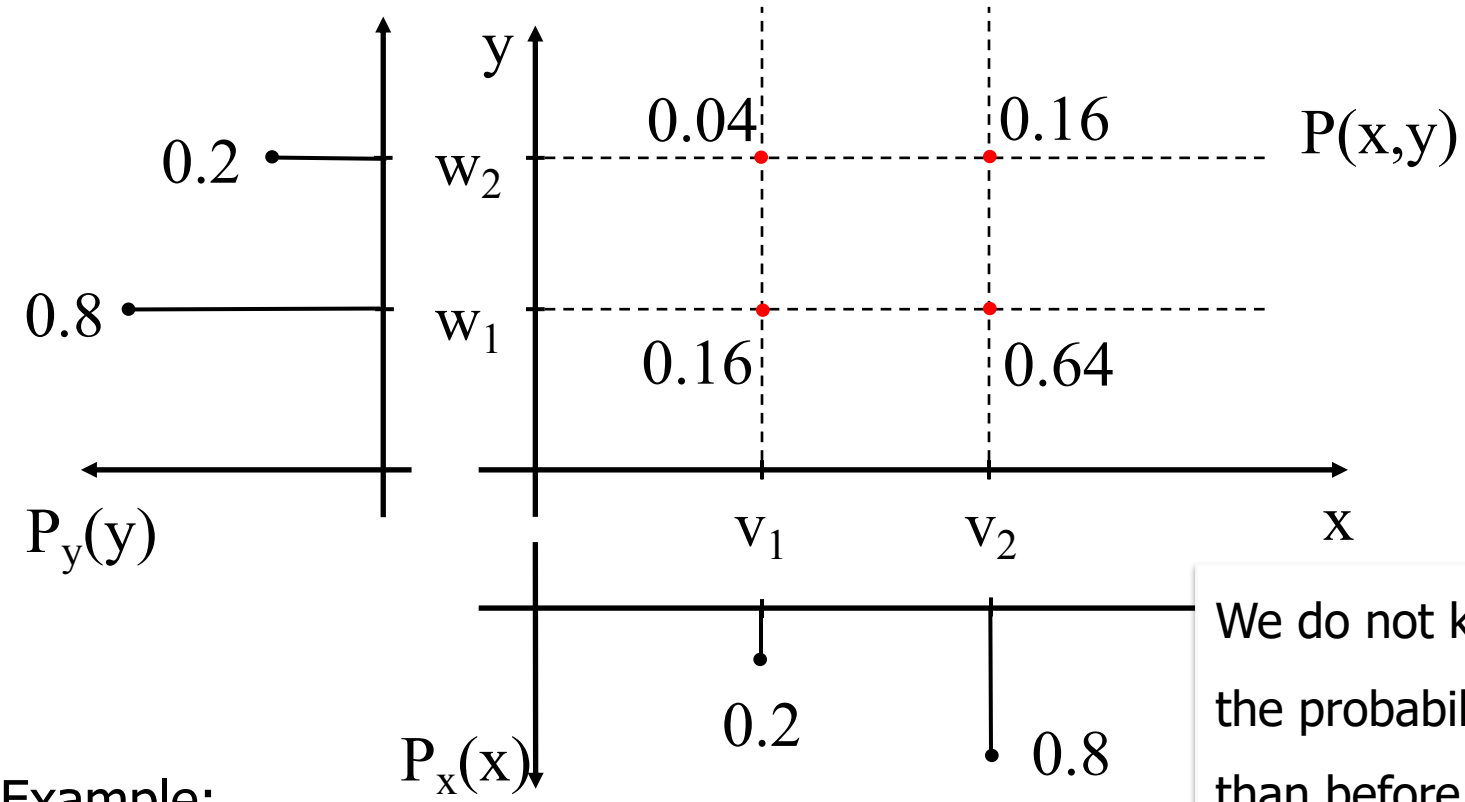
Example:

Let $x = v_1$ → $P(v_1, w_1) = 0.16$ $P(v_1, w_2) = 0.04$

1.5 Statistical independence

$$P(x, y) = P_x(x)\, P_y(y)$$

How could a distribution look like in case of statistical independence?



We do not know more about the probability of w_1 and w_2 than before.

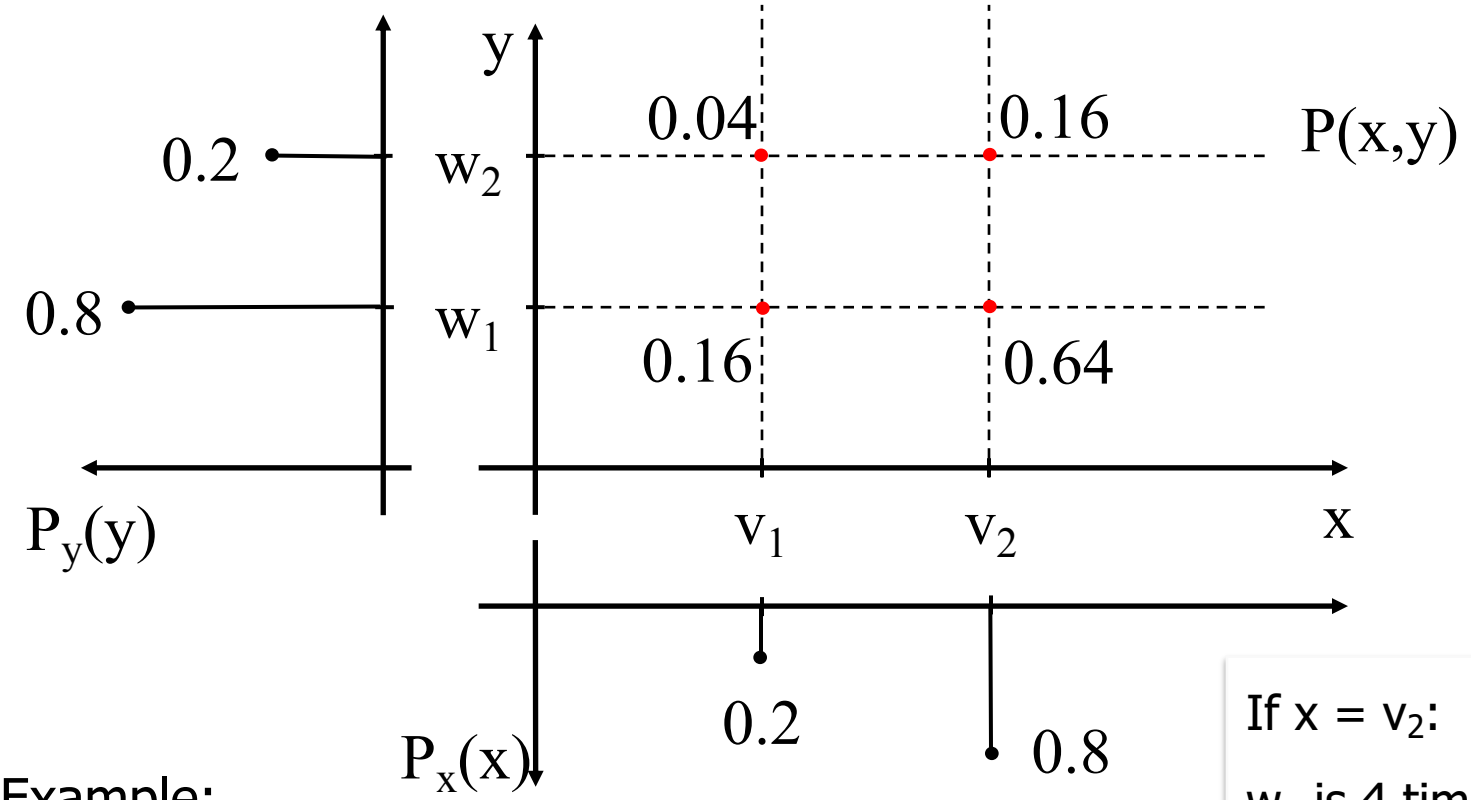
Example:

Let $x = v_1$ → $P(v_1, w_1) = 0.16$ $P(v_1, w_2) = 0.04$

1.5 Statistical independence

$$P(x, y) = P_x(x)\, P_y(y)$$

How could a distribution look like in case of statistical independence?



If $x = v_2$:

w_1 is 4 times as likely as w_2

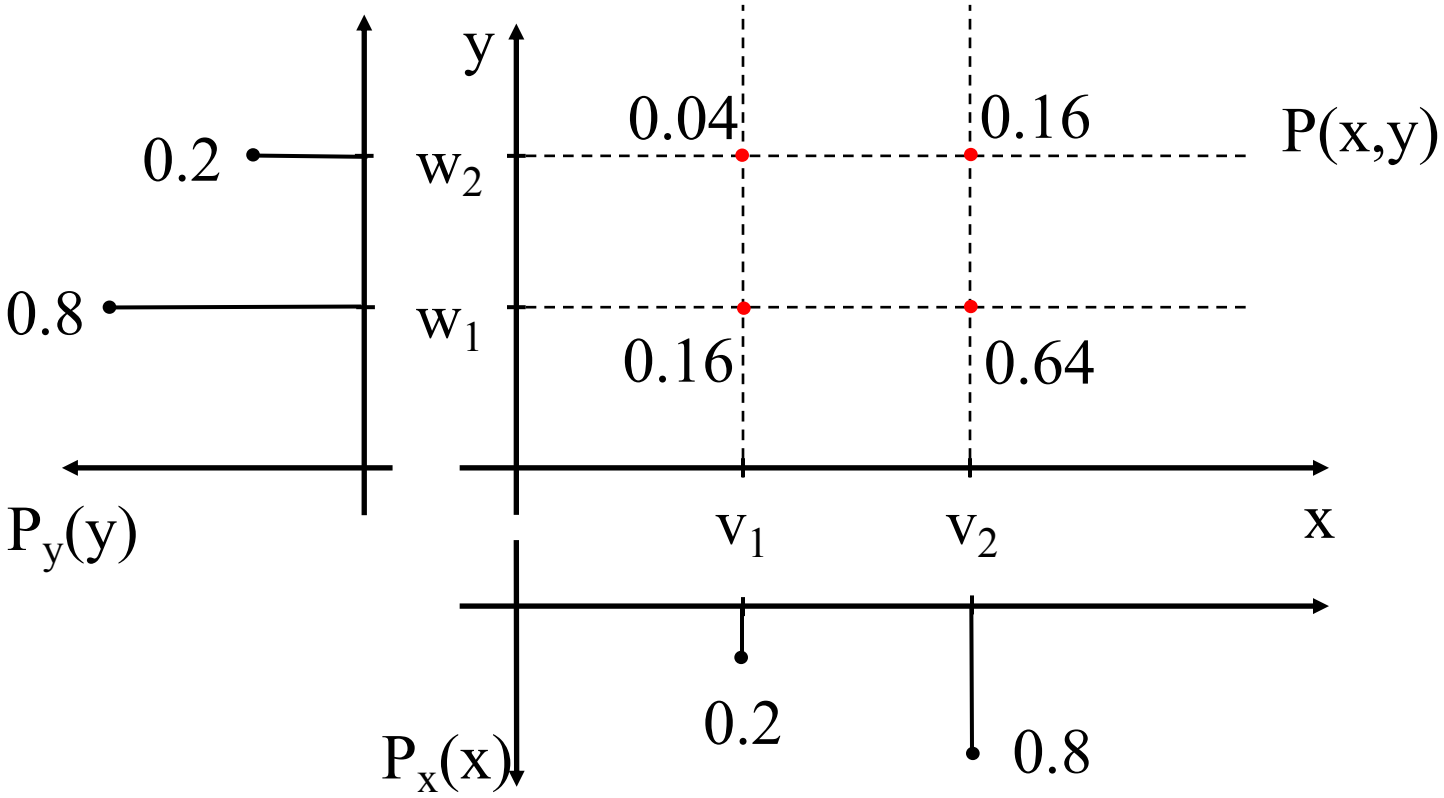
Example:

Let $x = v_2$ → $P(v_2, w_1) = 0.64$ $P(v_2, w_2) = 0.16$

1.5 Statistical independence

$$P(x, y) = P_x(x)\, P_y(y)$$

How could a distribution look like in case of statistical independence?



Knowledge about x does **not** give us additional information about y and vice versa.

1.6 Expected Values of Functions of Two Variables

The expected value of a **function f(x, y)** of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Interpretation 1:

A weighting of the function values at point (x,y) with the probability that this combination of x and y really occurs.

1.6 Expected Values of Functions of Two Variables

The expected value of a **function f(x, y)** of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Interpretation 2:

Function f(x, y) depends upon random variables.

1. Perform experiment → get one observation for x and y
2. Enter values into function and observe corresponding function value
3. Repeat a large number of times, compute mean observed function value

   → expected value of function f(x, y)

1.6 Expected Values of Functions of Two Variables

The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Remark: Result is one scalar expectation value!

Usual notation is somewhat unclear:

- Here, x and y are random variables
  - → Placeholder to show that f depends upon x and y
  - → Do not enter concrete values for x and y here
- Here, concrete values for x and y are used to determine the result

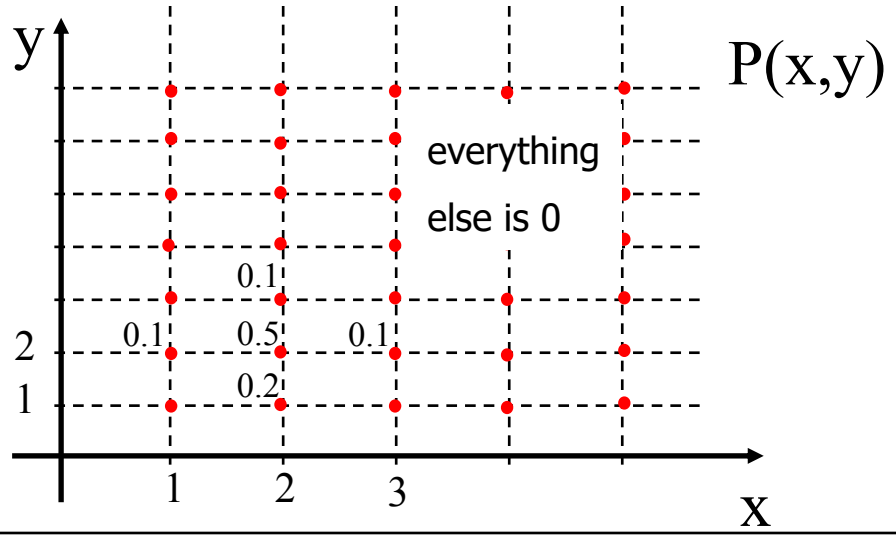
1.6 Expected Values of Functions of Two Variables

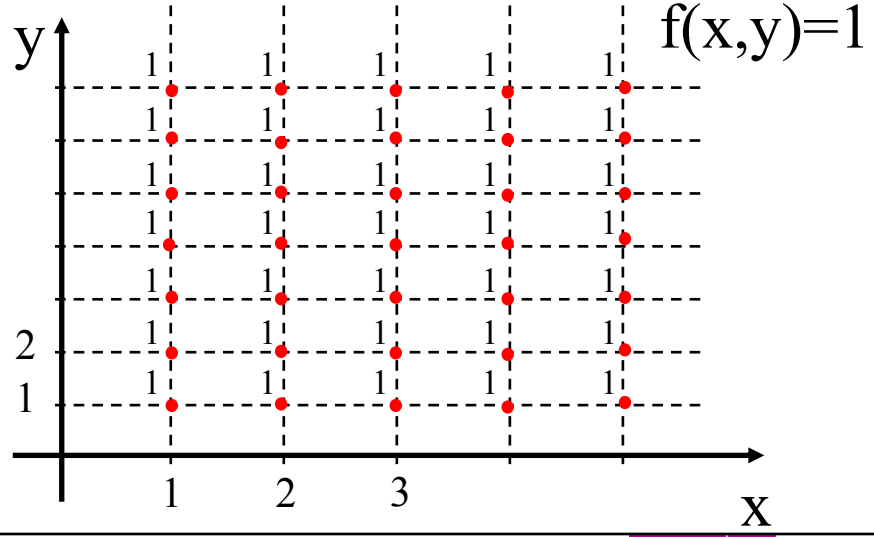
The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Example

Which value do you expect for function f(x, y) in this case?





1.6 Expected Values of Functions of Two Variables

The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Example

E [f(x, y)]= 1 · 0.1 + 1 · 0.2 + 1 · 0.5 + 1 · 0.1 + 1 · 0.1 = 1

→ Reasonable, since the function has always value 1. Therefore, the probability of (x, y) does not have an influence on the expected value.

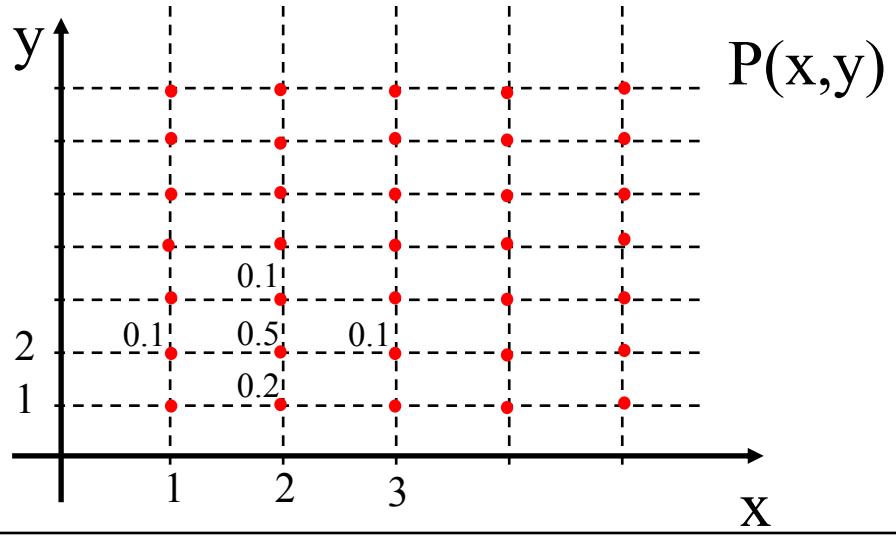
1.6 Expected Values of Functions of Two Variables

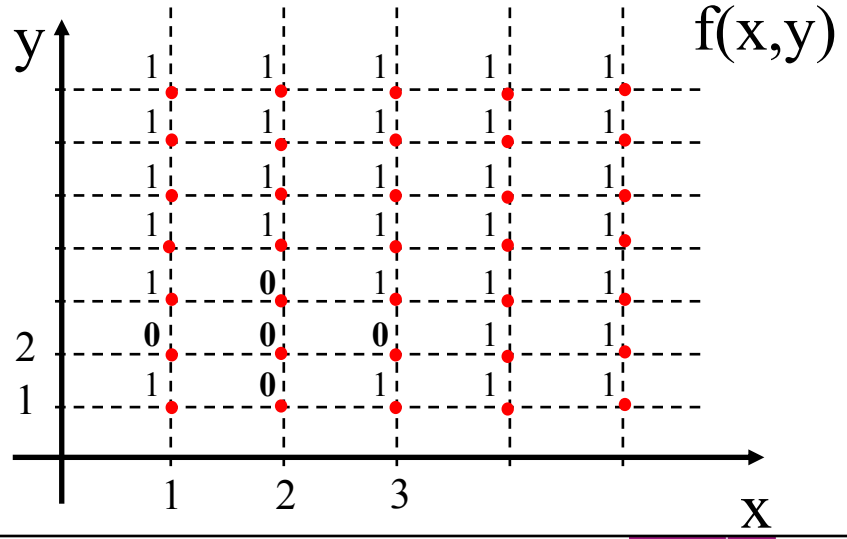
The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Example

Which value do you expect for function f(x, y) in this case?





1.6 Expected Values of Functions of Two Variables

The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Example

E [f(x, y)]= 1 · 0 + … + 1 · 0 + 0 · 0.1 + 0 · 0.2 + 0 · 0.5 + 0 · 0.1 + 0 · 0.1 = 0

- For all possible combinations of x and y the function value is 0.
- For impossible combinations, the function value does not matter.

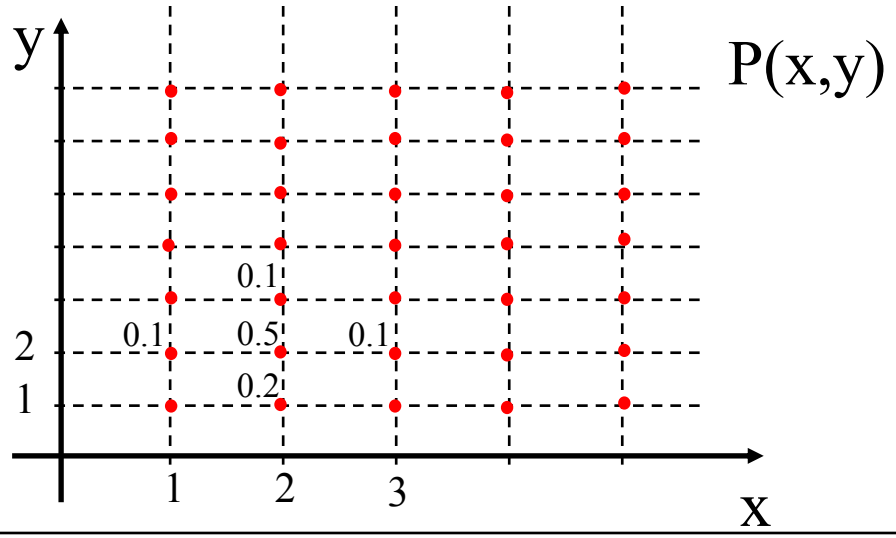
1.6 Expected Values of Functions of Two Variables

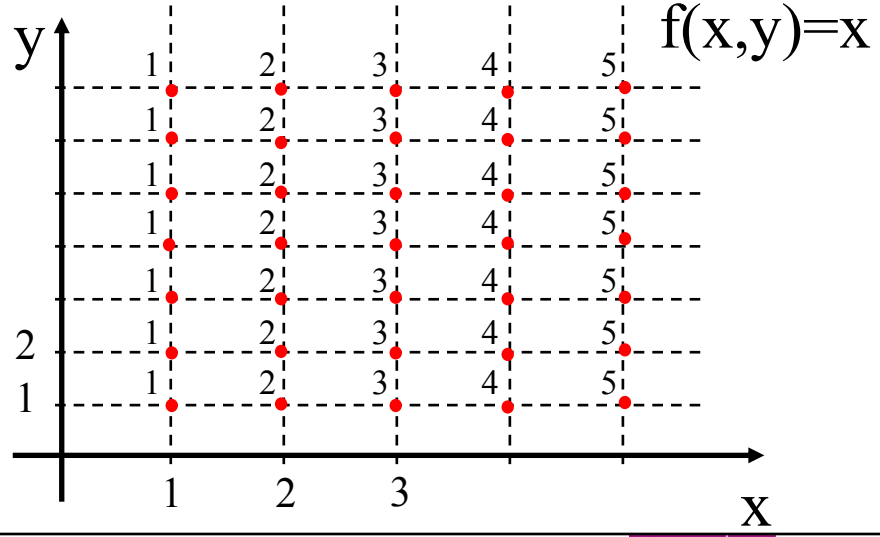
The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Example

Which value do you expect for function f(x, y) in this case?





1.6 Expected Values of Functions of Two Variables

The expected value of a function f(x, y) of two random variables is given by:

$$E[f(x,y)] = \sum_{x \in X} \sum_{y \in Y} f(x,y) P(x,y)$$

Example

E [f(x, y)]= 1 · 0.1 + 2 · 0.2 + 2 · 0.5 + 2 · 0.1 + 3 · 0.1 = 2

→ Expected function value 'in x-direction'

→ Mean or *first moment*

1.6 Expected Values of Functions of Two Variables

Means (first moments) and variances (second moments) are given by:

$$\mu_x = E[x] = \sum_{x \in X} \sum_{y \in Y} x P(x, y)$$

$$\mu_y = E[y] = \sum_{x \in X} \sum_{y \in Y} y P(x, y)$$

$$\sigma_x^2 = E[(x - \mu_x)^2] = \sum_{x \in X} \sum_{y \in Y} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = E[(y - \mu_y)^2] = \sum_{x \in X} \sum_{y \in Y} (y - \mu_y)^2 P(x, y)$$

1.6 Expected Values of Functions of Two Variables

Means (first moments) and variances (second moments) are given by:

How strong does a random variable deviate from its mean value.

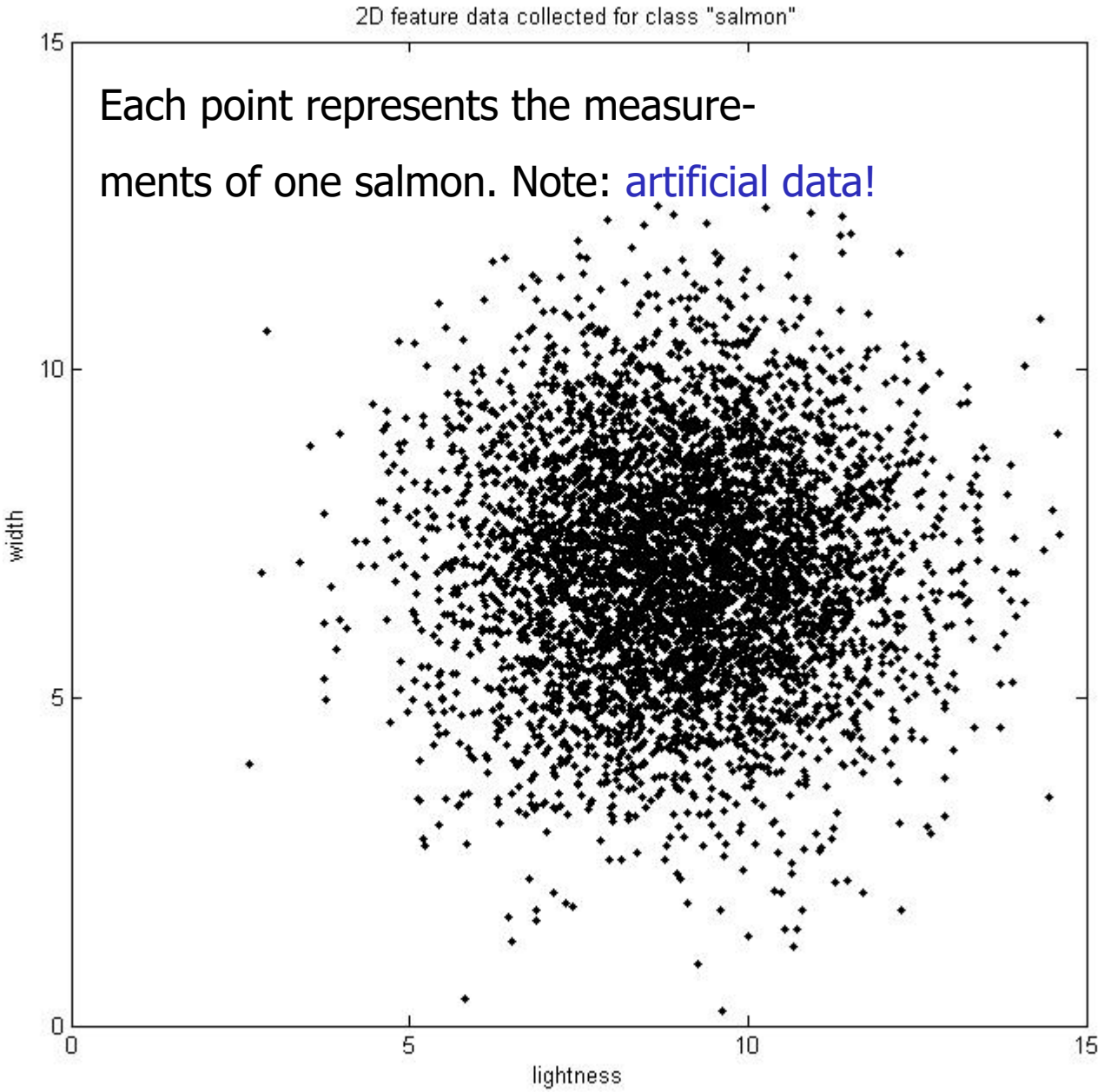
$$\sigma_x^2 = E[(x - \mu_x)^2] = \sum_{x \in X} \sum_{y \in Y} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = E[(y - \mu_y)^2] = \sum_{x \in X} \sum_{y \in Y} (y - \mu_y)^2 P(x, y)$$

Practical example:

x = lightness

y = width

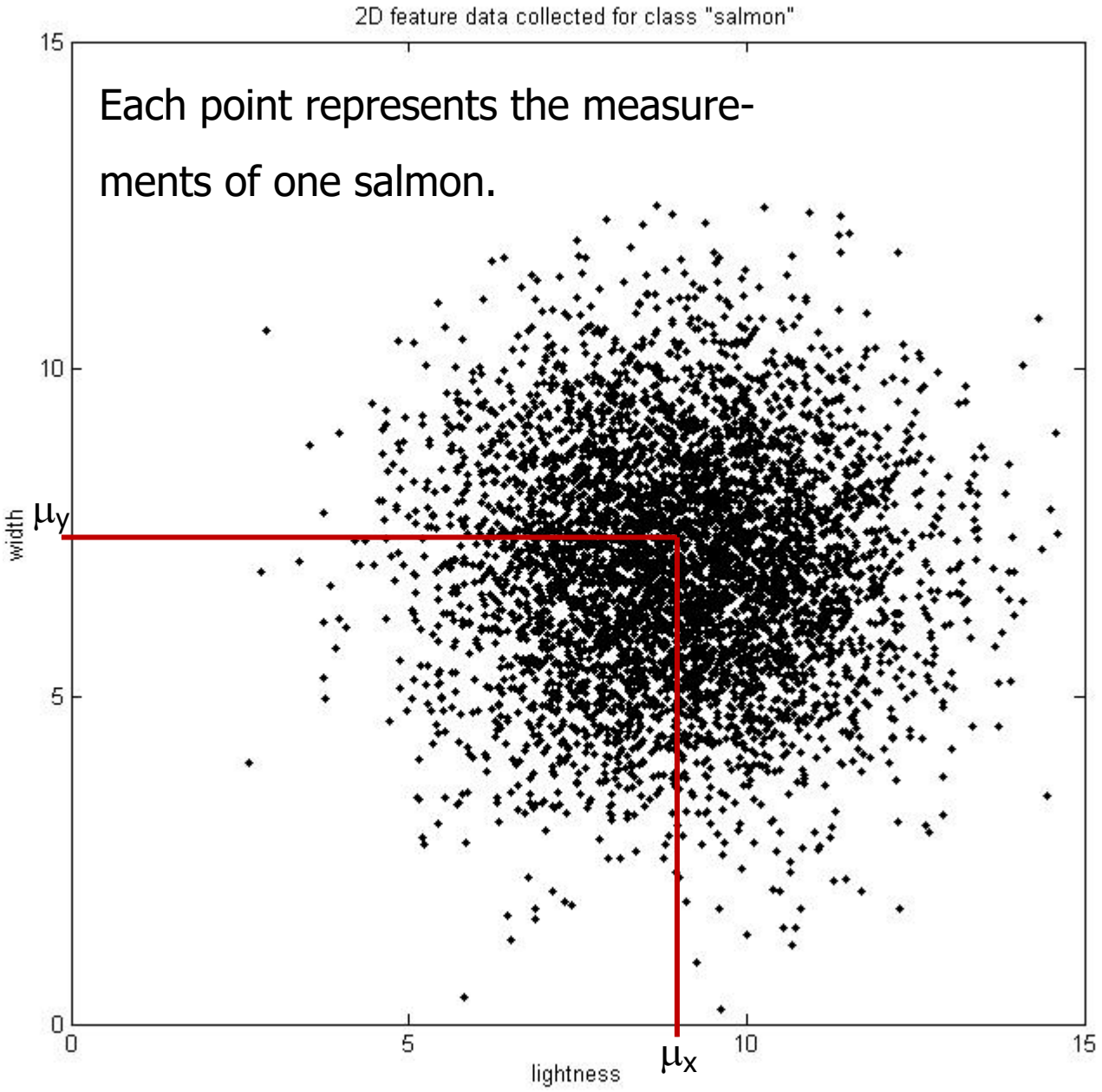


Practical example:

x = lightness

y = width

Means: μ_x , μ_y

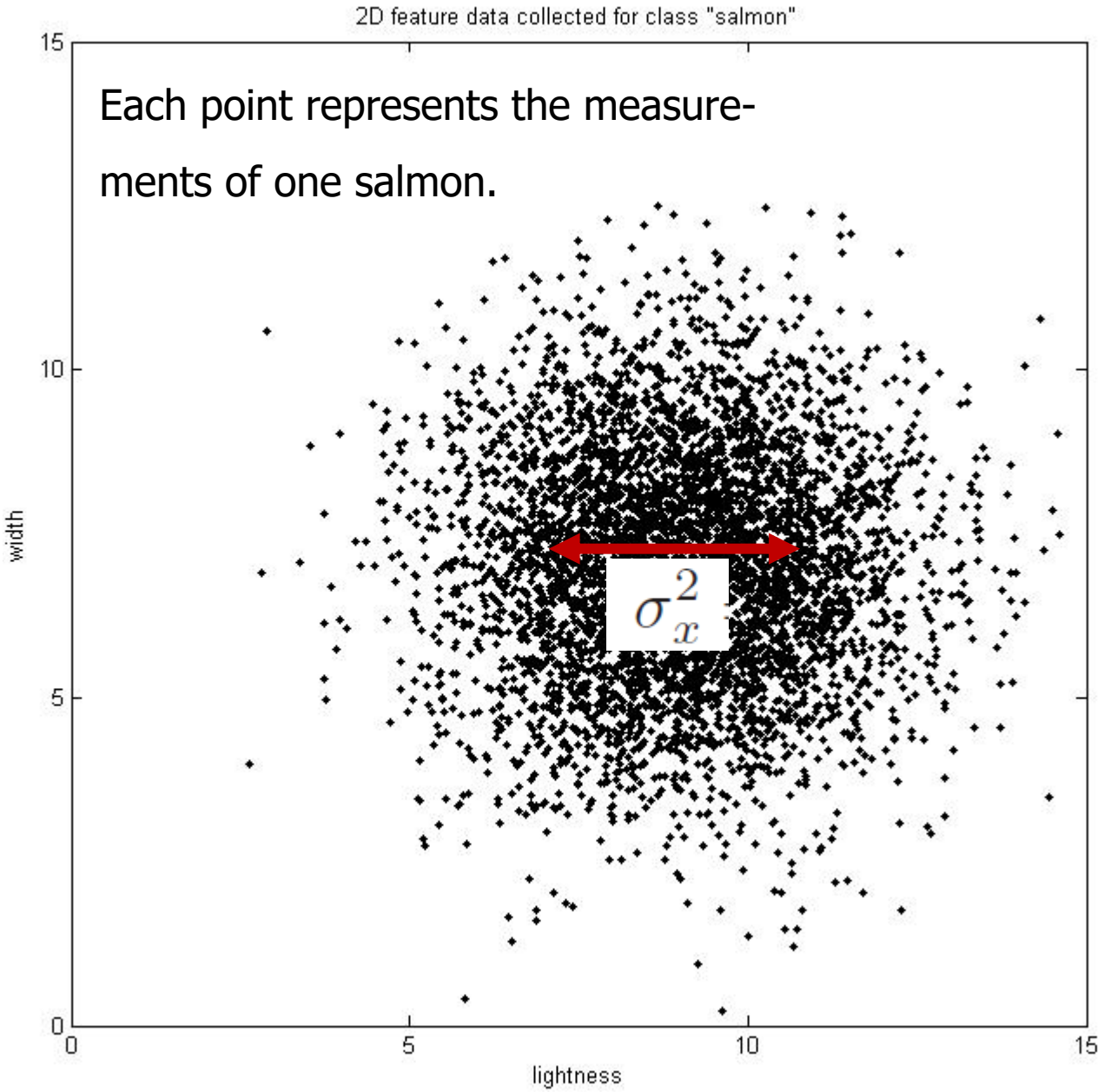


Practical example:

x = lightness

y = width

Variance: σ_x^2



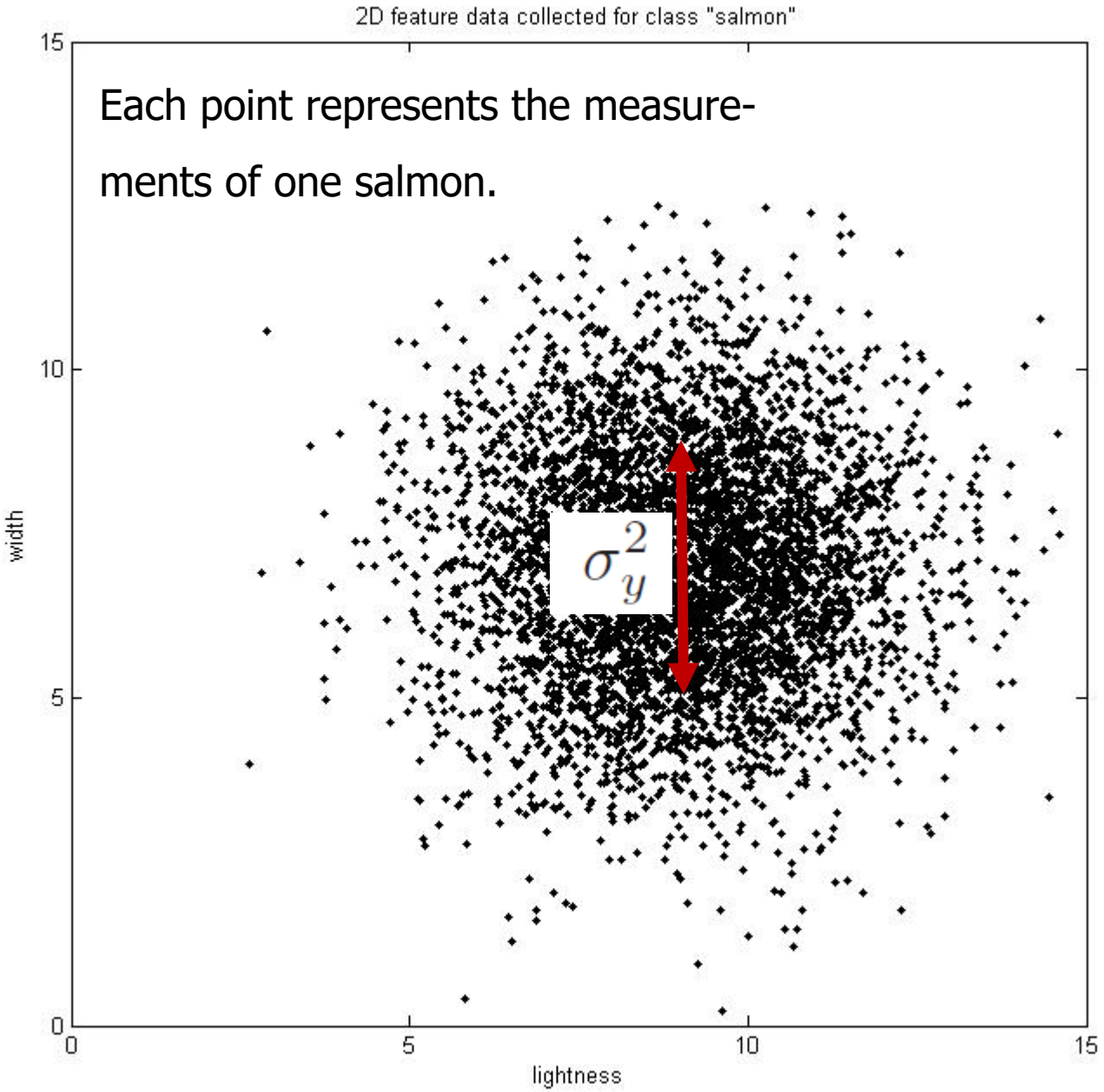
Practical example:

x = lightness

y = width

Variance: σ_y^2

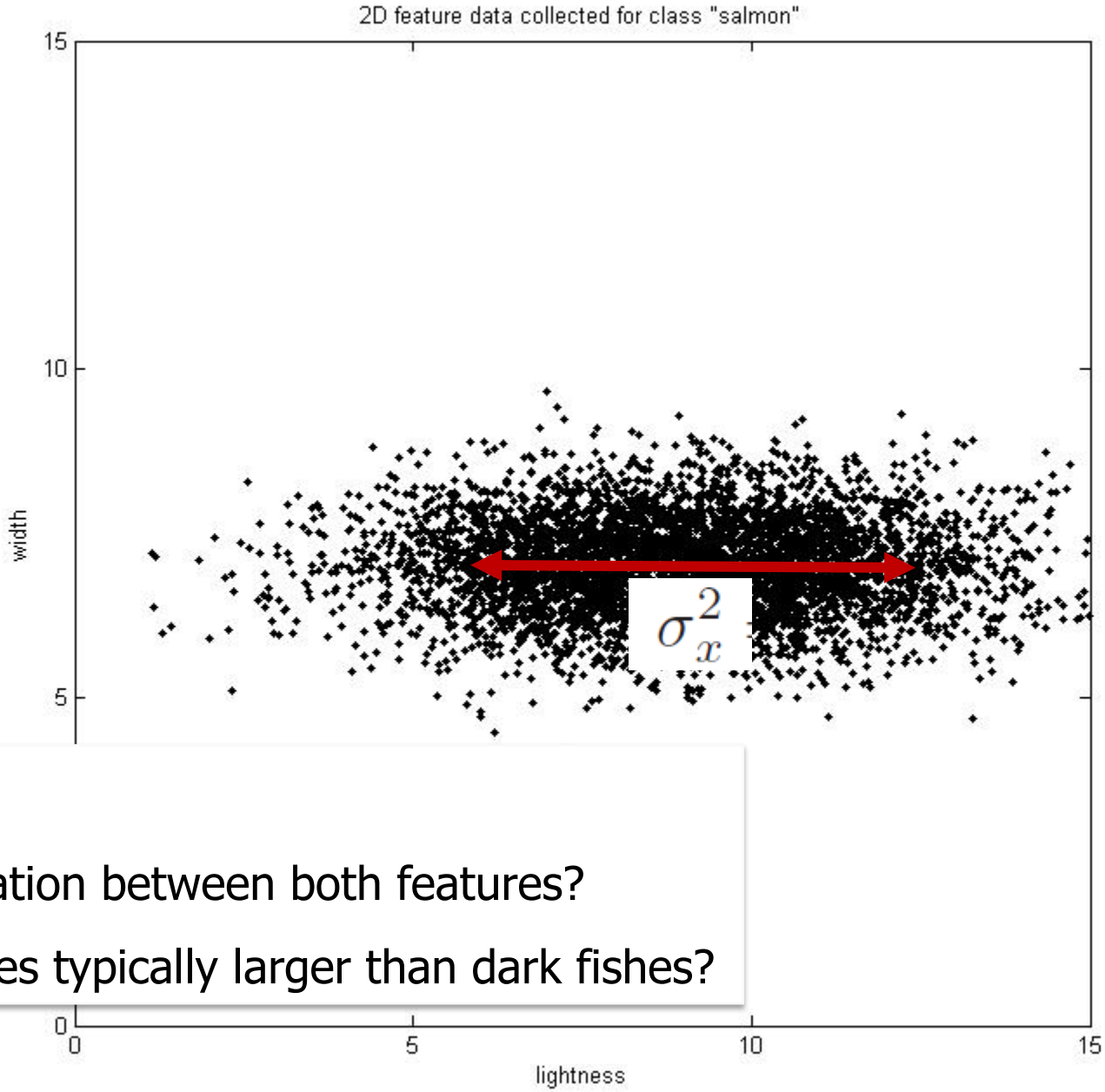
→ Same variance!



Practical example:

Another distribution:

→ stronger variance of lightness



Question:

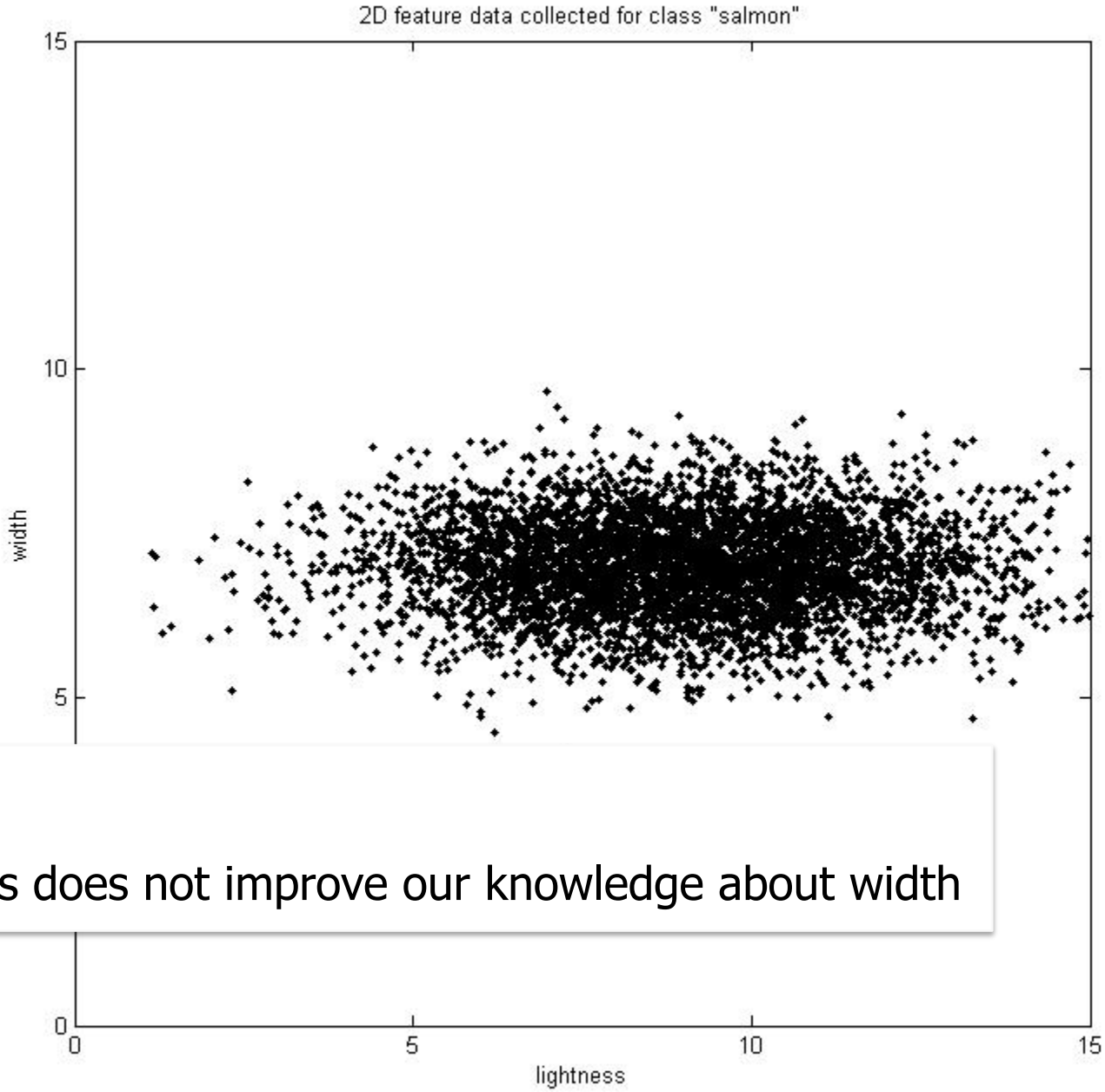
Is there any correlation between both features?

Or: Are bright fishes typically larger than dark fishes?

Practical example:

Another distribution:

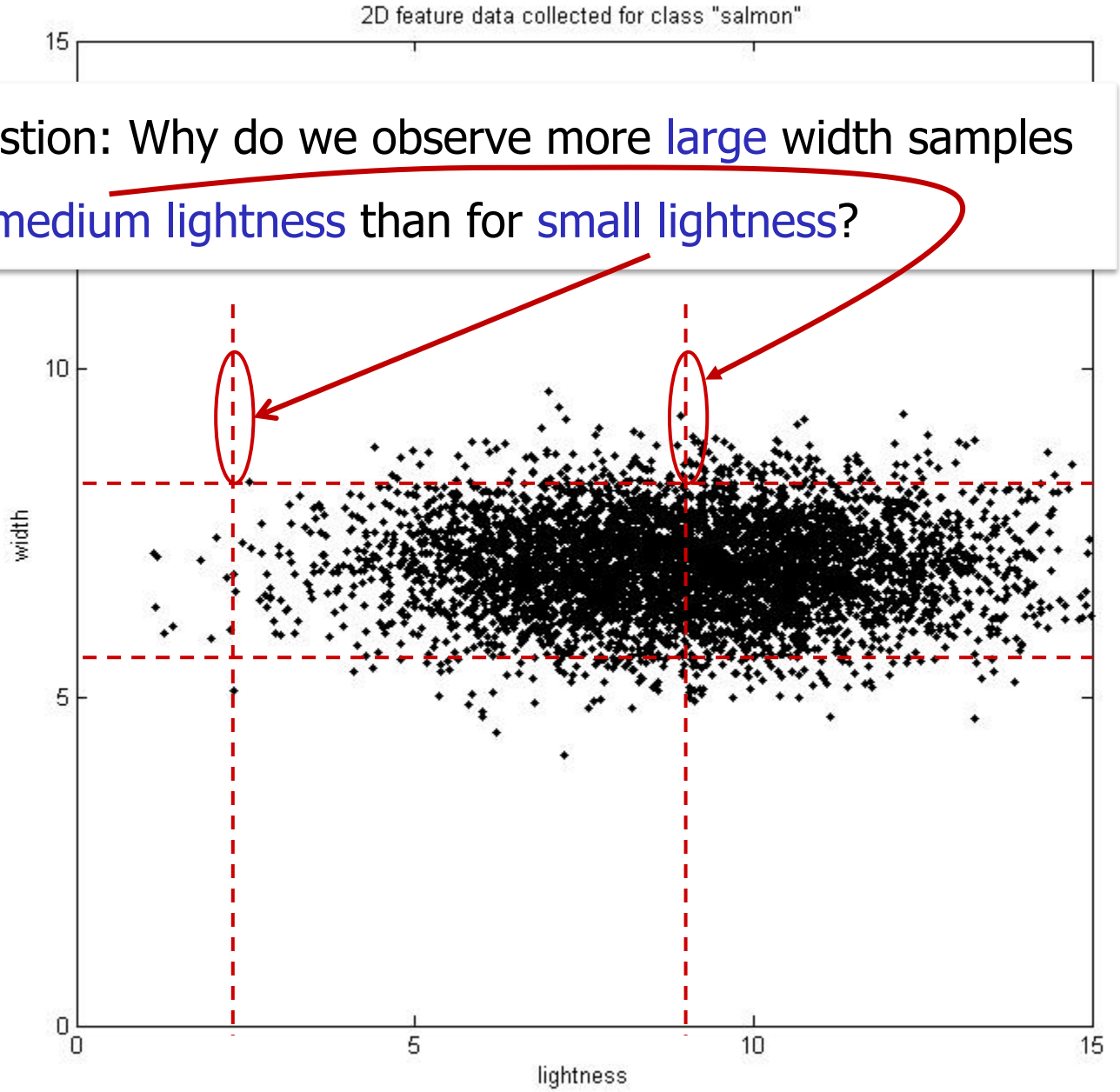
→ stronger variance of lightness



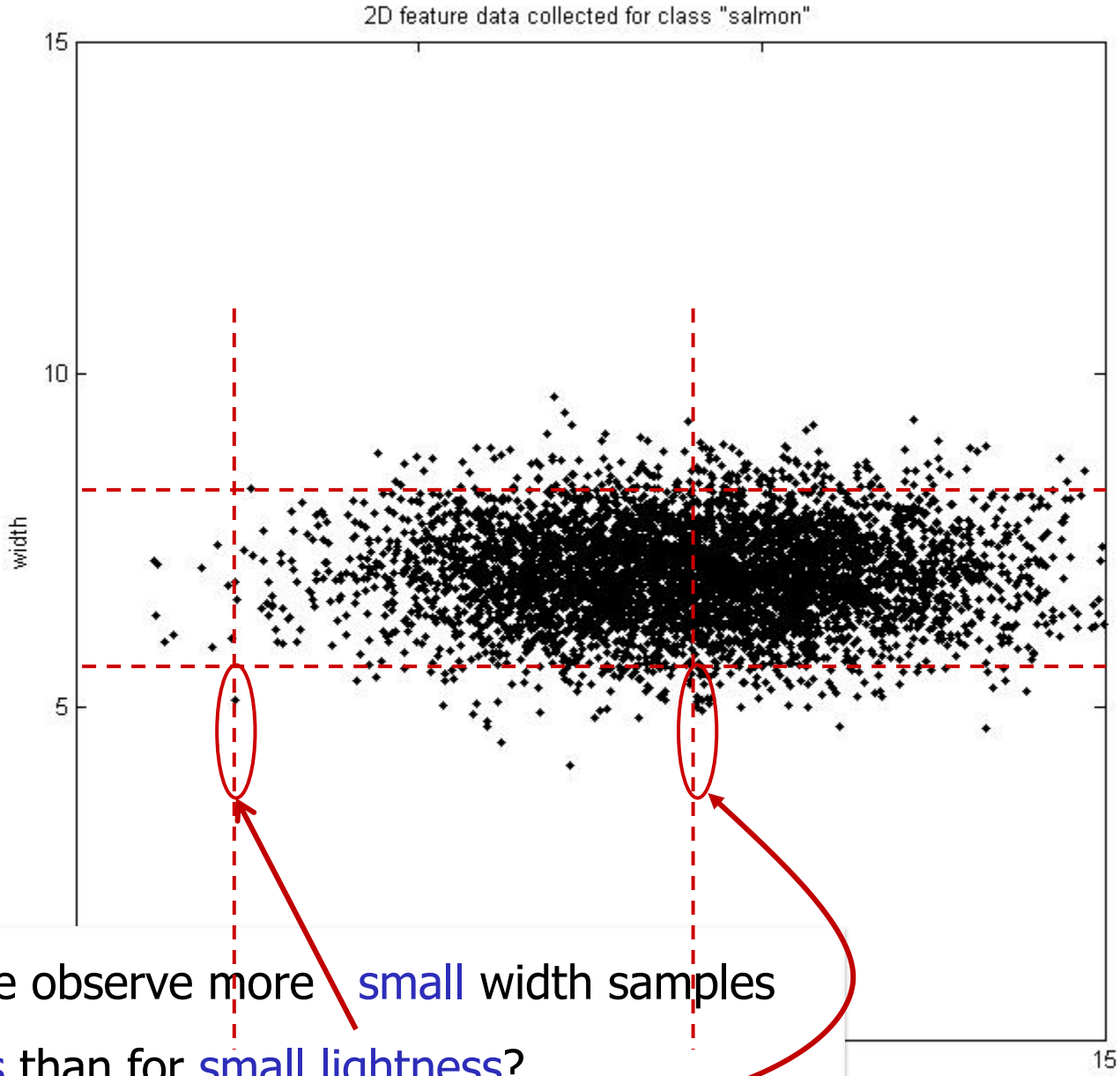
Answer: No!

Knowledge of lightness does not improve our knowledge about width

Practical example:



Practical example:

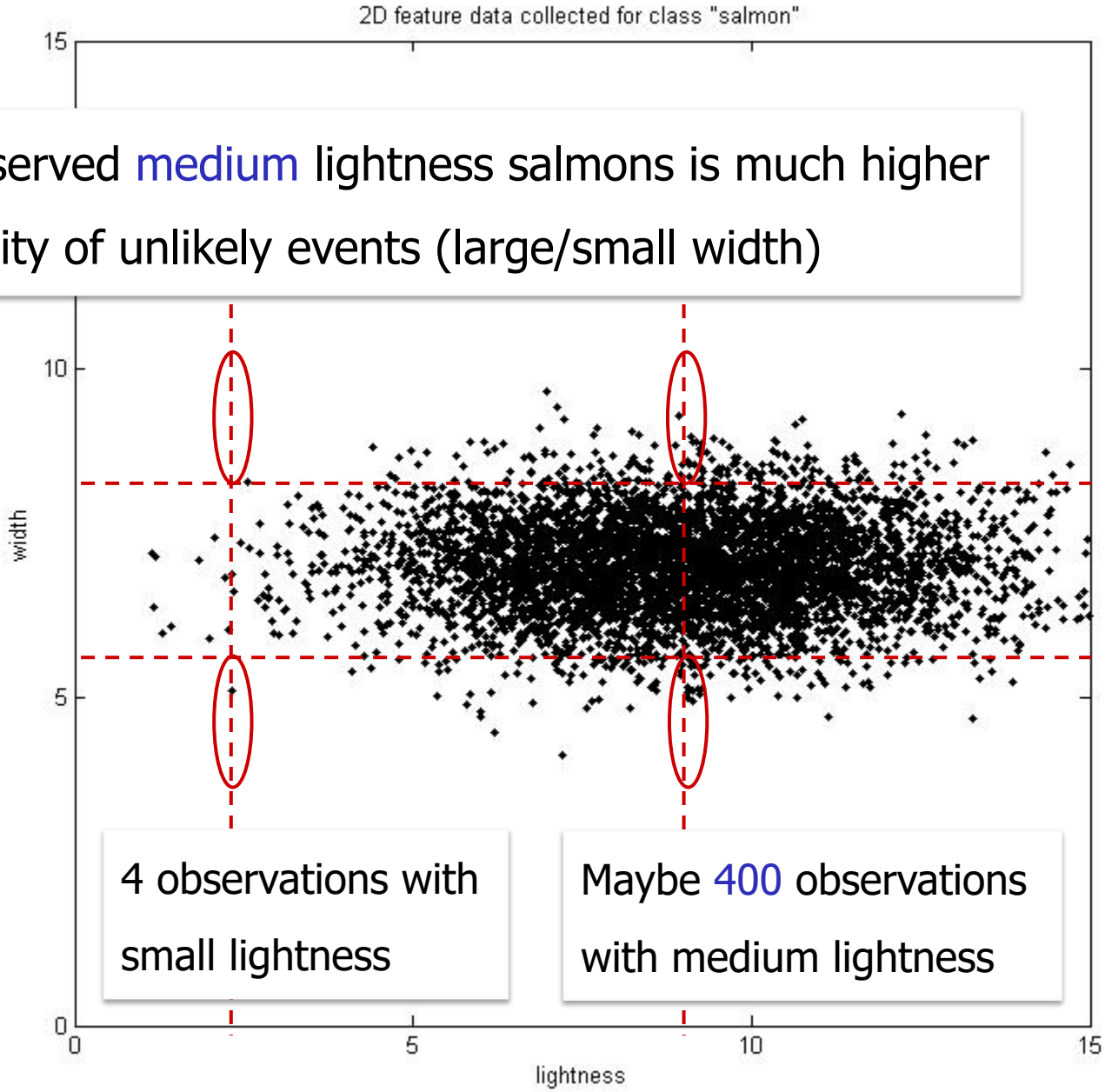


Question: Why do we observe more small width samples for medium lightness than for small lightness?

Practical example:

The number of observed medium lightness salmons is much higher

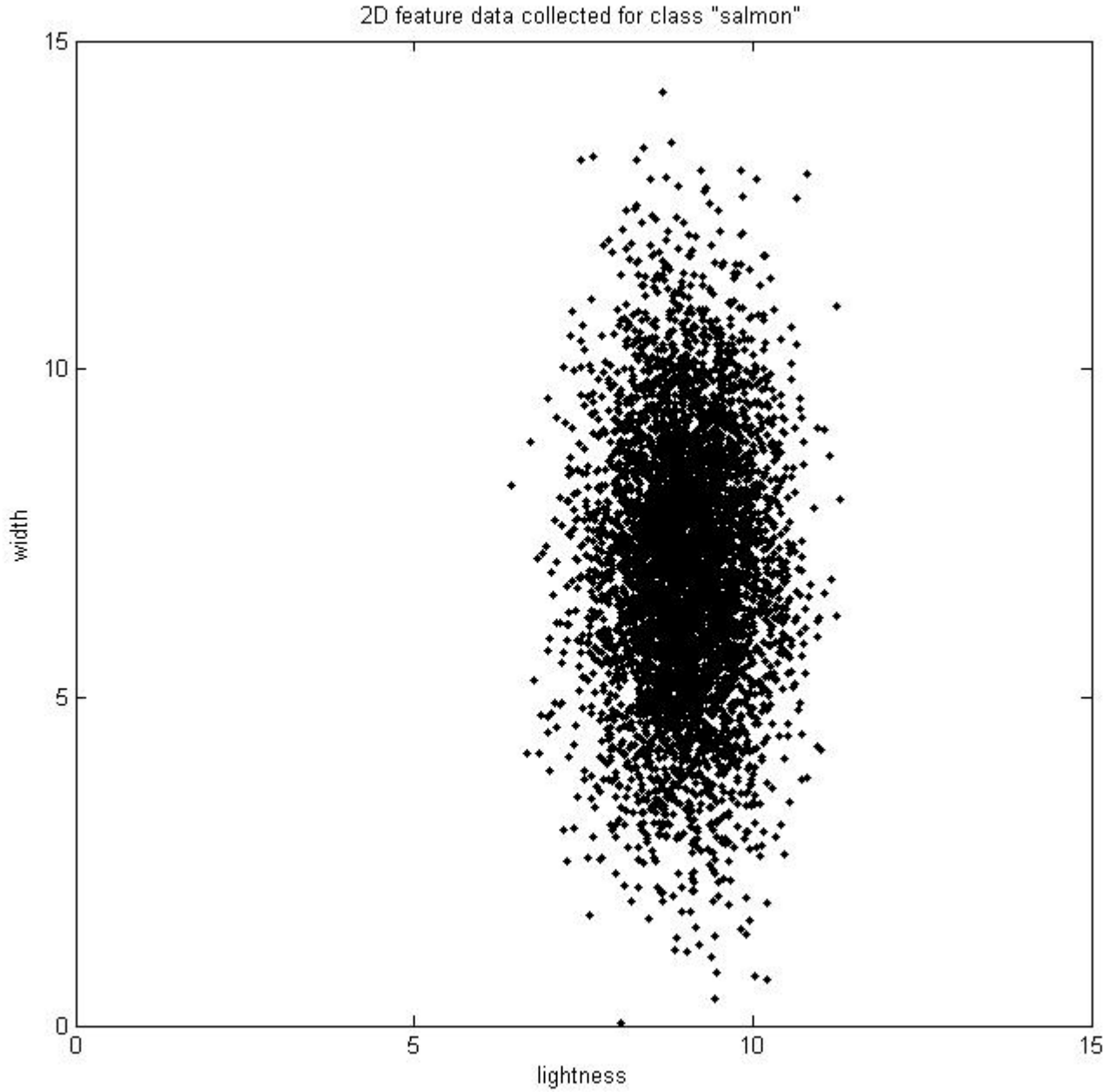
→ higher probability of unlikely events (large/small width)



Practical example:

Another distribution:

→ strong variance of width



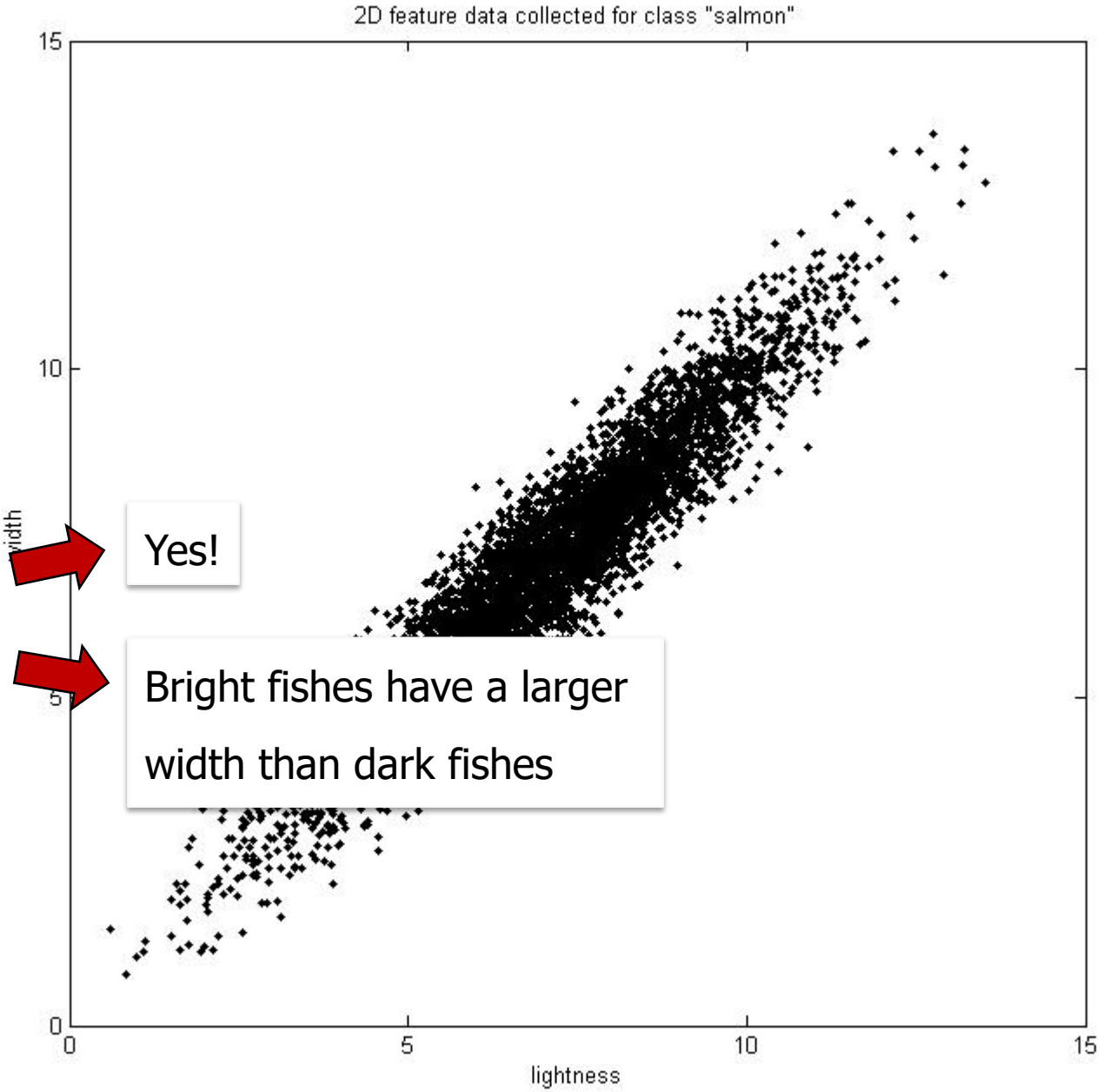
Practical example:

What about this distribution?

Comments?

→ Statistically dependend?

→ Practical interpretation?



Yes!

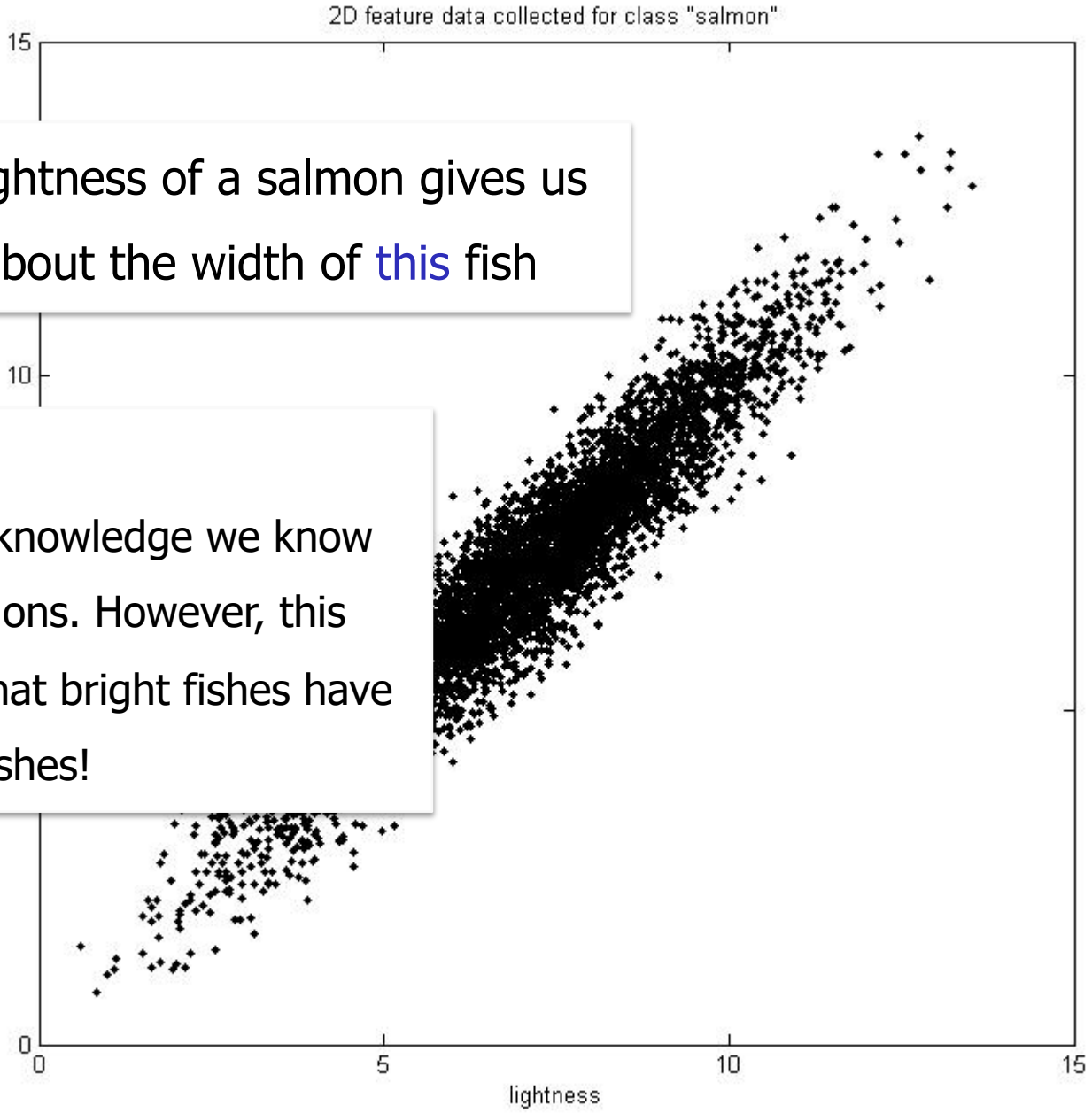
Bright fishes have a larger width than dark fishes

Practical example:

Knowledge about the lightness of a salmon gives us additional information about the width of this fish

Why additional?

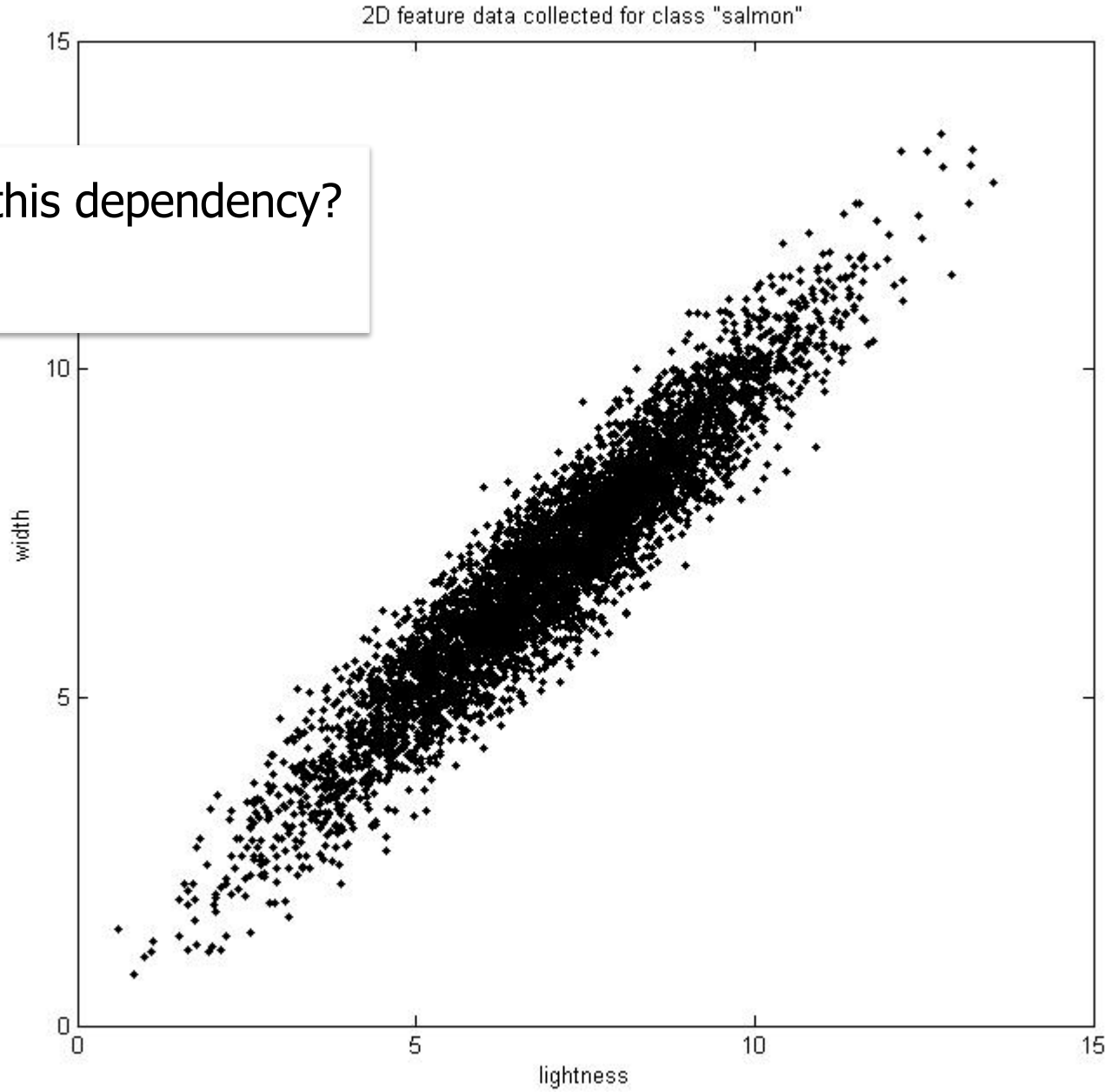
Because also without this knowledge we know the mean width of all salmons. However, this does not exploit the fact that bright fishes have a larger width than dark fishes!



Practical example:

How can we measure this dependency?

→ Covariance



1.7 Covariance

The covariance of x and y is a 'cross-moment' which is defined as:

$$\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$$

$$= \sum_{x \in X} \sum_{y \in Y} (x - \mu_x)(y - \mu_y) P(x, y)$$

Can be used for measuring the level of statistical dependency of x and y.

Normalized covariance: **Correlation coefficient**

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

→ Values between -1 and 1

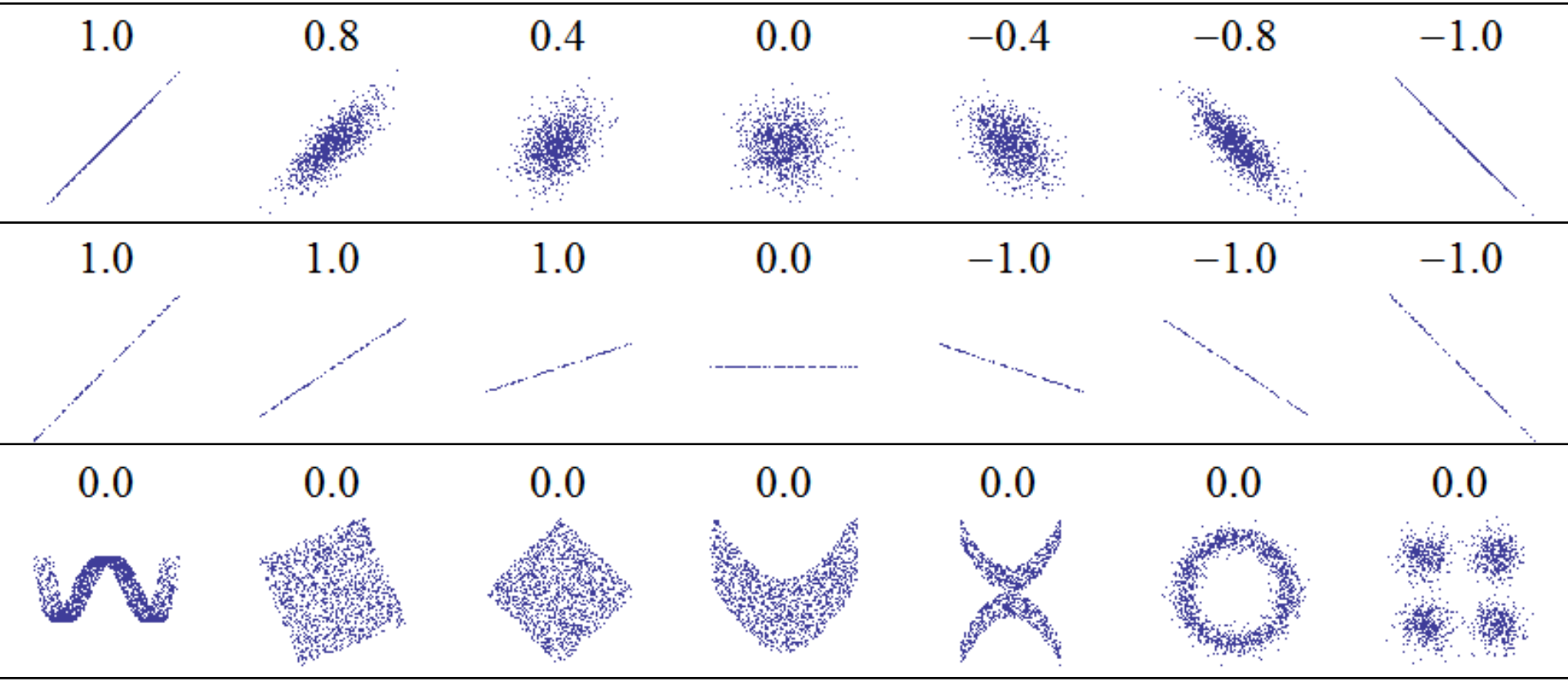
→ ρ = +1 maximally positively related

→ ρ = -1 maximally negatively related

→ ρ = 0 uncorrelated

1.7 Covariance

Some examples of 2D feature sample distribtions and correlation coefficient:



Source: Wikipedia

Last row: Examples for uncorrelated but statistically dependent distributions

1.7 Covariance

Covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$

- Diagonal elements are just the variances of x and y
- Statistically independent x and y → covariances σ_{xy} and σ_{yx} are zero

  → matrix is diagonal
- Covariance matrix is symmetric since $\sigma_{xy} = \sigma_{yx}$

1.8 Conditional probability

Let x and y be statistically dependent variables

→ knowing the value of one gives us better estimate of the value of the other

Expressed by definition of **conditional probability**:

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Explanation:

Probability of value x: $P(x)$

Probability of value x if we know the value of y: $P(x|y)$

1.8 Conditional probability

Let x and y be statistically dependent variables

→ knowing the value of one gives us better estimate of the value of the other

Expressed by definition of **conditional probability**:

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

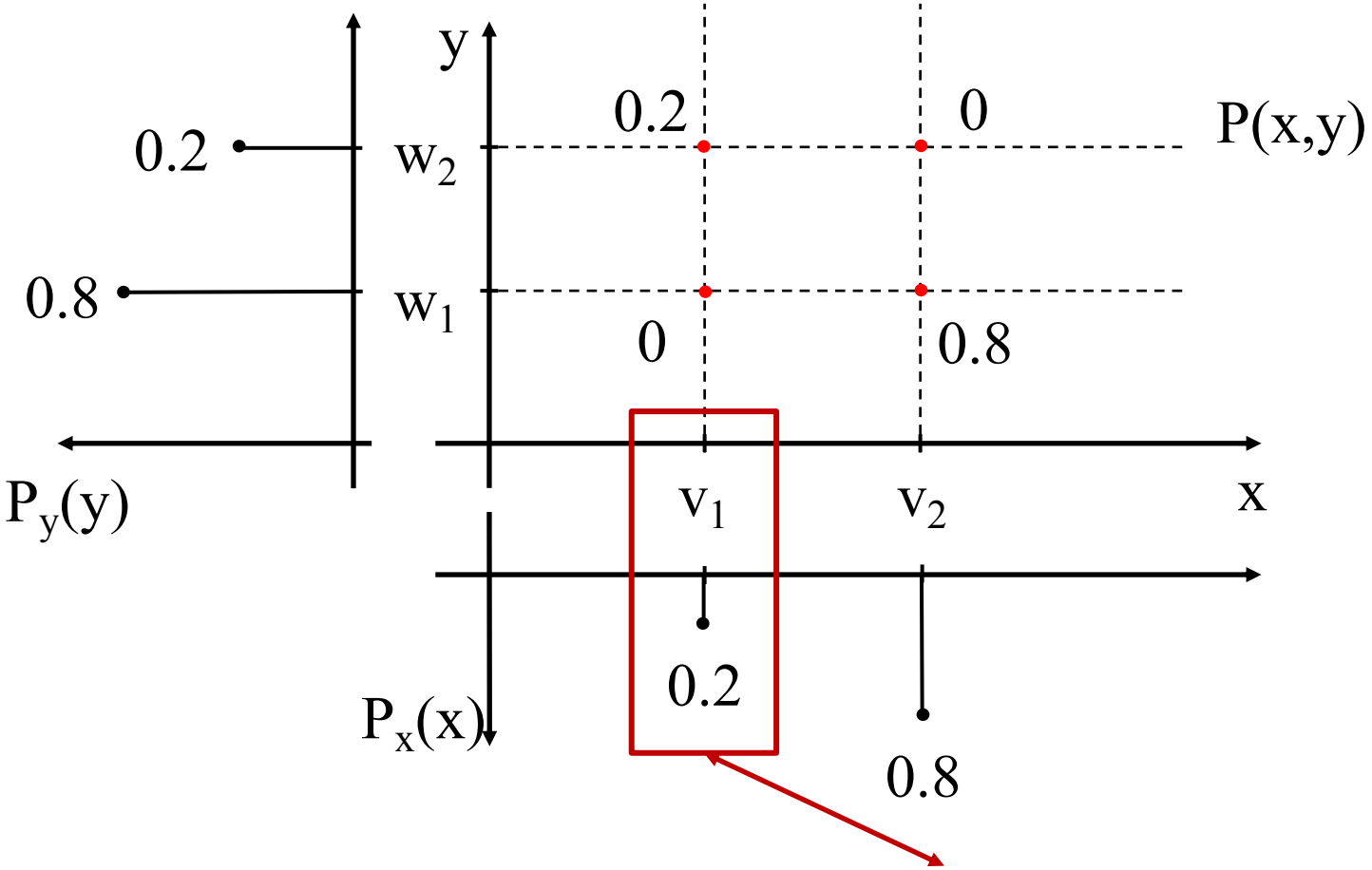
If x and y are statistically independent: $P(x|y) = P(x)$

The probability of x stays the same – independently of knowledge about y.

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **dependent**

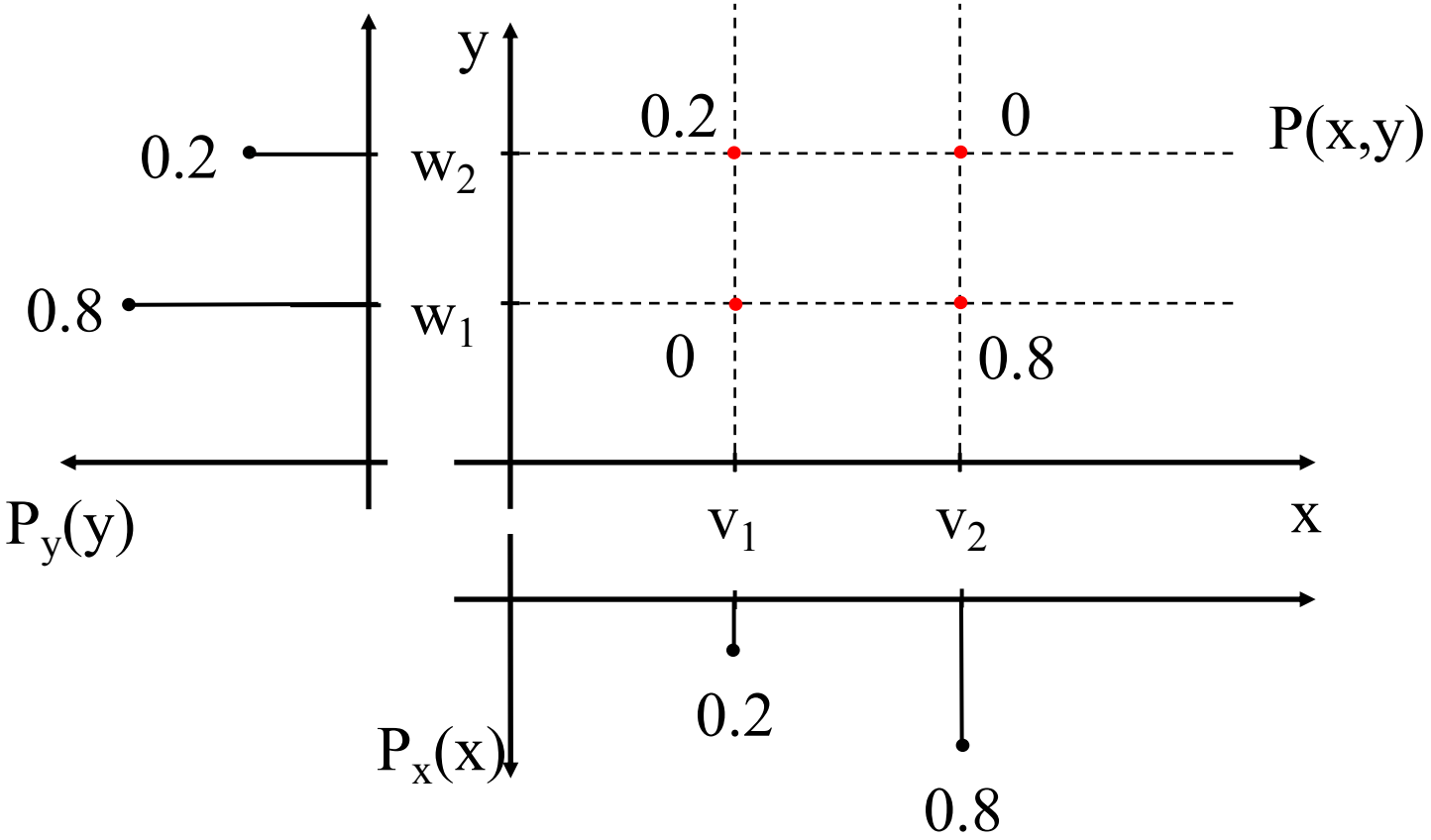


Probability of x = v_1 without knowledge about y: $P(x=v_1) = 0.2$

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **dependent**

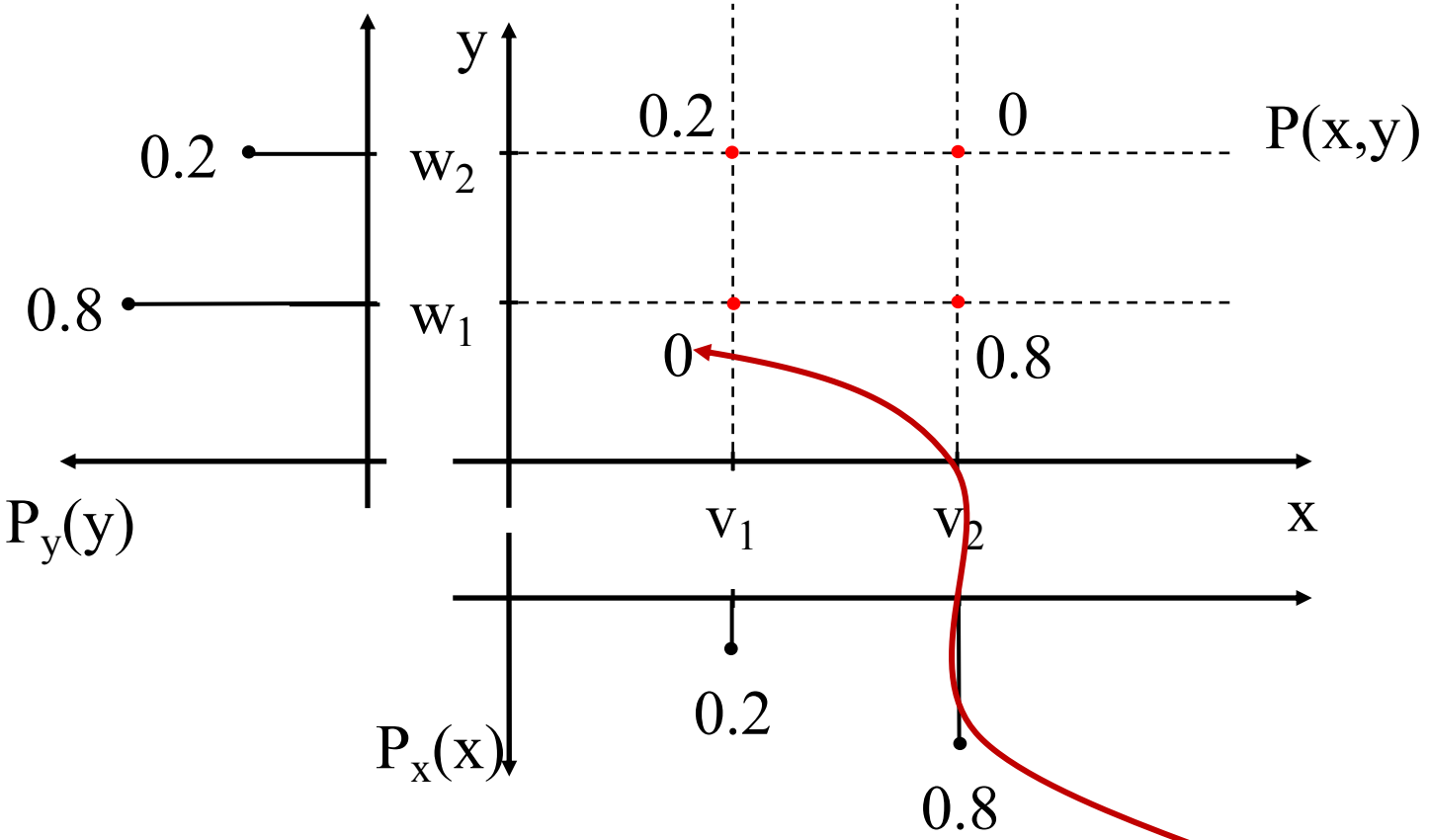


Probability of x = v_1 with knowledge about y: $P(x=v_1|y=w_1) = P(v_1, w_1) / P(w_1)$

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **dependent**

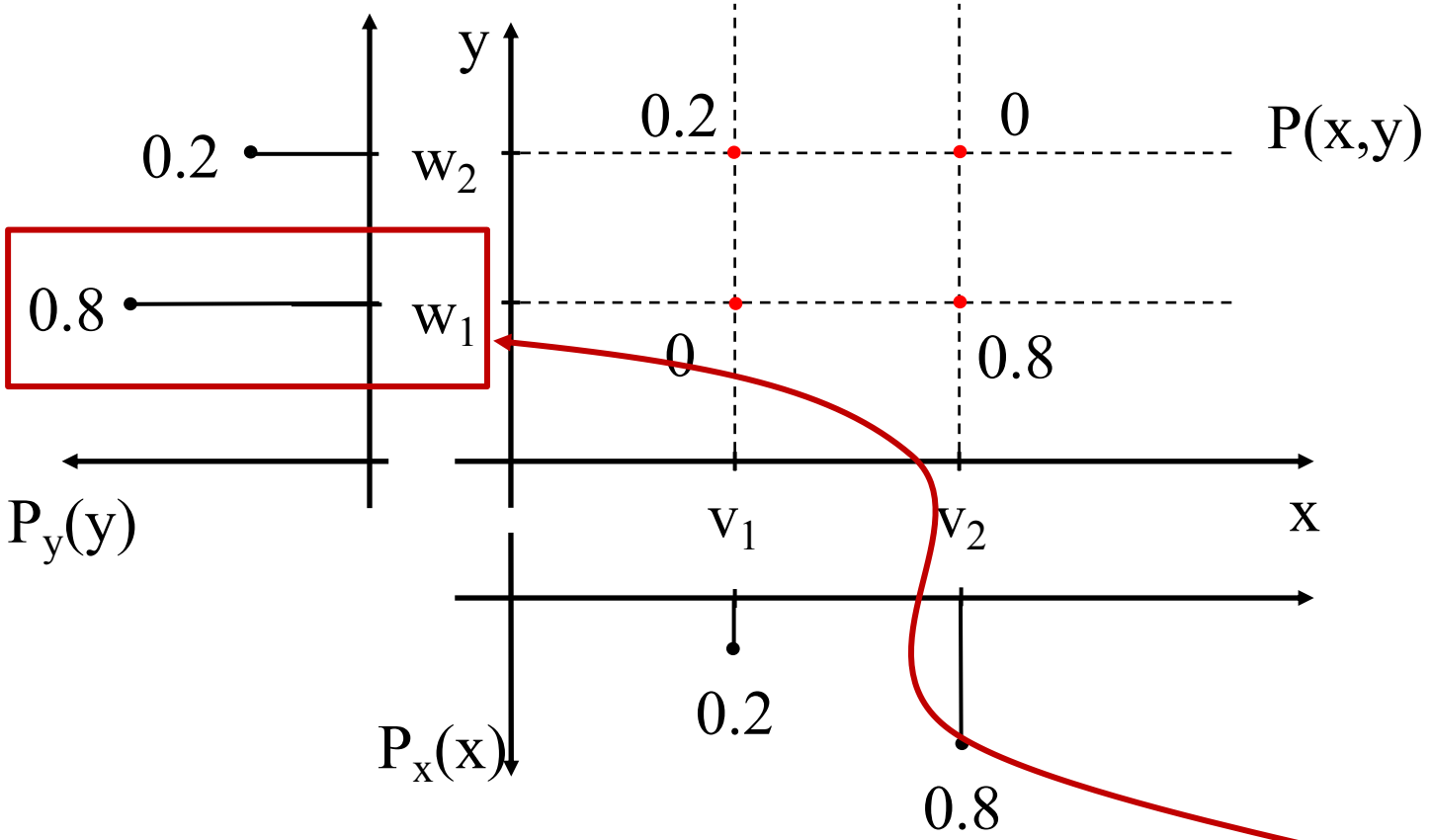


Probability of $x = v_1$ with knowledge about y: $P(x=v_1|y=w_1) = P(v_1, w_1) / P(w_1)$

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **dependent**

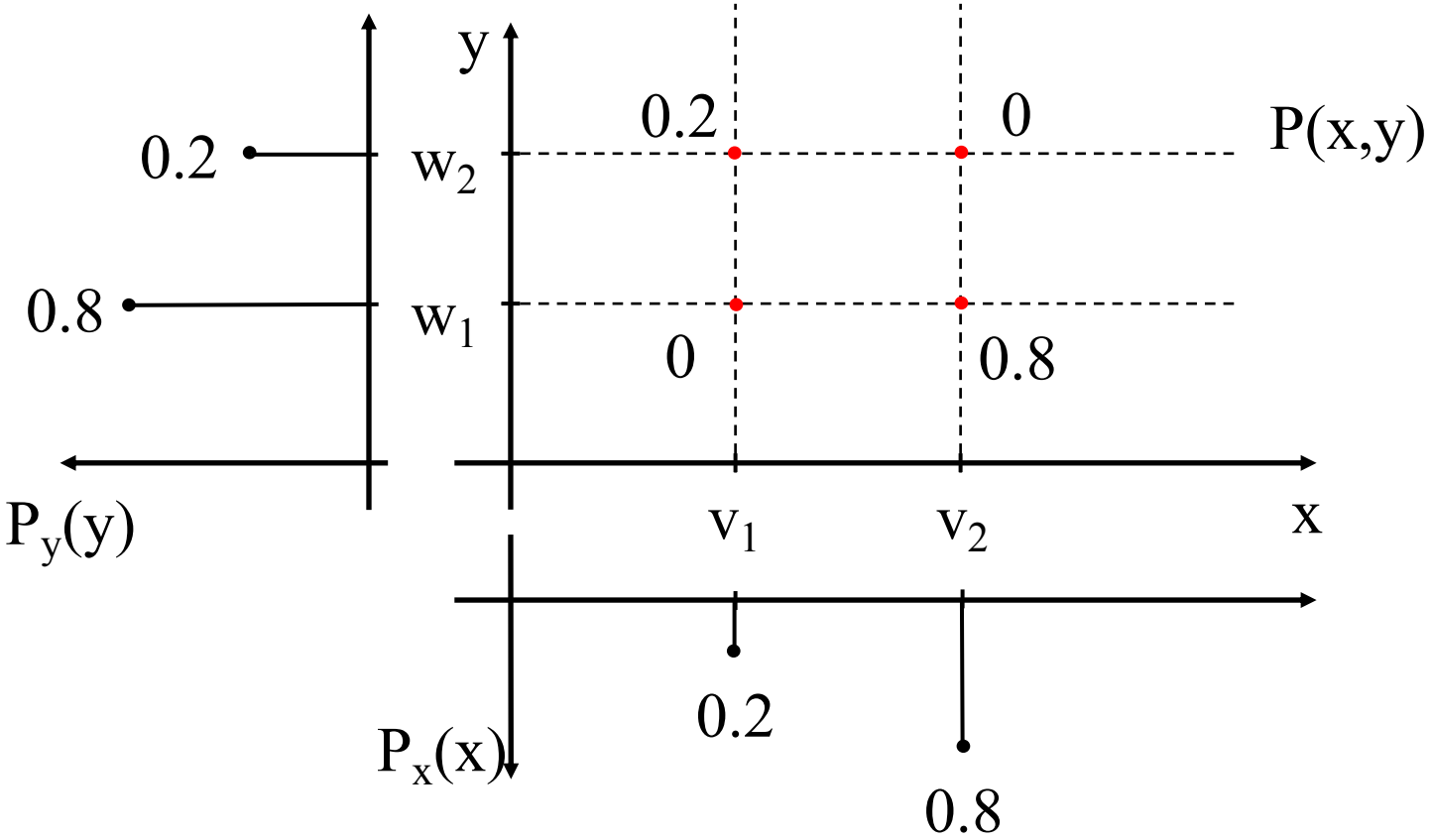


Probability of x = v_1 with knowledge about y: $P(x=v_1|y=w_1) = P(v_1, w_1) / P(w_1)$

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **dependent**

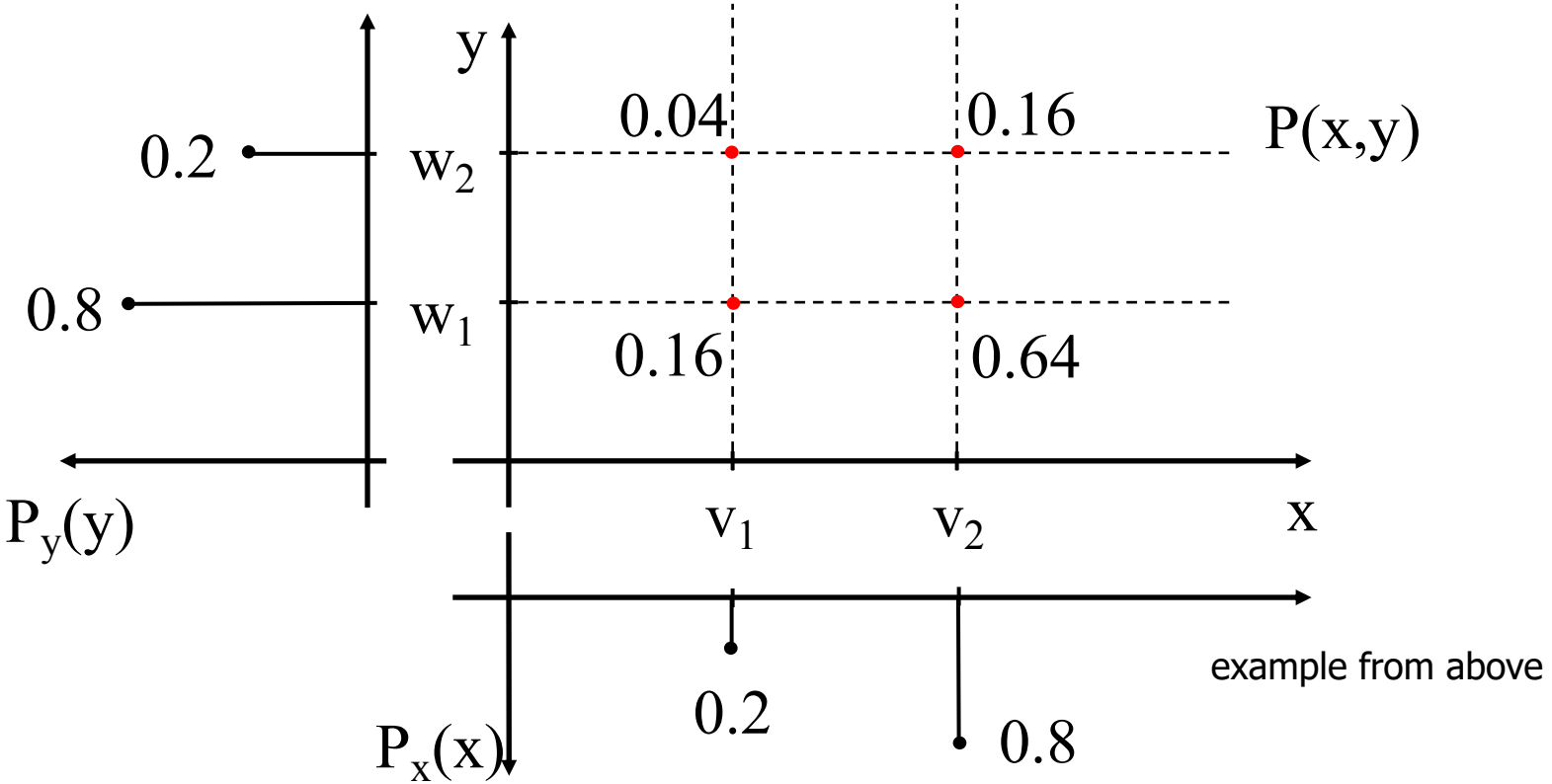


$P(x=v_1|y=w_1) = P(v_1, w_1) / P(w_1) = 0 / 0.8 = 0$ ≠ $P(x=v_1) = 0.2$

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **independent**

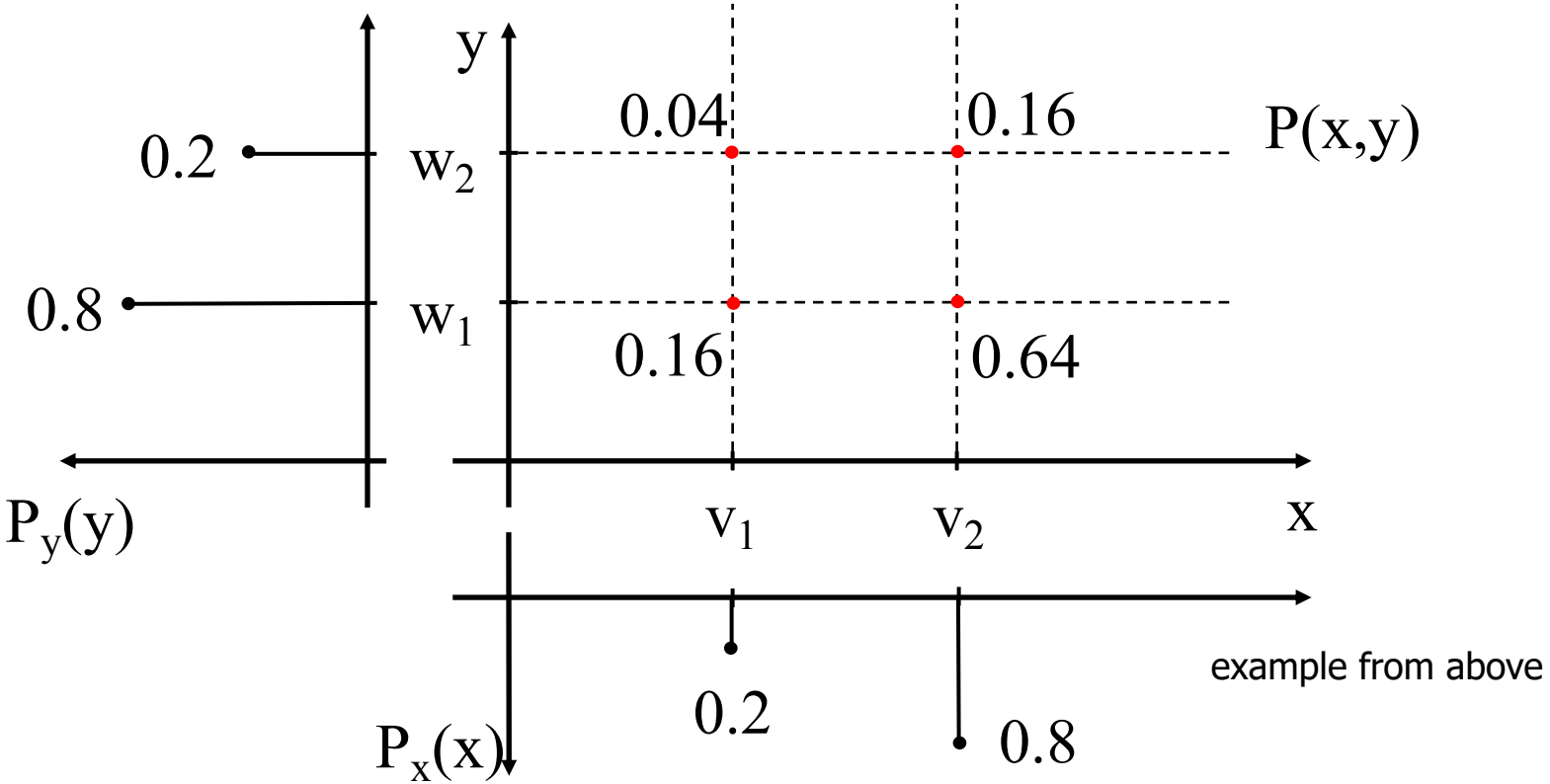


$P(x=v_1|y=w_1) = P(v_1, w_1) / P(w_1) = 0.16 / 0.8 = 0.2$ = $P(x=v_1) = 0.2$

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **independent**

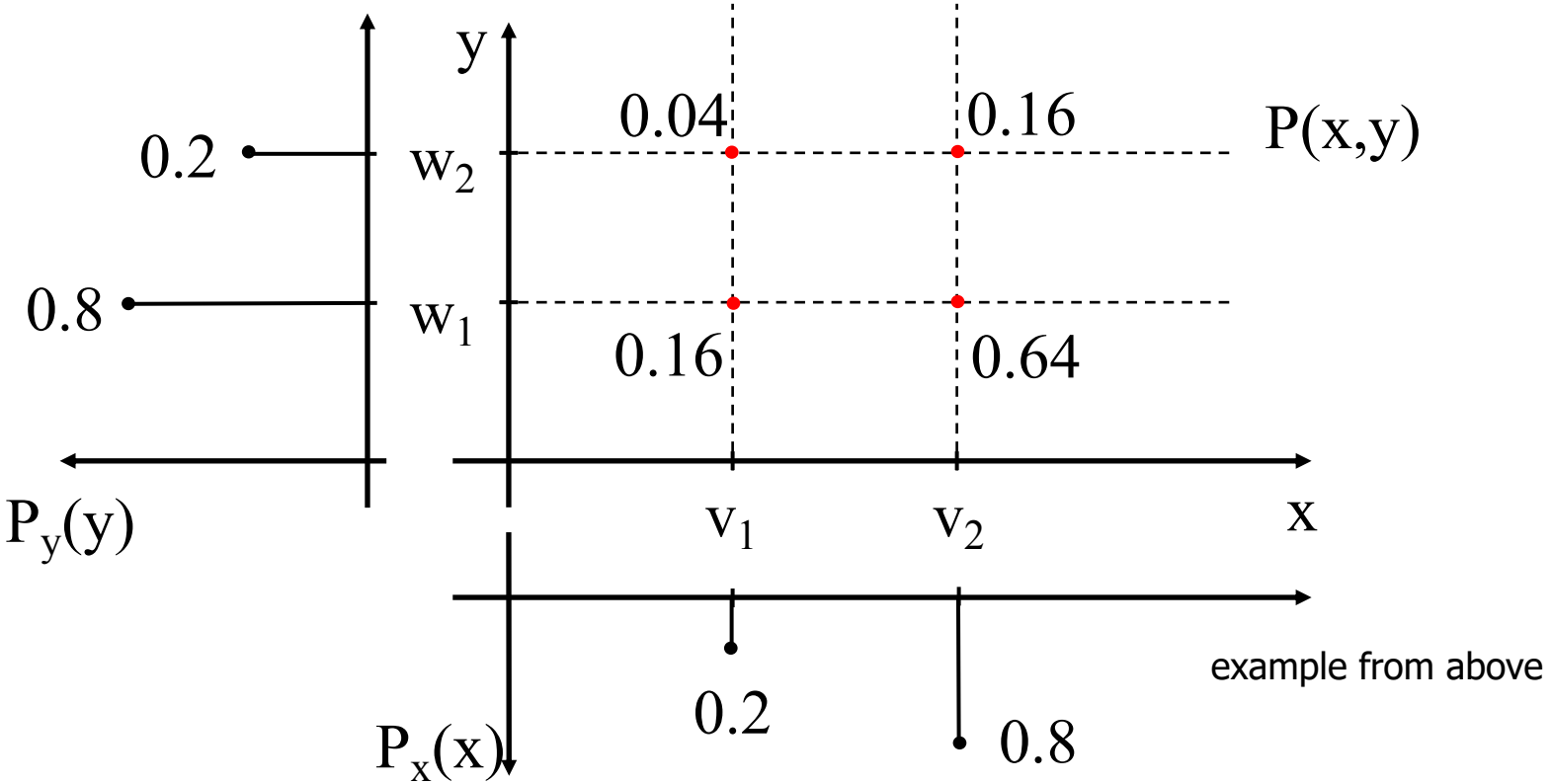


We do not need the knowledge of y to determine the probability of $x=v_1$.

1.8 Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

Example: x and y are statistically **independent**



$P(x=v_2|y=w_2) = P(v_2, w_2) / P(w_2) = 0.16 / 0.2 = 0.8$ = $P(x=v_2) = 0.8$

1.9 Vector random variables

In pattern recognition applications we typically need more than one feature

➡ Feature vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^T$

Probability mass function satisfies $P(\boldsymbol{x}) \geq 0$

$$\sum_{\boldsymbol{x}} P(\boldsymbol{x}) = 1$$

sum extends over all possible values for the vector **x**

1.9 Vector random variables

Probability mass function can be expressed by conditional probabilities:

$$\begin{aligned} P(x_1, x_2, x_3, x_4, x_5) &= P(x_1, x_2, x_3, x_4 | x_5) P(x_5) \\ &= P(x_1, x_2, x_3 | x_4, x_5) P(x_4 | x_5) P(x_5) \\ &= \dots \end{aligned}$$

Remark: We will use this in the context of spam filtering.

Simplification, if the random variables x_i are statistically independent:

$$P(\boldsymbol{x}) = P_{x_1}(x_1)\, P_{x_2}(x_2) \cdots P_{x_d}(x_d) = \prod_{i=1}^{d} P_{x_i}(x_i)$$

subscripts are used to denote that the $P(x_i)$ will in general have a different form

1.9 Vector random variables

Mean vector and Covariance matrix

Mean vector

$$\boldsymbol{\mu} = E[\boldsymbol{x}] = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_d] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} = \sum_{\boldsymbol{x}} \boldsymbol{x} P(\boldsymbol{x})$$

1.9 Vector random variables

Covariance matrix

covariance of variables x_1, x_2

variance of variable x_2

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{11} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d1} & \dots & \sigma_{dd} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{11} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d1} & \dots & \sigma_d^2 \end{bmatrix}$$

with covariances $\sigma_{ij} = \sigma_{ji} = E[(x_i - \mu_i)(x_j - \mu_j)]$

$$i, j = 1, \dots, d$$

1.9 Vector random variables

The covariance matrix is diagonal in case of statistically independent variables $x_1, \ldots, x_d$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{11} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_d^2 \end{bmatrix}$$

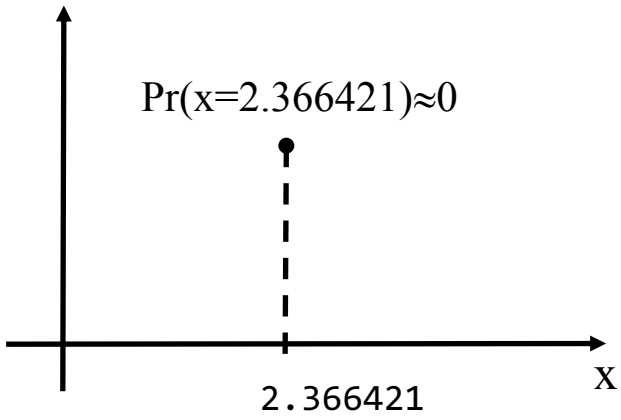
1.10 Continuous random variables

Now: random variable x can take values in the continuum (e.g. sensor signal)

It does not make sense to talk about probability that x has a particular value

Why?

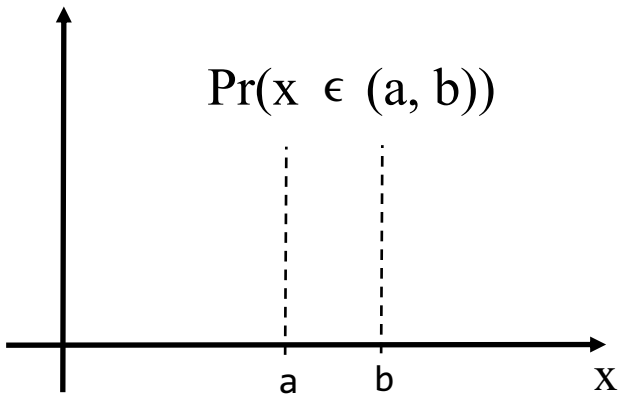
→ Probability of any particular exact value (e.g. 2.366421) will almost always be zero.



Because of the infinite number of different real number values hardly anyone of these values will be observed in experiments.

1.10 Continuous random variables

We rather talk about probability that x falls into some interval (a, b)

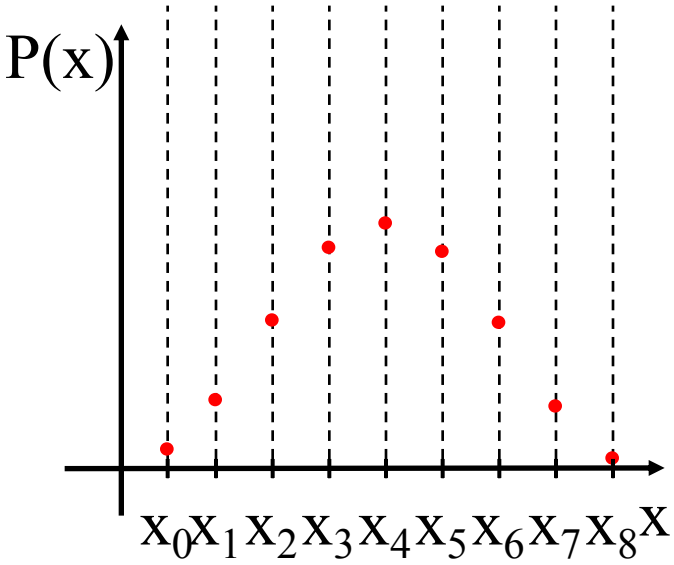


1.10 Continuous random variables

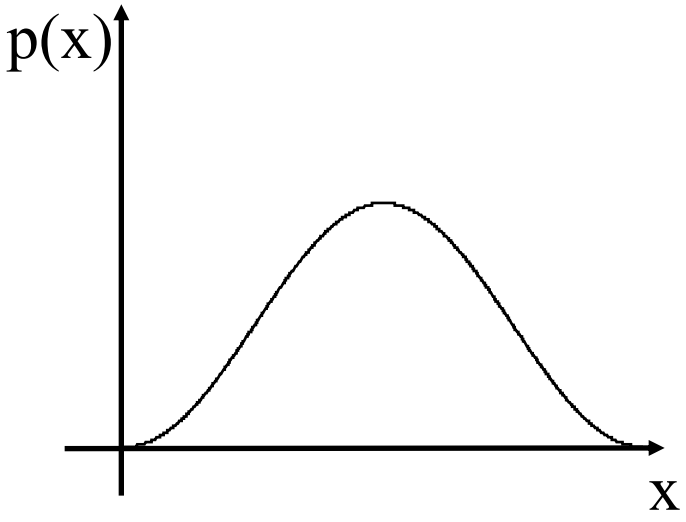
We rather talk about probability that x falls into some interval (a, b)

→ Probability Mass Function becomes a **Probability Density Function**

analogy with material density



Probability Mass Function



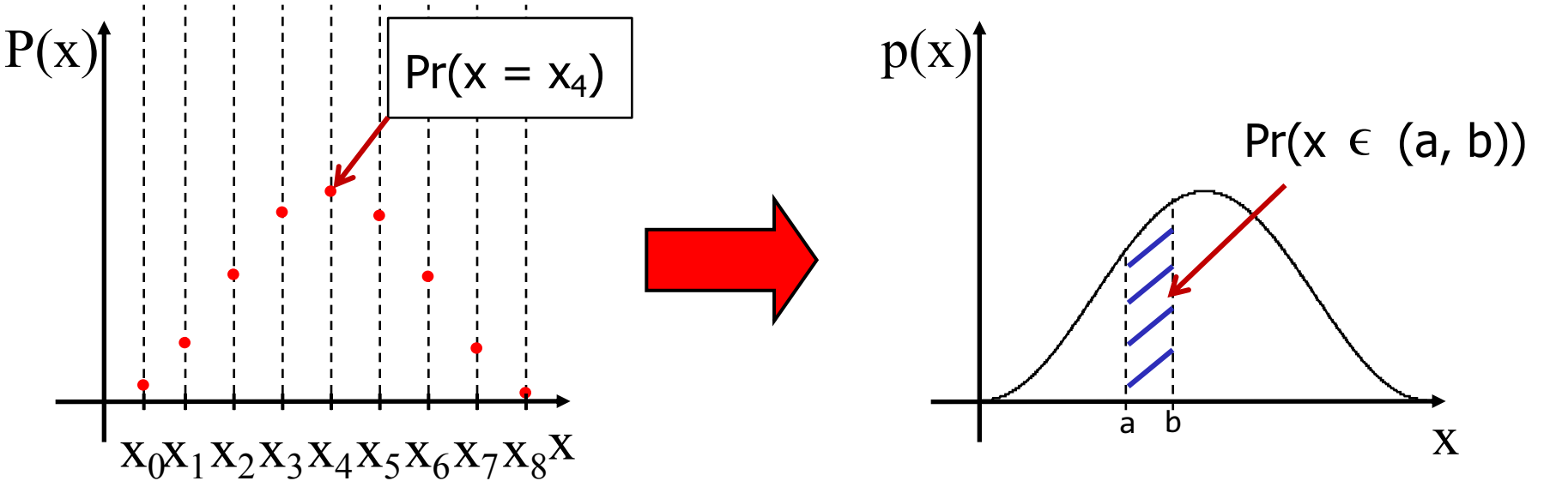
Probability Density Function

1.10 Continuous random variables

We rather talk about probability that x falls into some interval (a, b)

→ Probability Mass Function becomes a **Probability Density Function**

analogy with material density



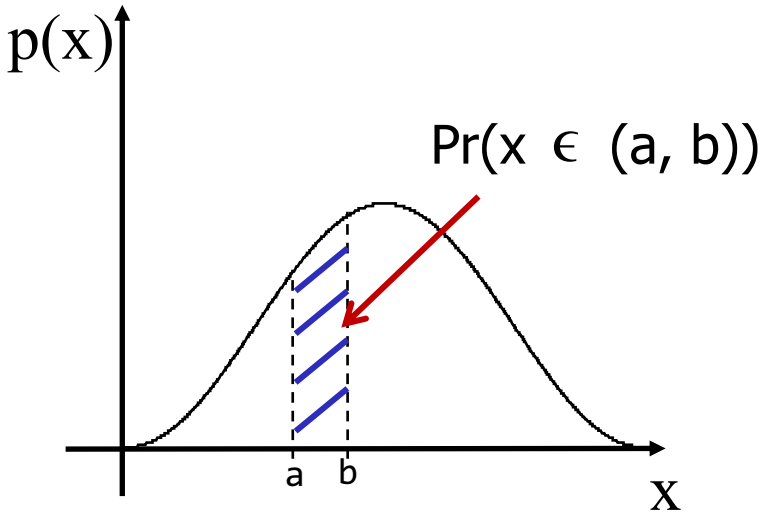
Probability Mass Function

Probability Density Function

1.10 Continuous random variables

Correspondence between probability Pr( ) and probability density p( ):

$$Pr(x \in (a,b)) = \int_a^b p(x)dx$$

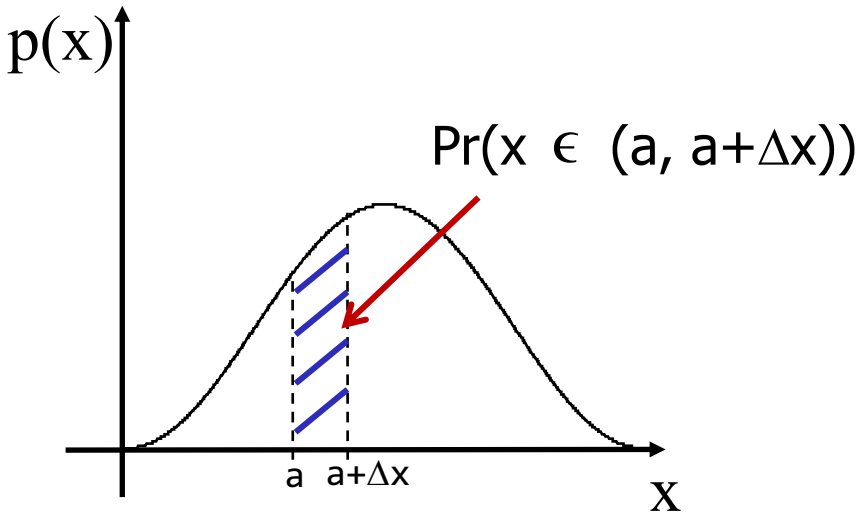


Probability Density Function

1.10 Continuous random variables

Correspondence between probability Pr( ) and probability density p( ):

$$p(x) = \lim_{\Delta x \to 0} \frac{Pr\{x \in (a, a + \Delta x)\}}{\Delta x}$$

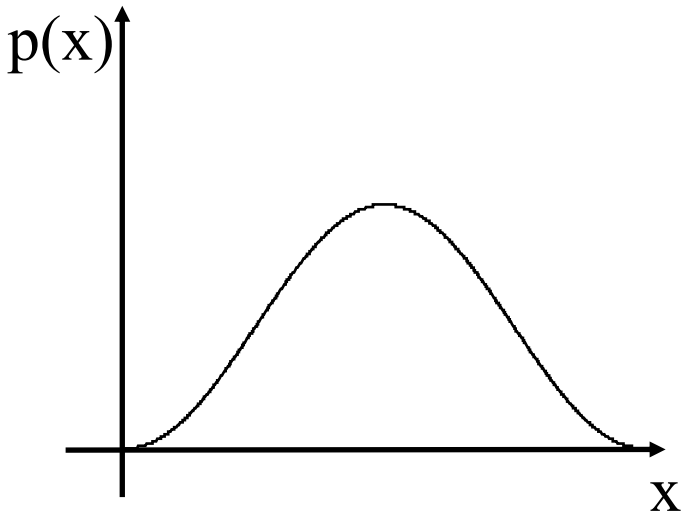


Probability Density Function

1.10 Continuous random variables

Properties of the probability density function:

$$p(x) \geq 0$$

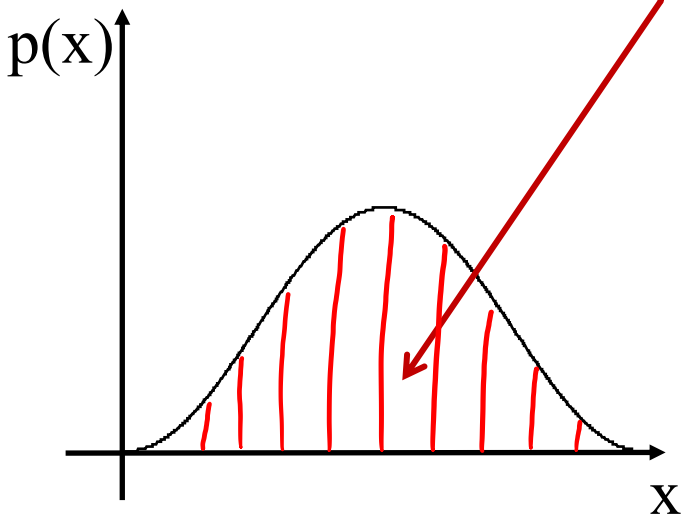


Probability Density Function

1.10 Continuous random variables

Properties of the probability density function:

$$\int_{-\infty}^{\infty} p(x)dx = 1$$



Probability Density Function

1.10 Continuous random variables

Most of the definitions and formulas for discrete random variables carry over to continuous random variables with sums replaced by integrals:

Expected value of a function f(x):

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$$

Mean

$$\mu = E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

Variance

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, p(x)\, dx$$

1.11 Normal distributions

Important result of the probability theory:

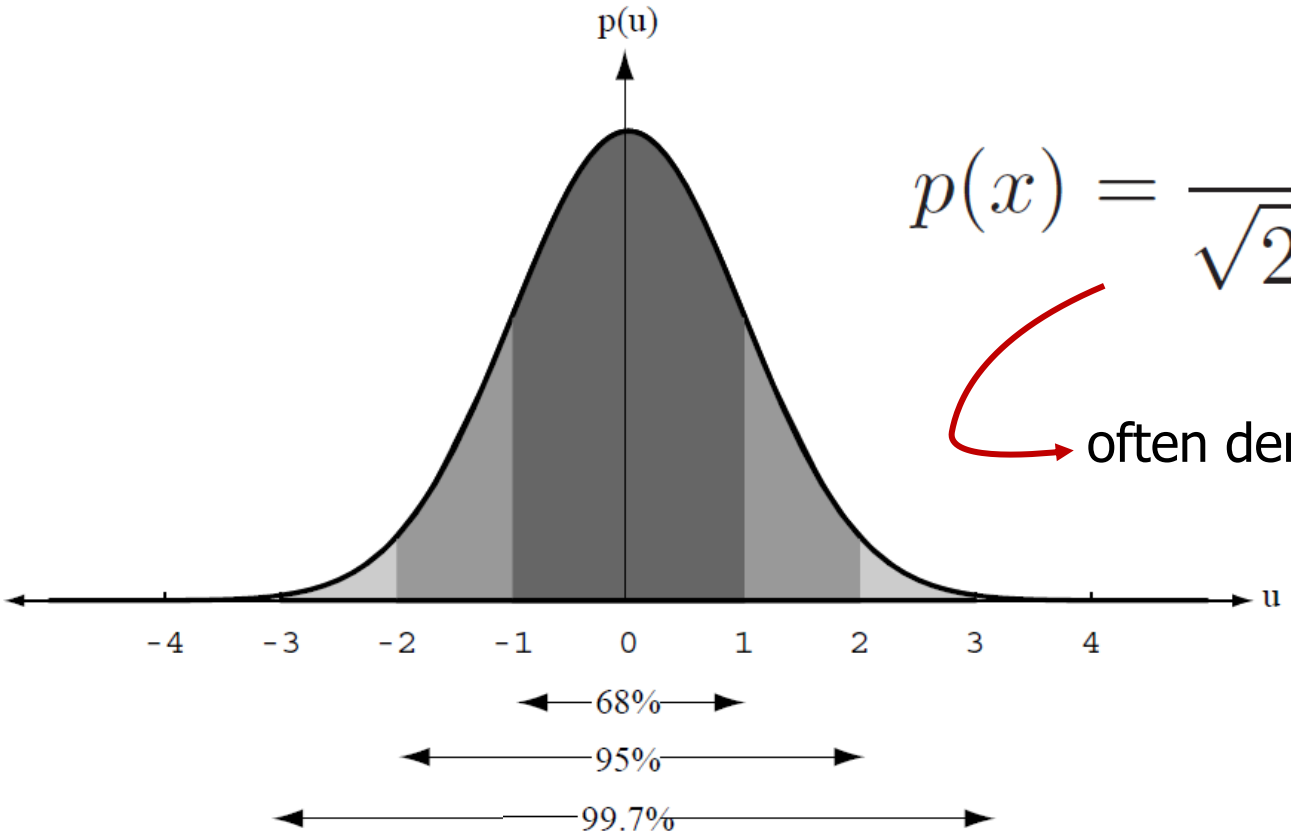
Central Limit Theorem:

Random variables generated under a large number of statistically independent influences can be considered *normally distributed:*

Representation by a Normal or **Gaussian** Probability Density Function

1.11 Normal distributions

Univariate Gaussian Distribution → 1-dimensional random variable x

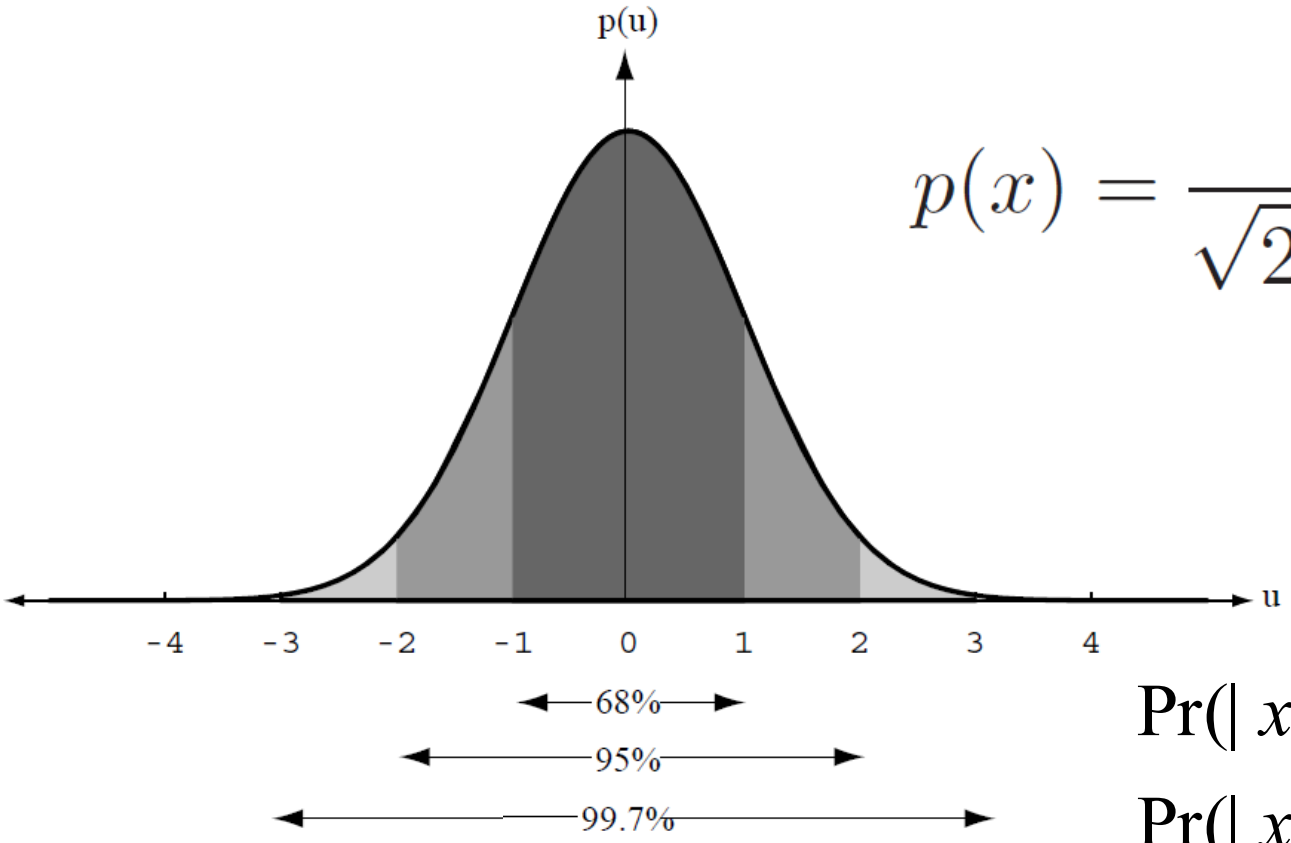


$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

often denoted as $N(\mu, \sigma^2)$

1.11 Normal distributions

Univariate Gaussian Distribution



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\Pr(|x-\mu| \leq \sigma) \cong 0.68$$

$$\Pr(|x-\mu| \leq 2\sigma) \cong 0.95$$

$$\Pr(|x-\mu| \leq 3\sigma) \cong 0.997$$

1.12 Mahalanobis Distance

Measure of the distance between an **observation x** and a **data set** with mean μ and standard deviation σ. „Is x similar to the data set?"

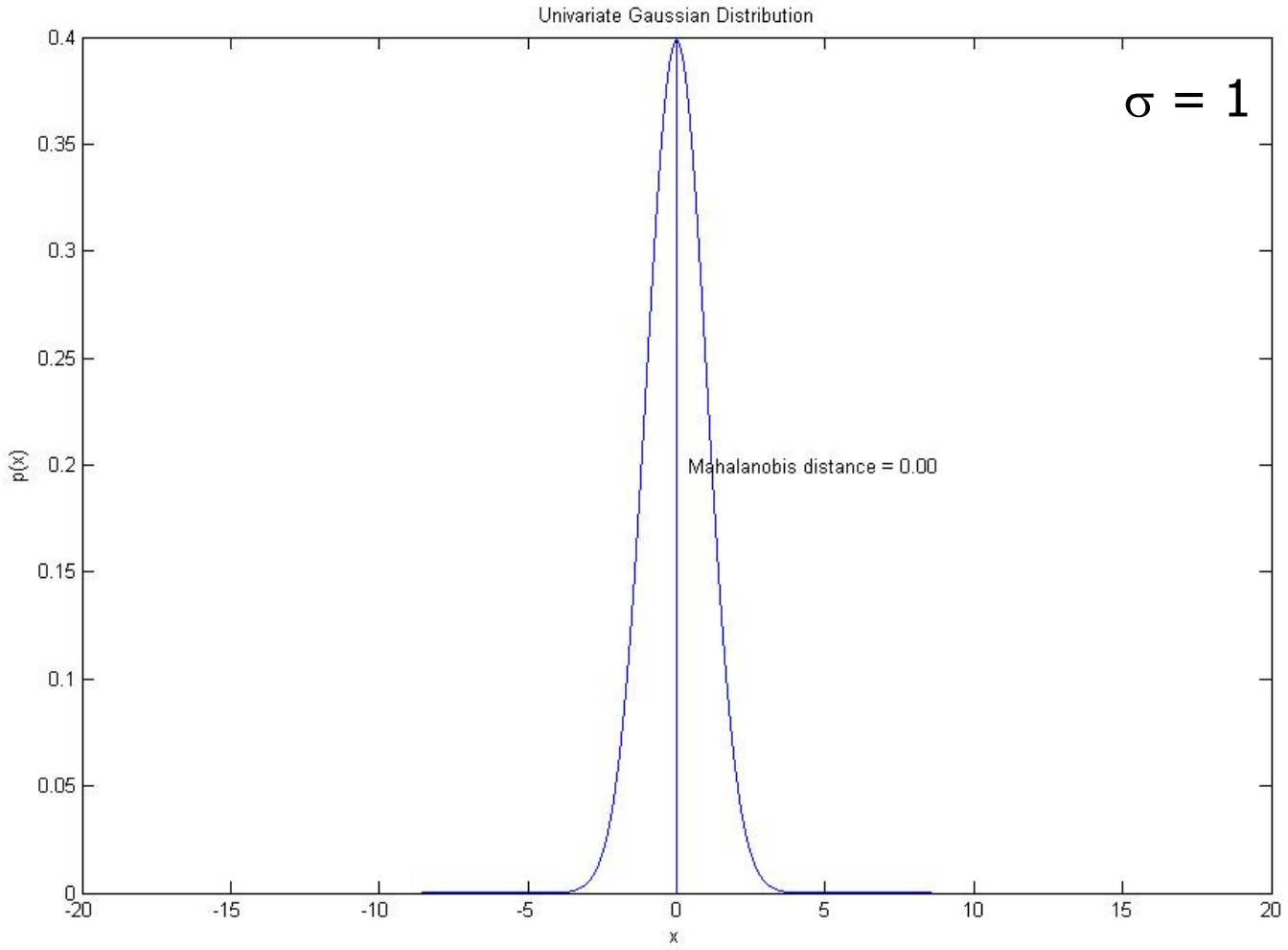
Based on the distance $|x-\mu|$ measured in units of the standard deviation σ:

Mahalanobis distance: $$d_M(x,\mu) = \frac{|x-\mu|}{\sigma}$$

→ smaller distance in case of larger variance

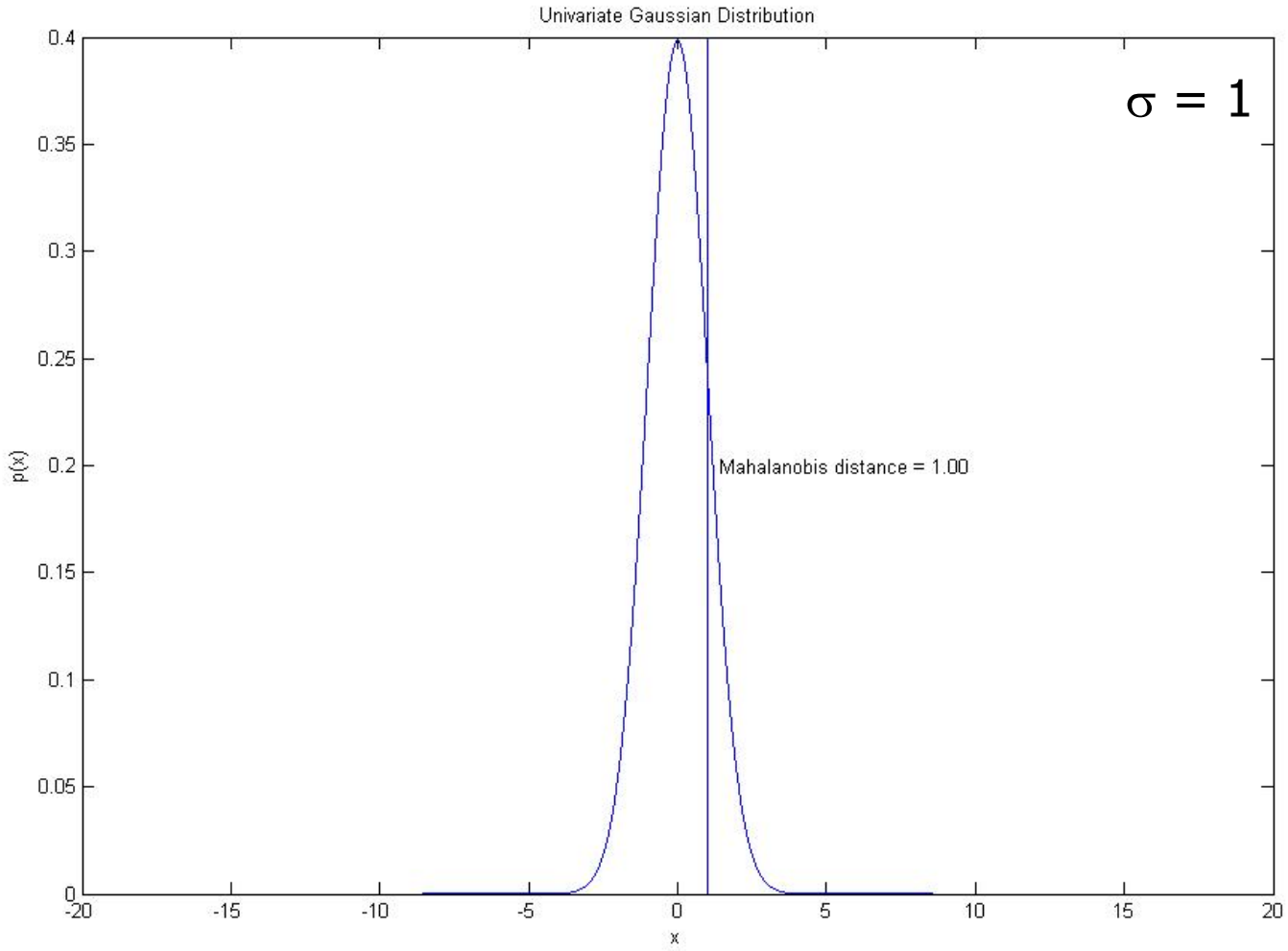
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



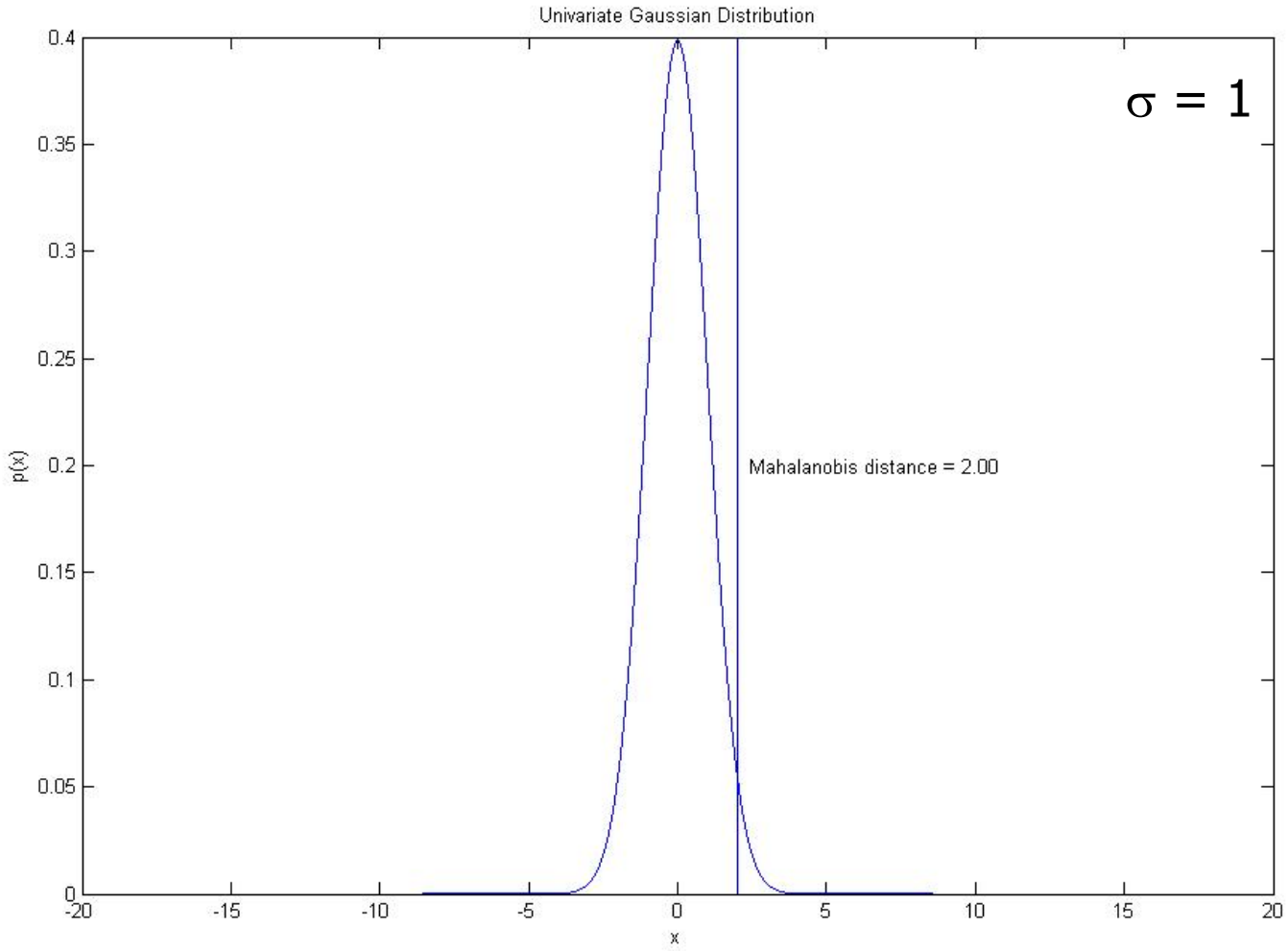
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



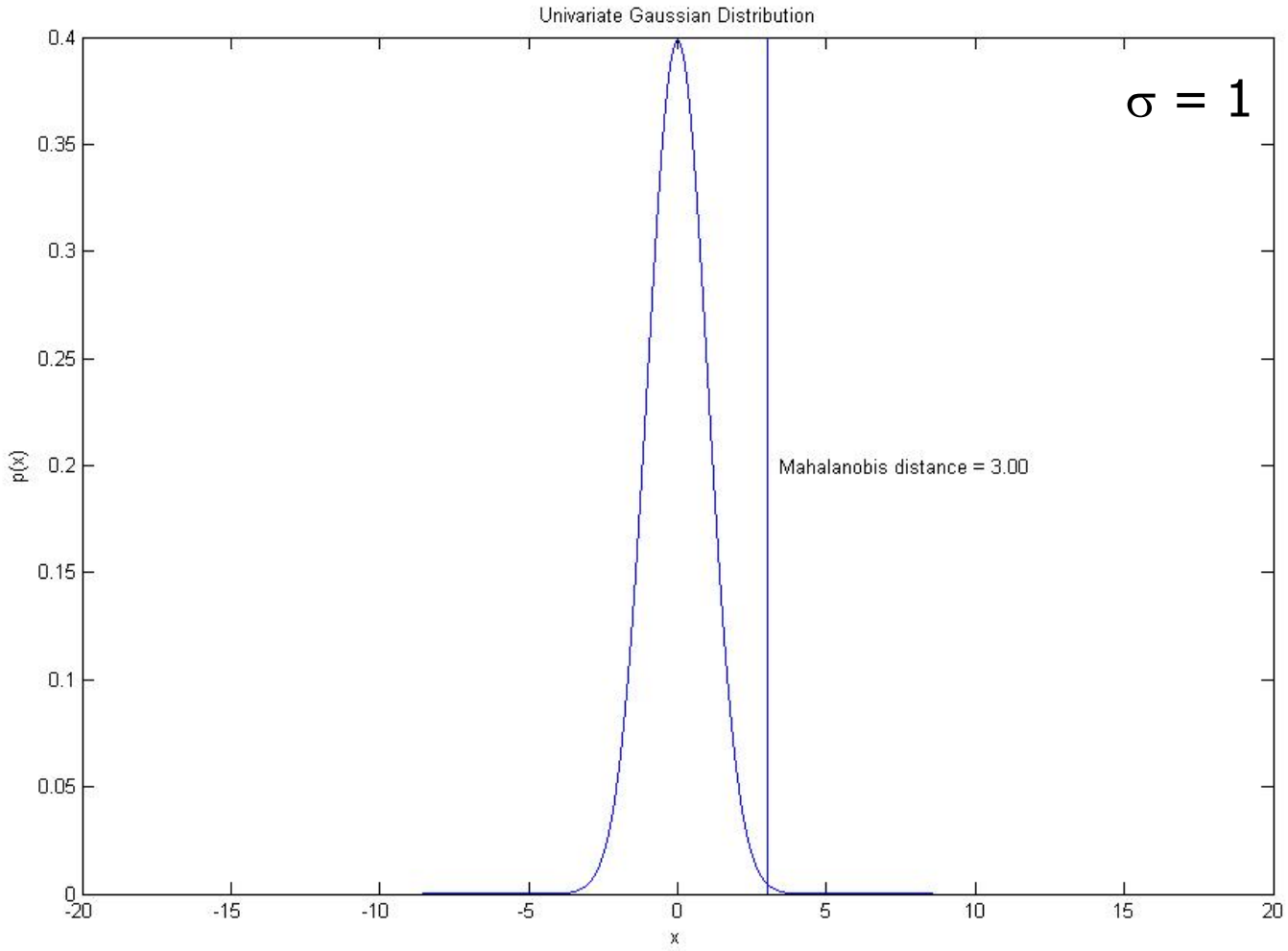
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



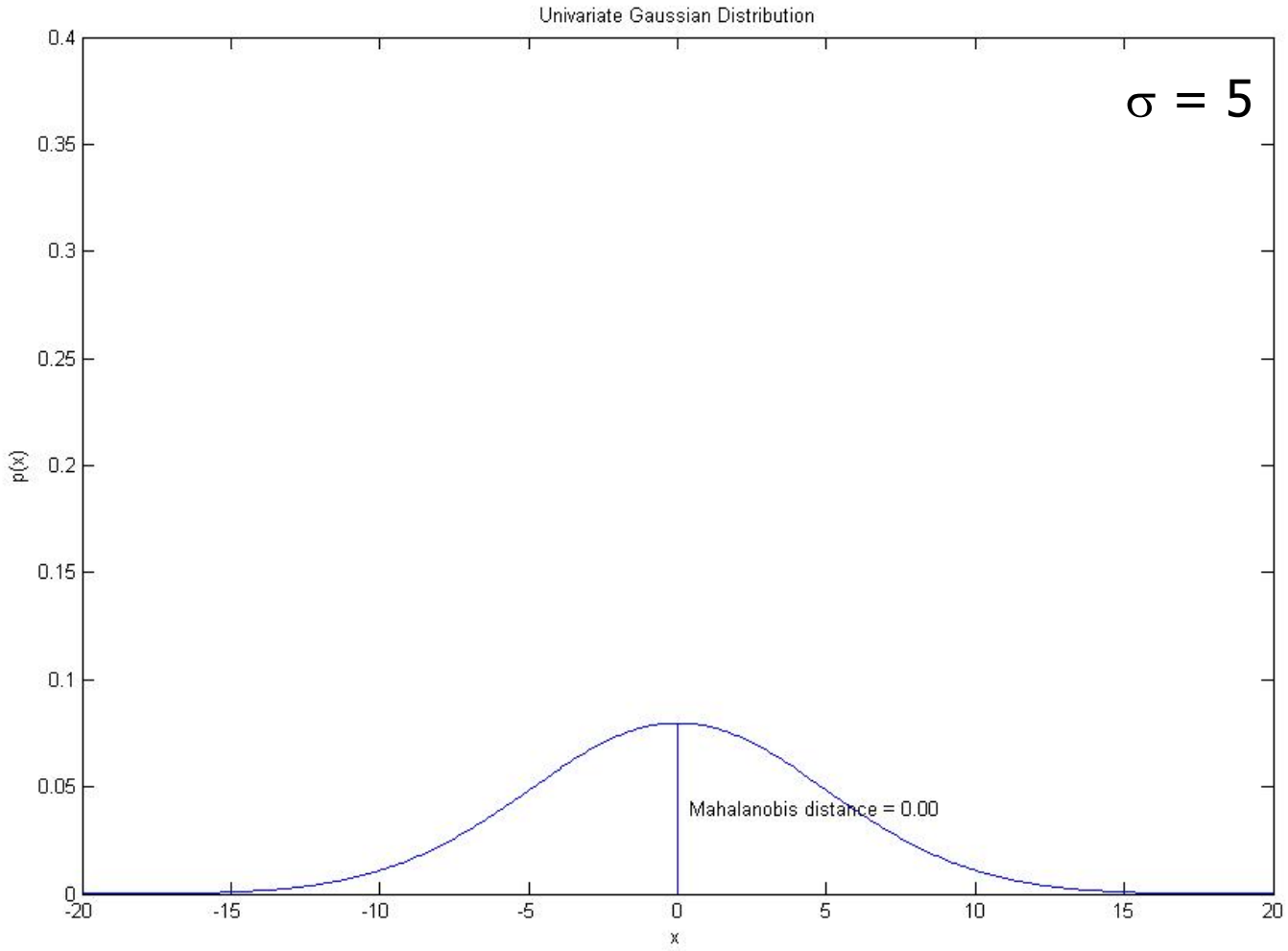
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



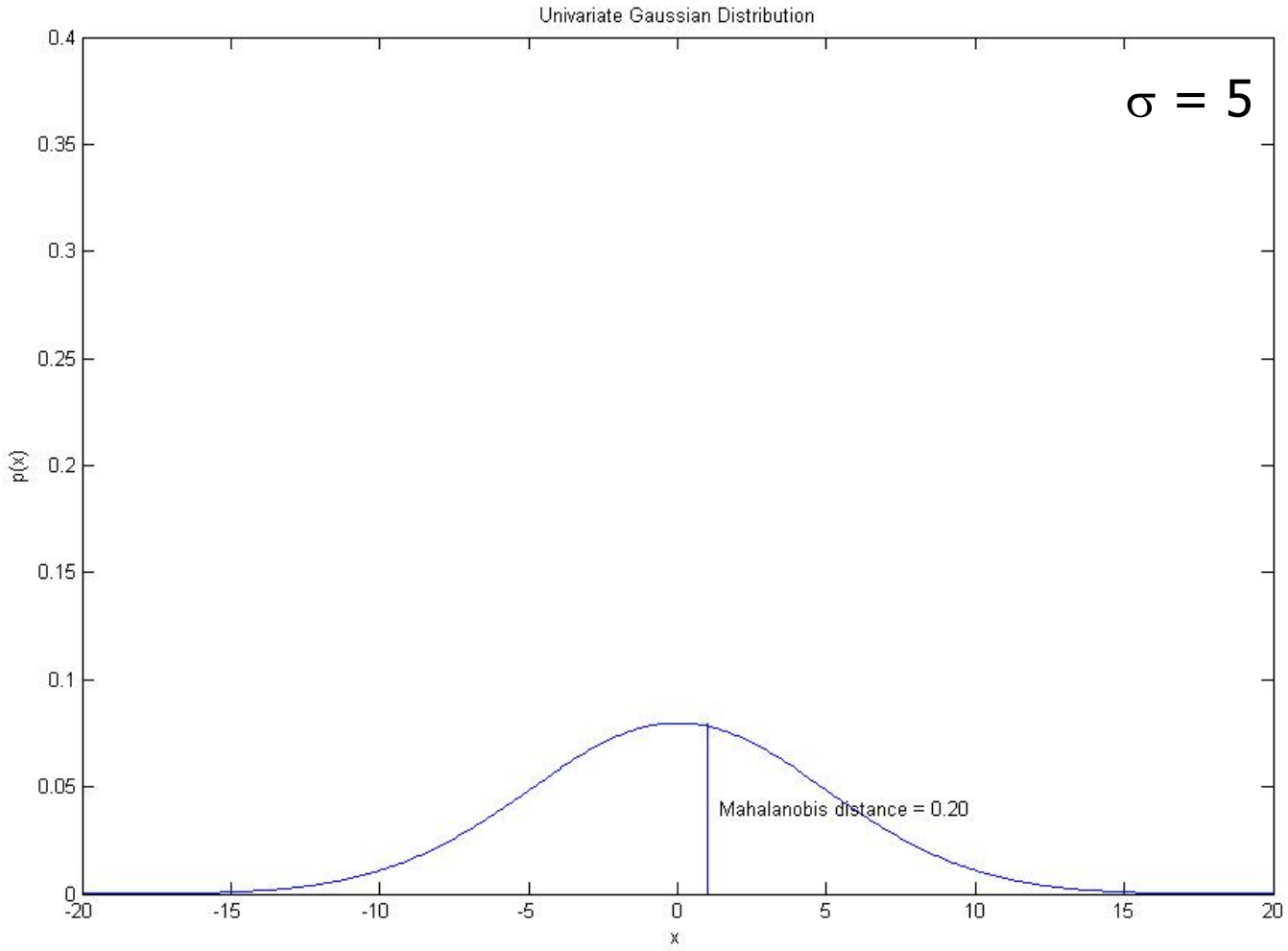
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



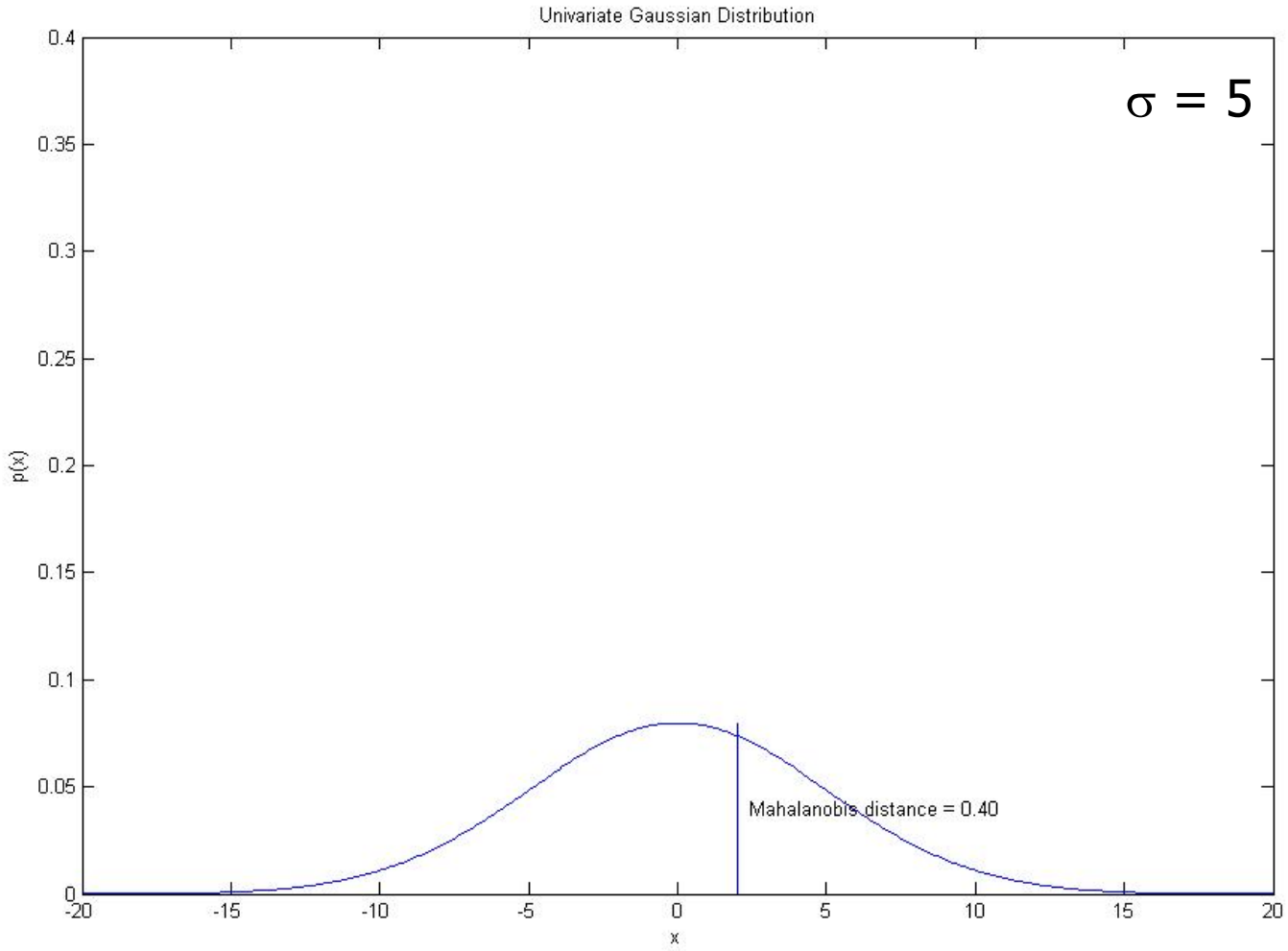
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



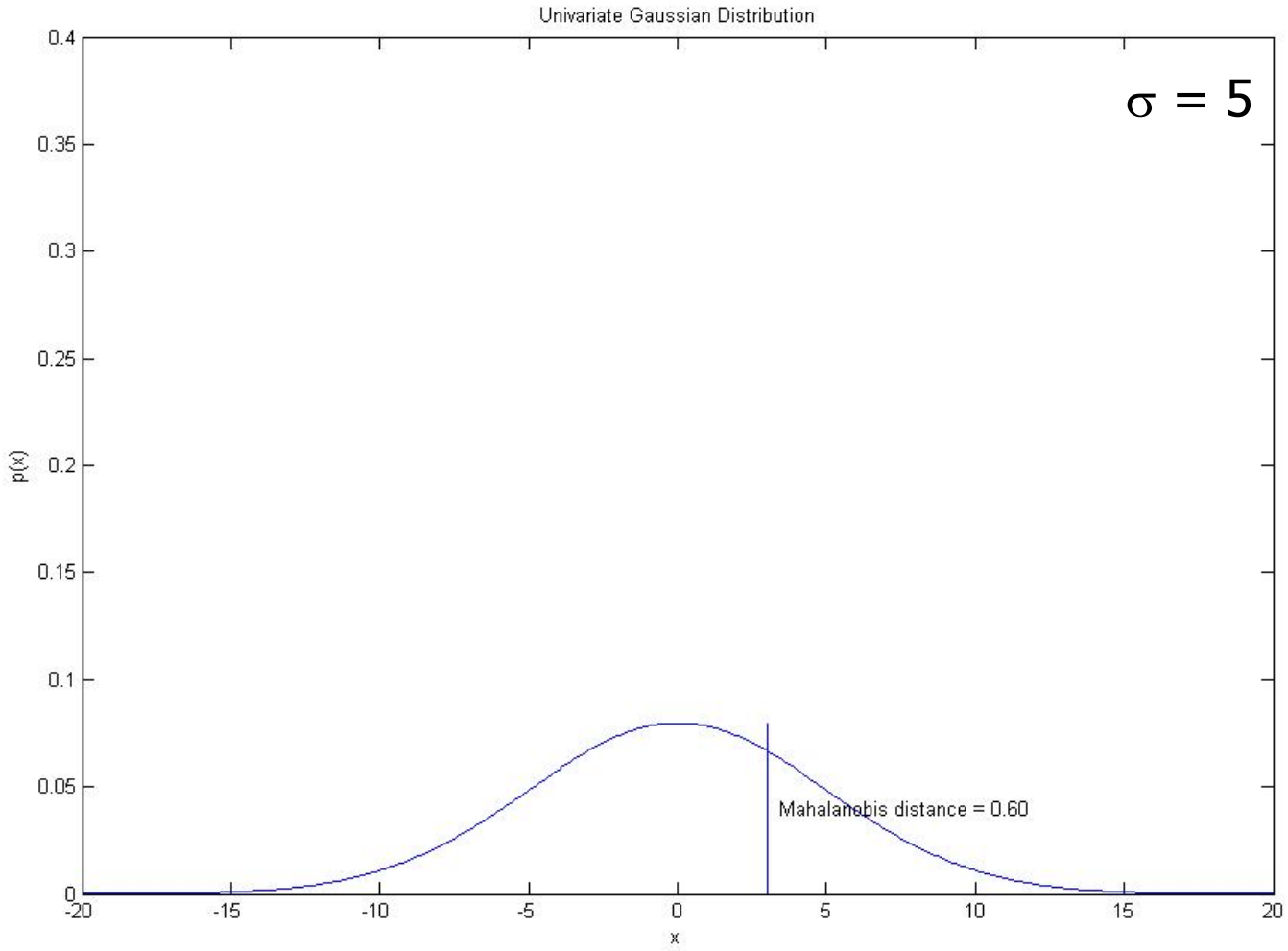
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



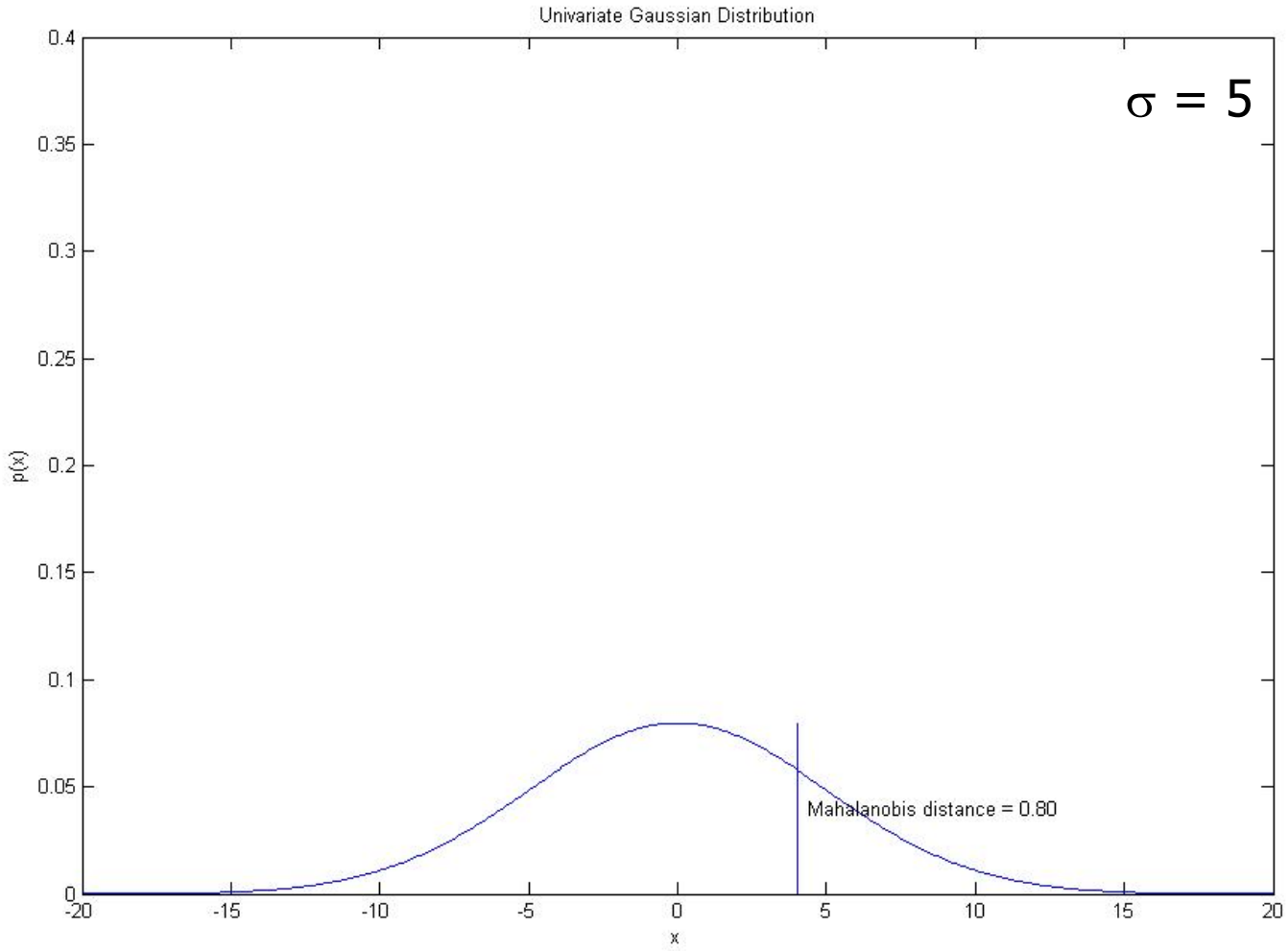
1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



1.12 Mahalanobis Distance

$$d_M(x, \mu) = \frac{|x - \mu|}{\sigma}$$



1.13 Multivariate Normal Densities

more than one random variable

1.13 Multivariate Normal Densities

We consider d normally distributed random variables x_i, each with its own mean μ_i and variance σ_i^2: $p(x_i) \sim N(\mu_i, \sigma_i^2)$

1.13 Multivariate Normal Densities

We consider d normally distributed random variables x_i, each with its own mean μ_i and variance σ_i^2: $p(x_i) \sim N(\mu_i, \sigma_i^2)$

Description of the joint probability density function? $p(x_1, x_2, \ldots, x_d)$

1.13 Multivariate Normal Densities

We consider d normally distributed random variables x_i, each with its own mean μ_i and variance σ_i^2:

$$p(x_i) \sim N(\mu_i, \sigma_i^2)$$

Description of the joint probability density function?

$$p(x_1, x_2, \ldots, x_d)$$

Case 1: Statistically independent variables

$$p(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\cdot\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$

Multivariate Gaussian Density With Independent Components

1.13 Multivariate Normal Densities

We consider d normally distributed random variables x_i, each with its own mean μ_i and variance σ_i^2:

$$p(x_i) \sim N(\mu_i, \sigma_i^2)$$

Description of the joint probability density function?

$$p(x_1, x_2, \ldots, x_d)$$

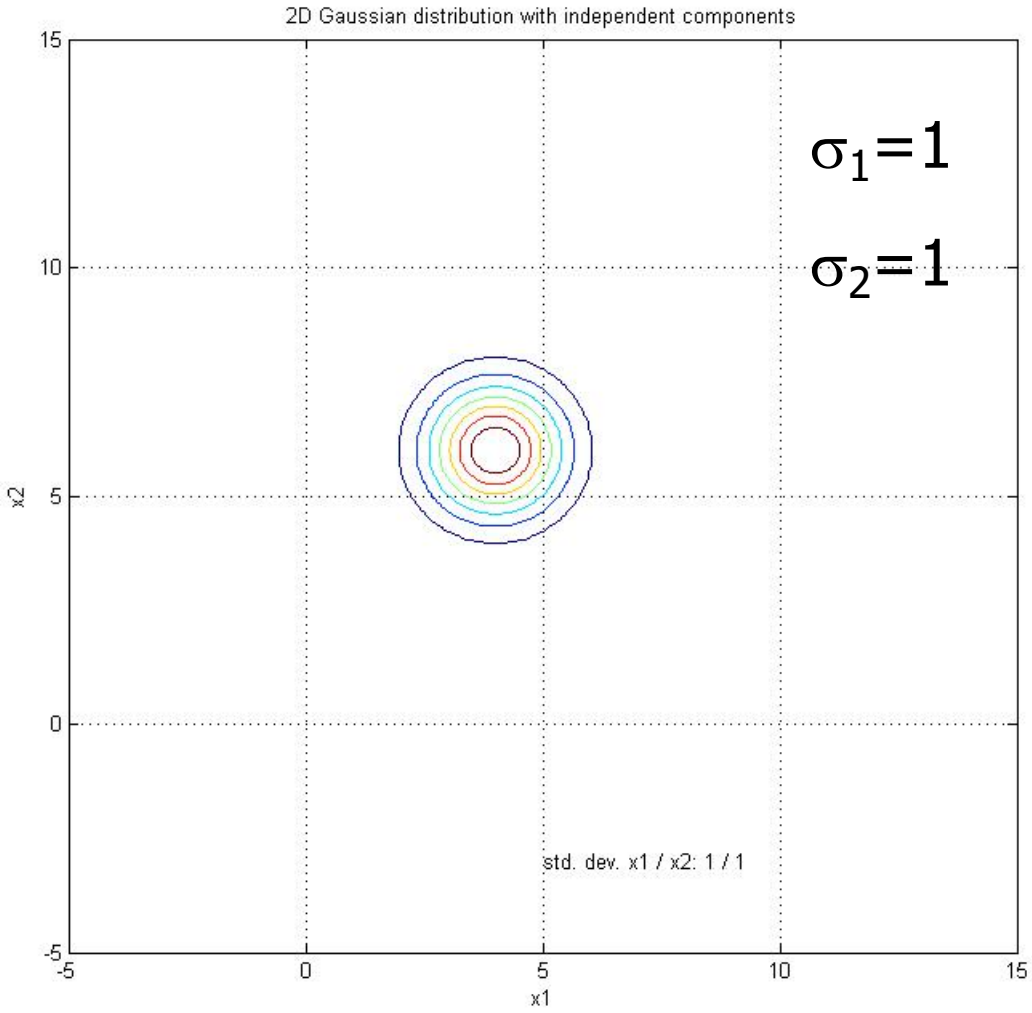
Case 1: Statistically independent variables

$$p(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} p(x_i) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\cdot\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2}$$

2-D case: $p(x_1, x_2) = p(x_1) \cdot p(x_2)$

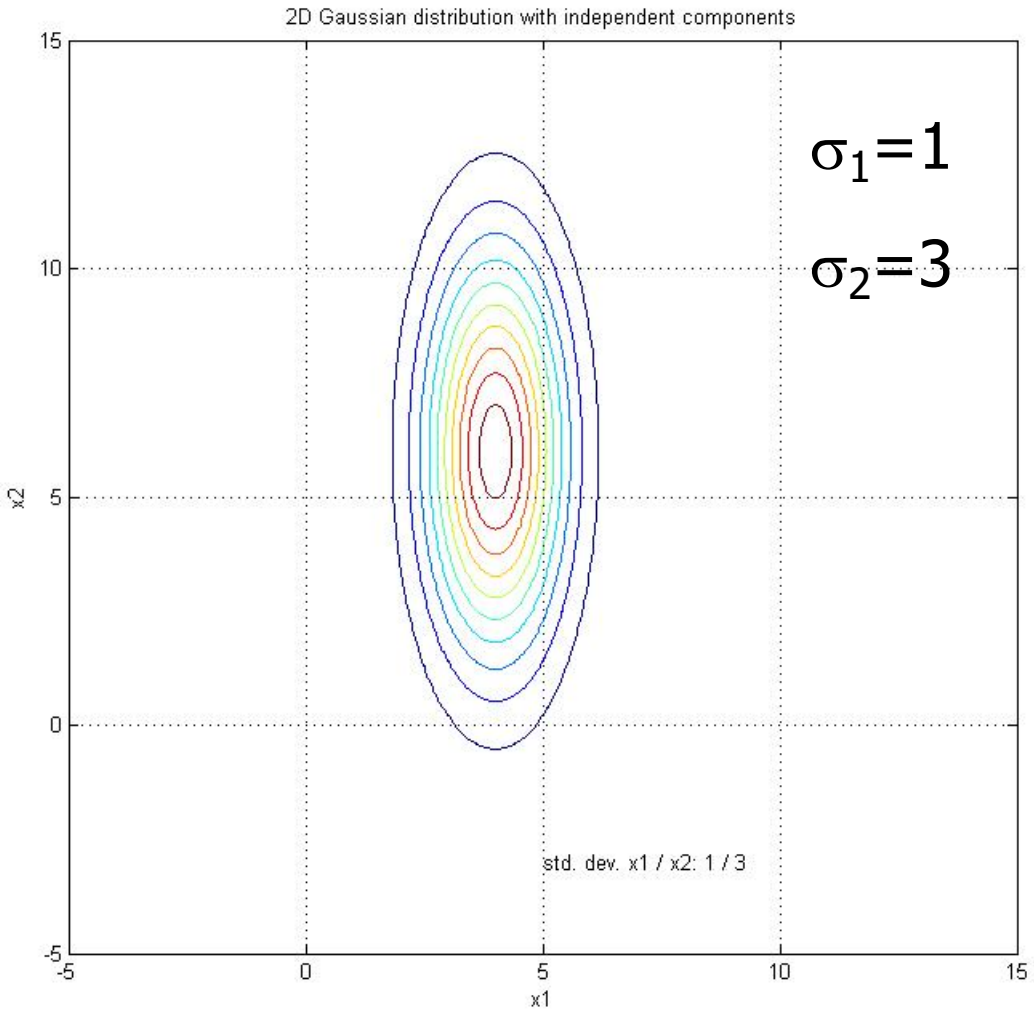
1.13 Multivariate Normal Densities

2D-Example



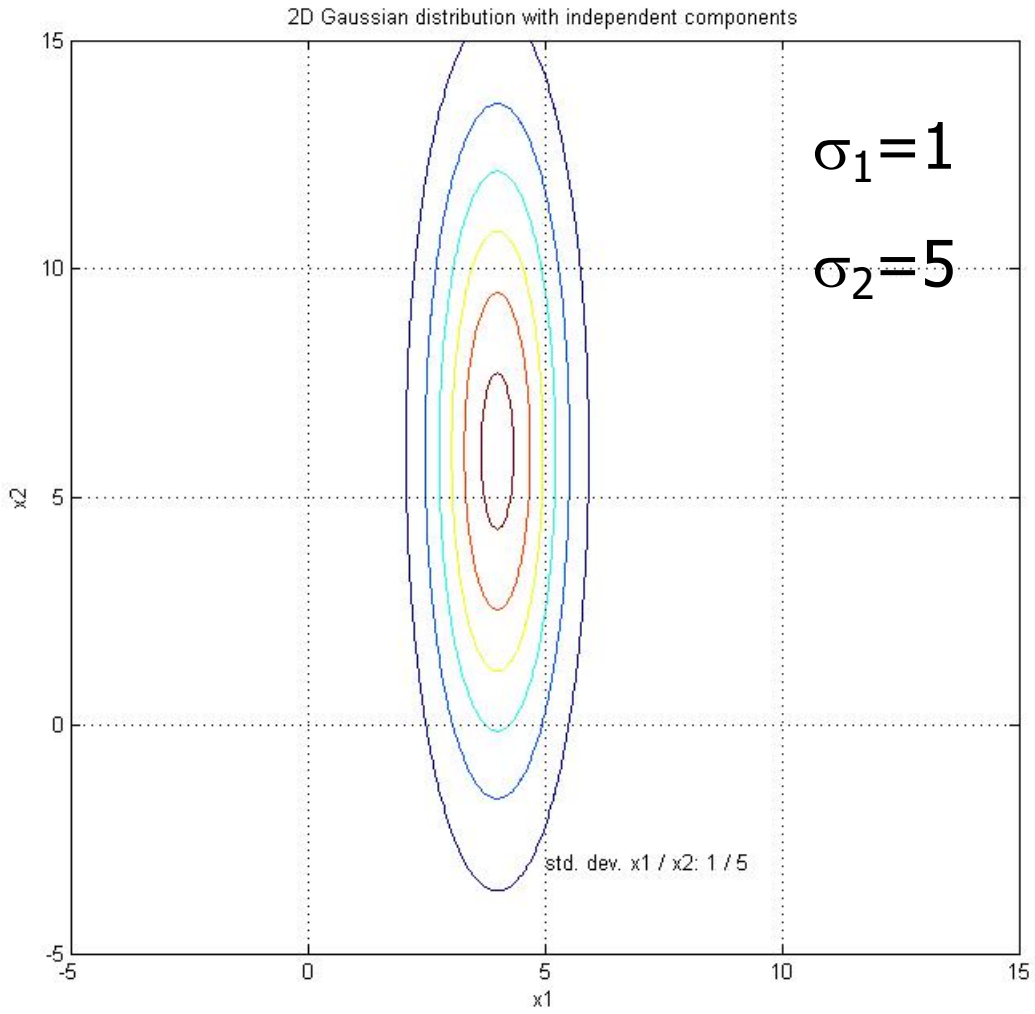
1.13 Multivariate Normal Densities

2D-Example



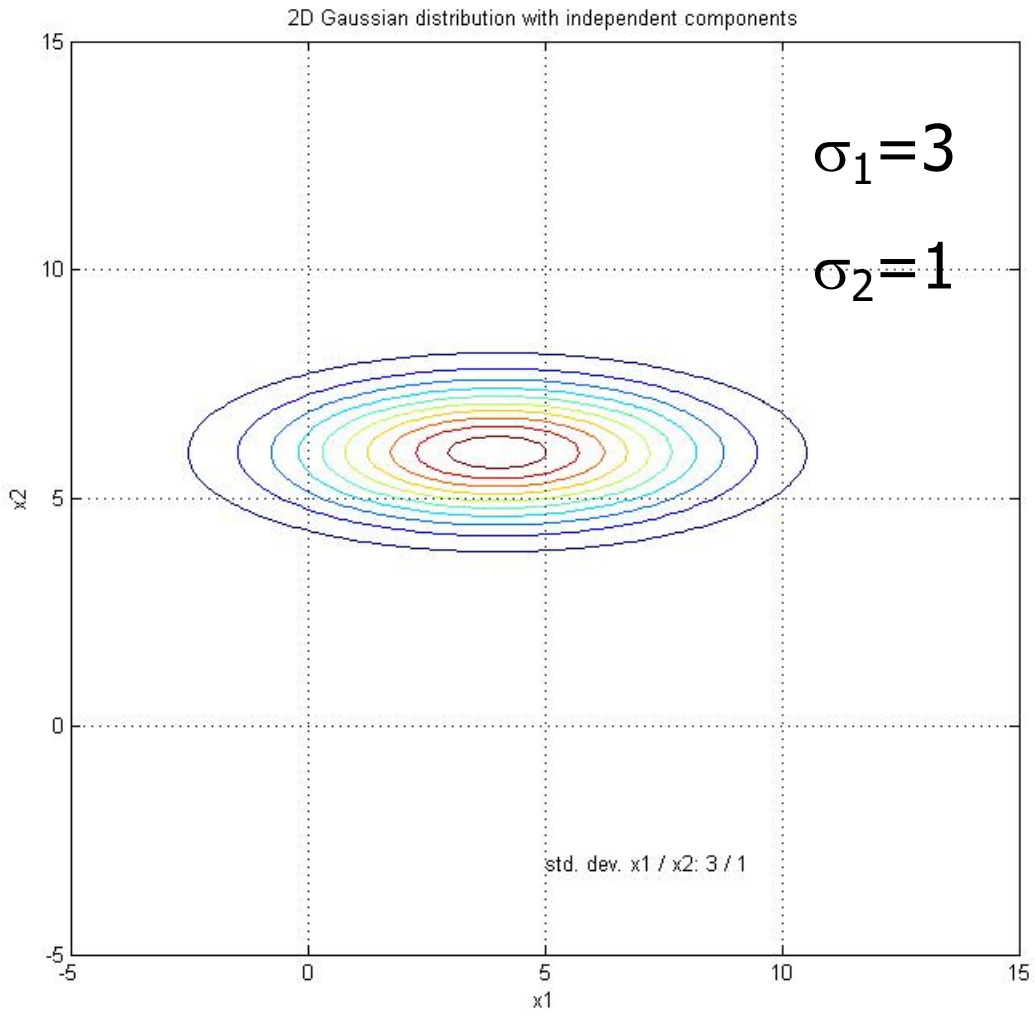
1.13 Multivariate Normal Densities

2D-Example



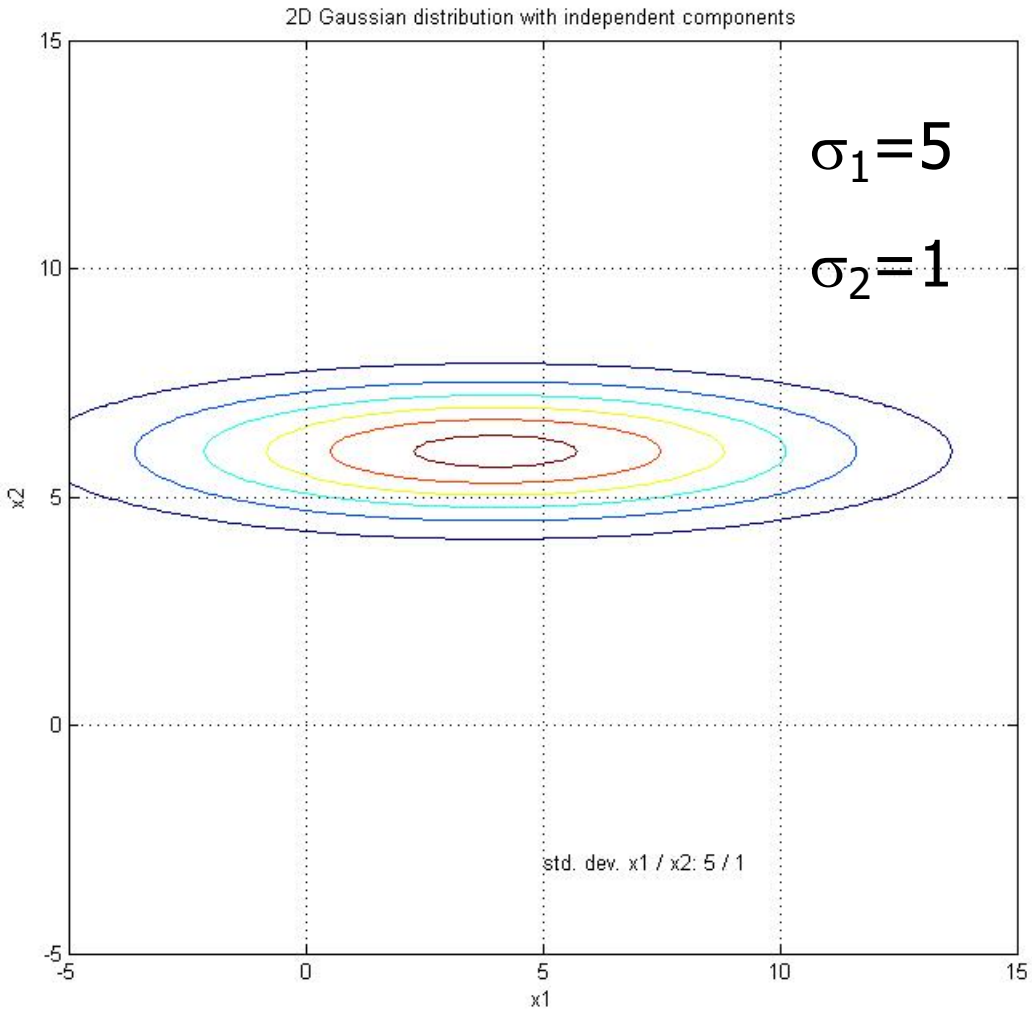
1.13 Multivariate Normal Densities

2D-Example



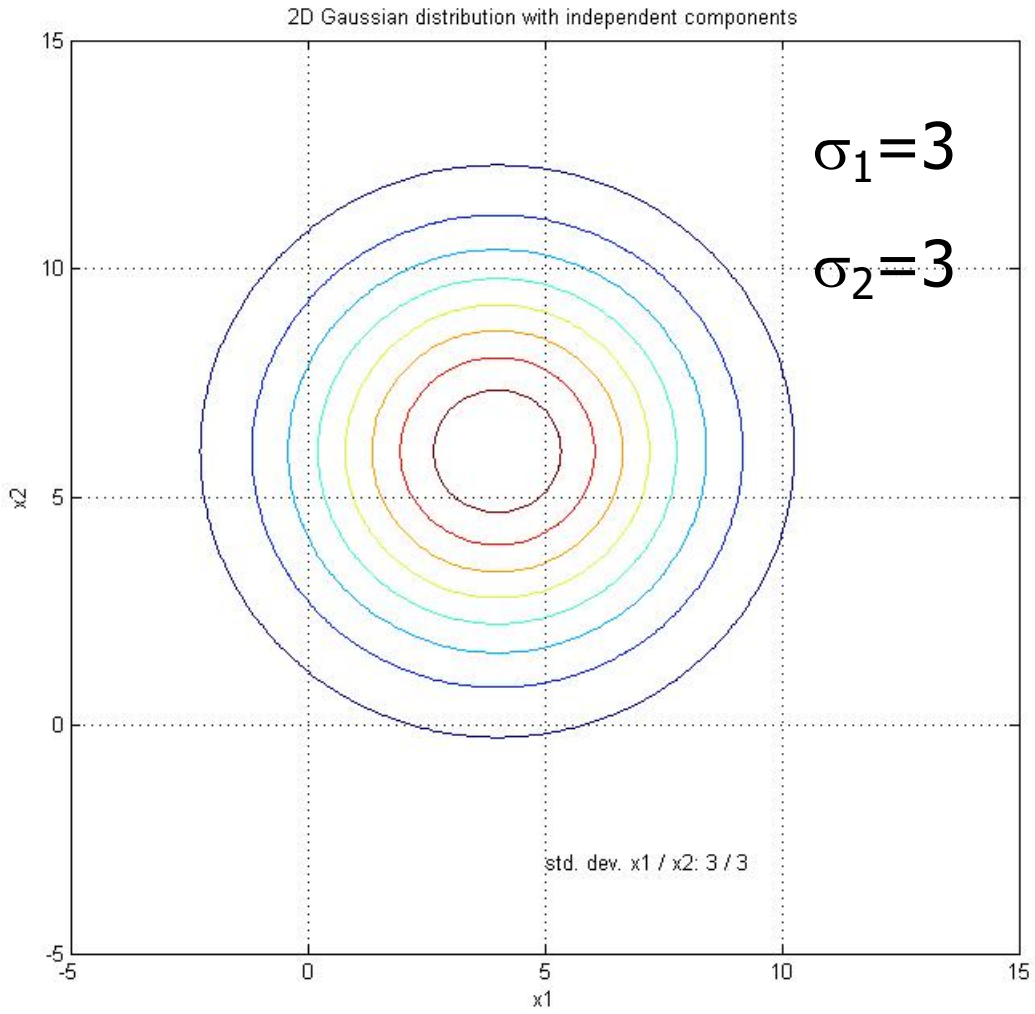
1.13 Multivariate Normal Densities

2D-Example



1.13 Multivariate Normal Densities

2D-Example



1.13 Multivariate Normal Densities

Why do we need this?

→ Application example: **Fish classification**

Example: Representation of fish samples using Gaussian Densities

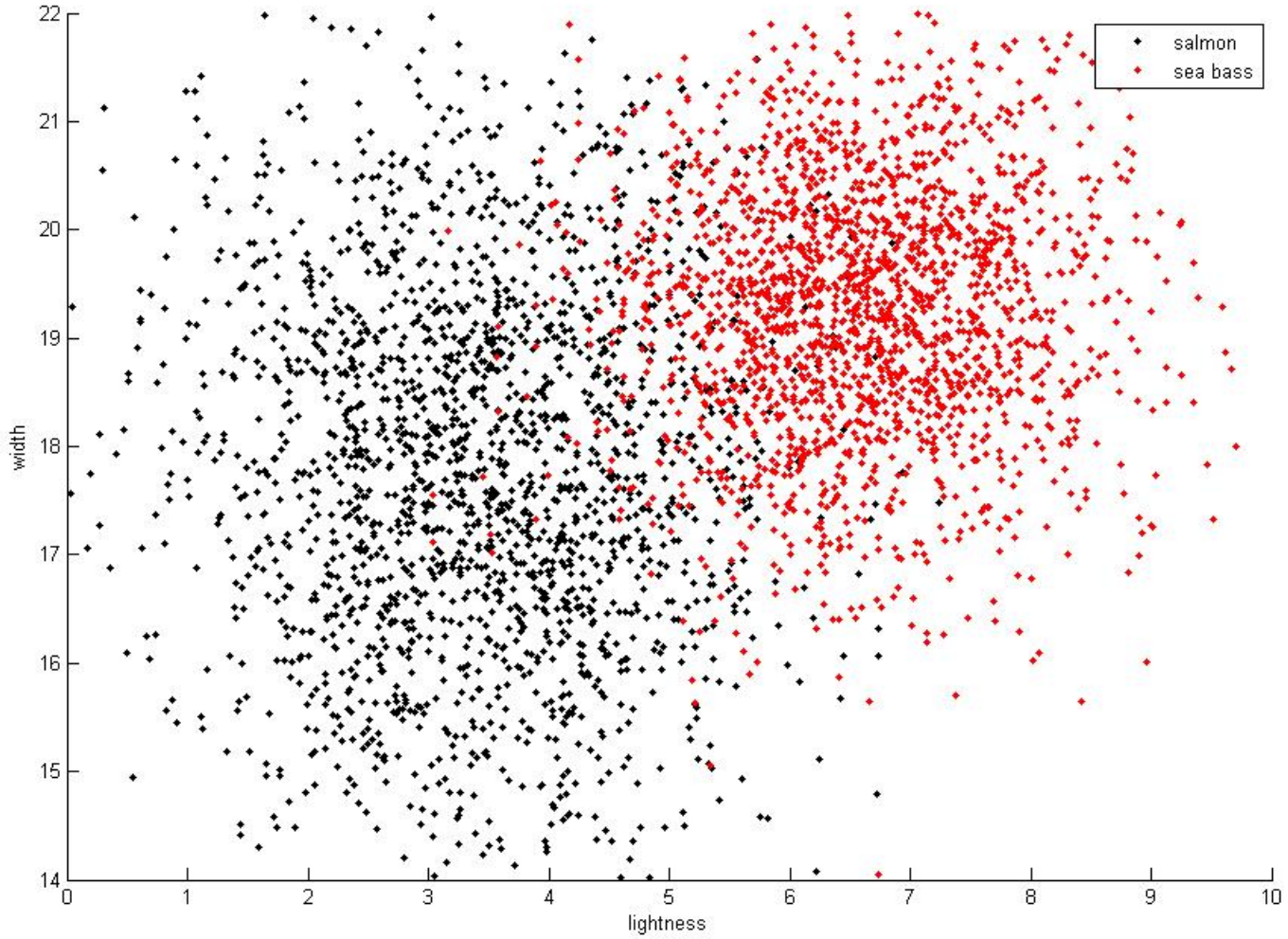
Step 1: Collection of 2D sensor data (lightness, width) for large training set

Practical realization: Data is organized in a Matlab array

```
>> salmon(1:5,:)
ans =
    3.6642   20.0939
    4.9162   15.3323
    3.9818   16.6113
    5.3453   15.4816
    2.5478   18.4773
```

First salmon: Lightness = 3.6642, Width = 20.0939

Example: Representation of fish samples using Gaussian Densities



Example: Representation of fish samples using Gaussian Densities

Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

2a. Compute mean and variance of each random variable:

```
>> load salmon_artificial.mat       % available in OLAT material folder

>> m1 = mean(salmon(:,1));          % lightness mean

>> s1 = sqrt(var(salmon(:,1)));     % lightness standard deviation

>> m2 = mean(salmon(:,2));          % width mean

>> s2 = sqrt(var(salmon(:,2)));     % width standard deviation
```

Example: Representation of fish samples using Gaussian Densities

Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

2b. Compute marginal distributions

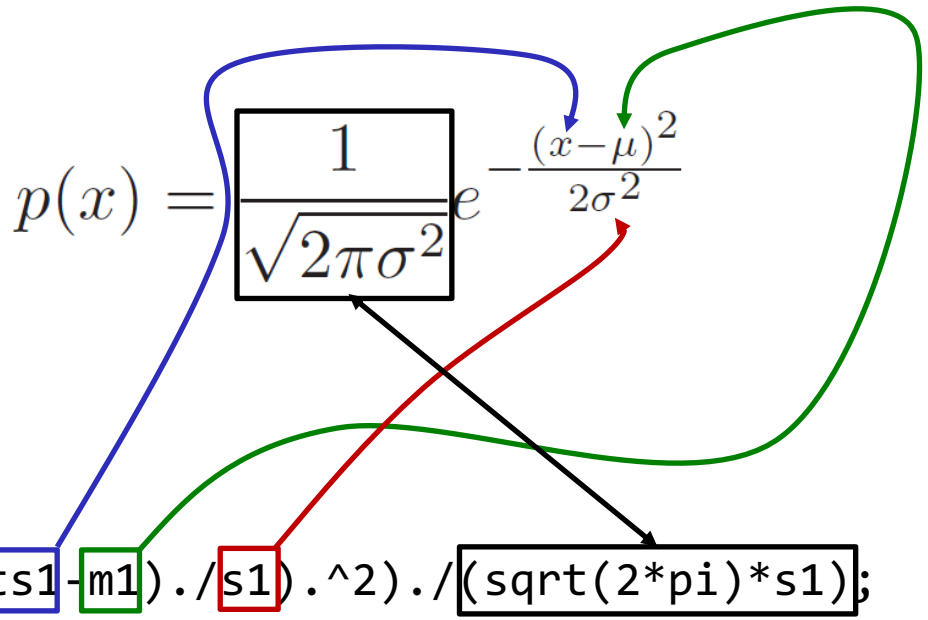
```
%  Compute lightness values

>> pts1 = min(salmon(:,1)):0.1:max(salmon(:,1));

%  Compute width values

>> pts2 = min(salmon(:,2)):0.1:max(salmon(:,2));

%  Compute marginals: px1 is p(lightness), px2 is p(width)

>> px1 = exp(-0.5*((pts1-m1)./s1).^2)./(sqrt(2*pi)*s1);

>> px2 = exp(-0.5*((pts2-m2)./s2).^2)./(sqrt(2*pi)*s2);
```

Example: Representation of fish samples using Gaussian Densities

Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

2b. Compute marginal distributions




```
>> px1 = exp(-0.5*((pts1-m1)./s1).^2)./(sqrt(2*pi)*s1);

>> px2 = exp(-0.5*((pts2-m2)./s2).^2)./(sqrt(2*pi)*s2);
```

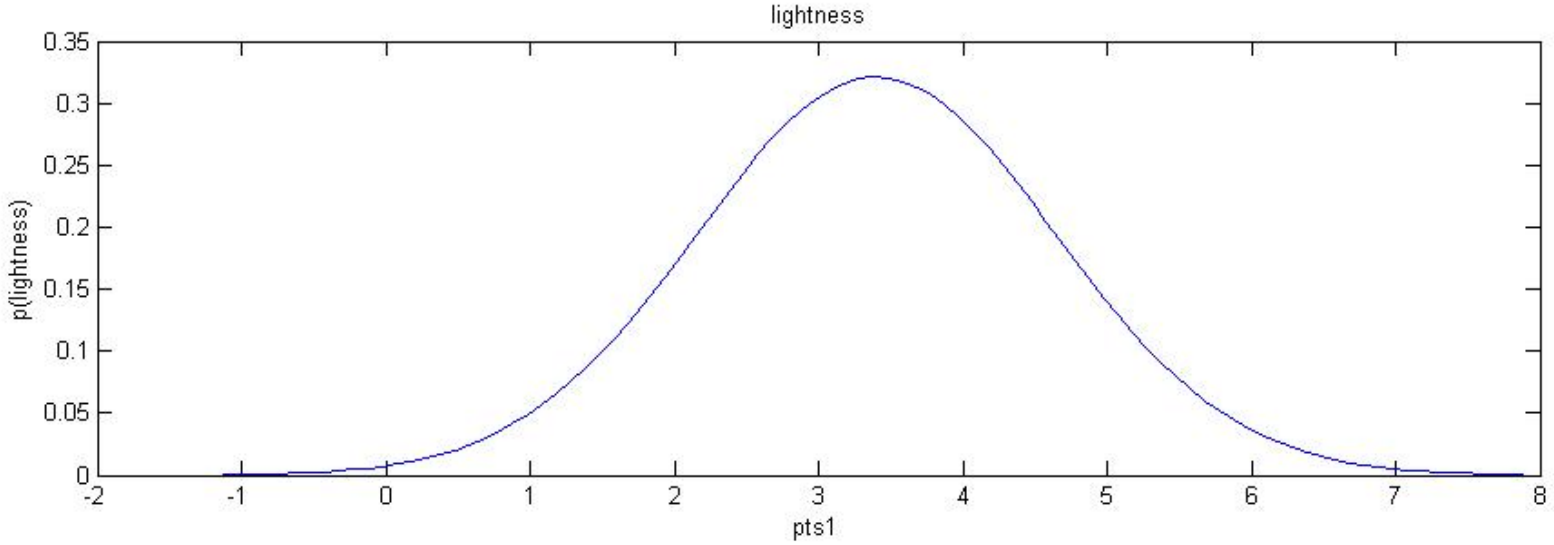
Example: Representation of fish samples using Gaussian Densities

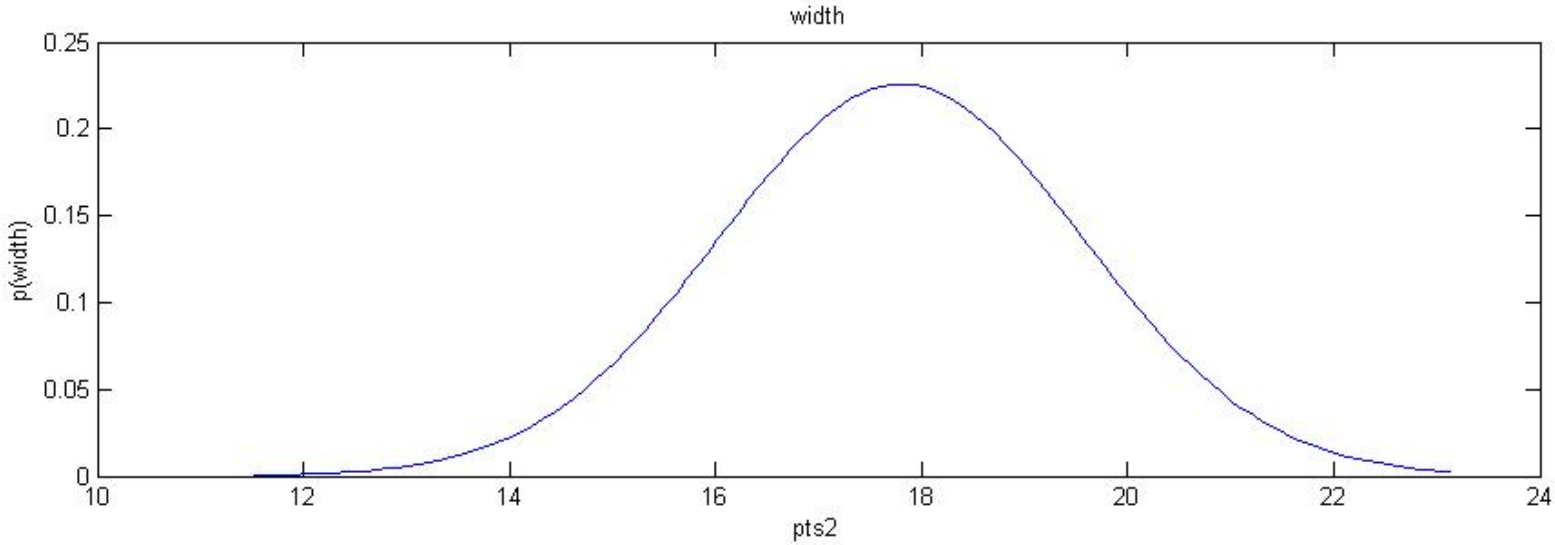
Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

2b. Visualize marginal distributions

```
>> subplot(311);
>> plot(pts1, px1)
>> title('lightness')
>> xlabel('pts1')
>> ylabel('p(lightness)')
>> subplot(312);
>> plot(pts2, px2)
>> title('width')
>> xlabel('pts2')
>> ylabel('p(width)')
>> xlabel('pts2')
```


lightness
p(lightness)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-2
-1
0
1
2
3
4
5
6
7
8
pts1


width
p(width)
0.25
0.2
0.15
0.1
0.05
0
10
12
14
16
18
20
22
24
pts2

Example: Representation of fish samples using Gaussian Densities

Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

2c. Compute joint probability p(lightness, width) and plot contour lines

```
%  joint probability

>> p_x1x2_sal = px2'*px1;  % column vector * row vector = matrix
```

$$P(x, y) = P_x(x)\ P_y(y)$$

Example:

```
>> a = [1, 2]'
>> b = [3, 4]
>> a * b
     3    4
     6    8
```

Example: Representation of fish samples using Gaussian Densities

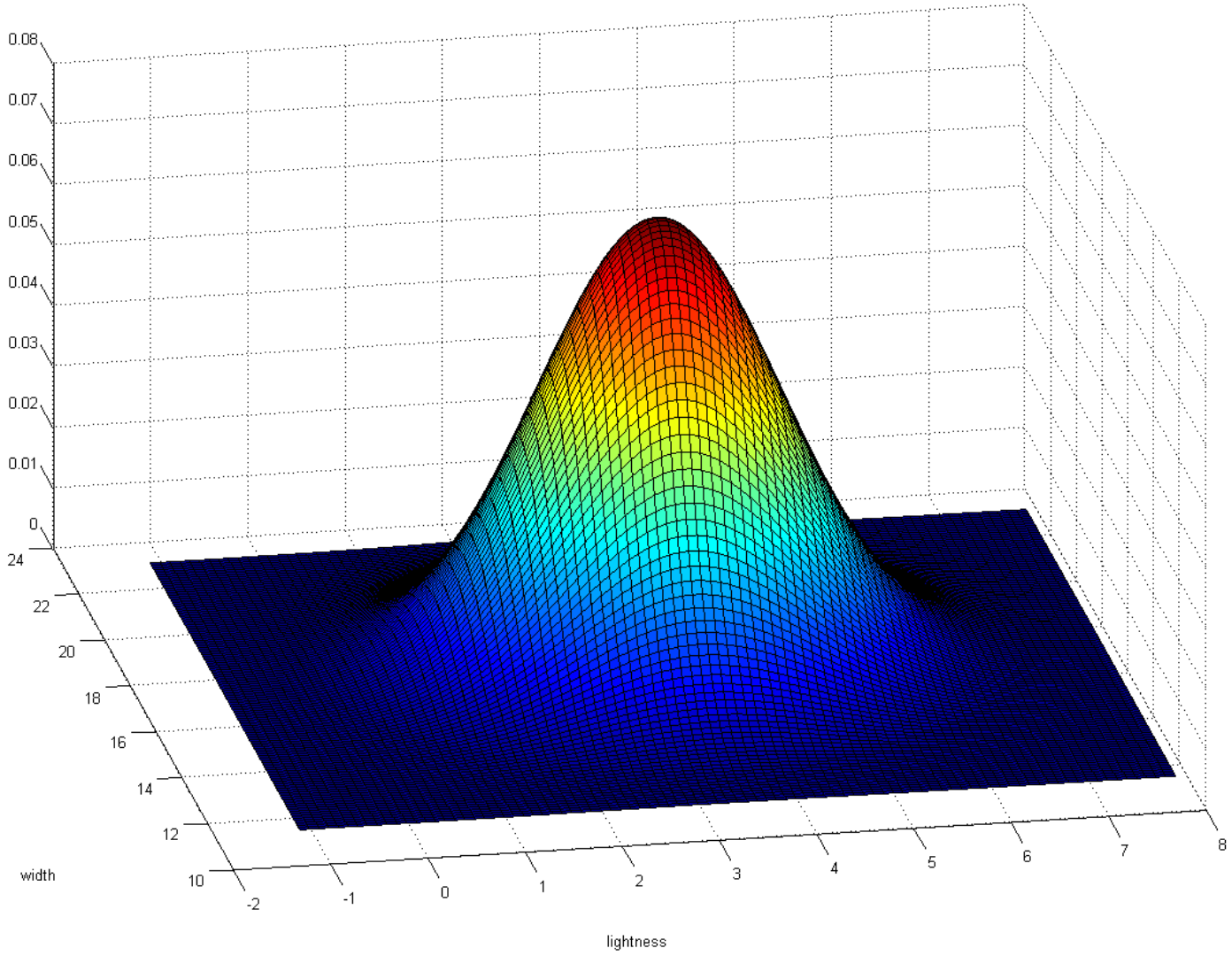
Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

2c. Compute joint probability p(lightness, width) and plot contour lines

```
%  visualize joint probability
>> subplot(313);
>> surf(pts1,pts2,p_x1x2_sal);
>> title('2-D Gaussian density with independent components');
>> xlabel('x1');
>> ylabel('x2');
>> xlabel('lightness');
>> ylabel('width');
```

2-D Gaussian density with independen components



Example: Representation of fish samples using Gaussian Densities

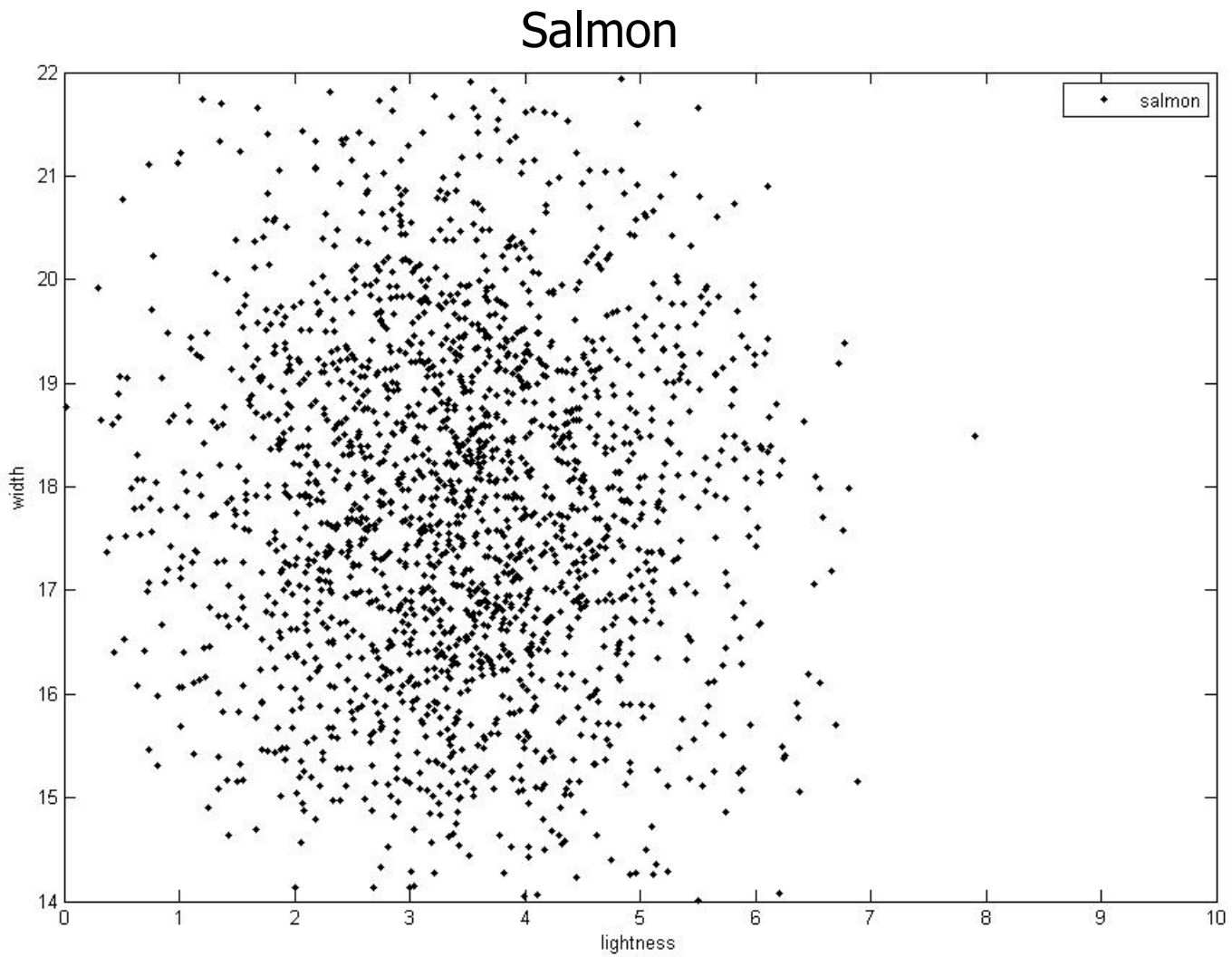
Step 1: Collection of 2D sensor data (lightness, width) for large training set

Step 2: Modeling of the data points in each class with a Gaussian Density

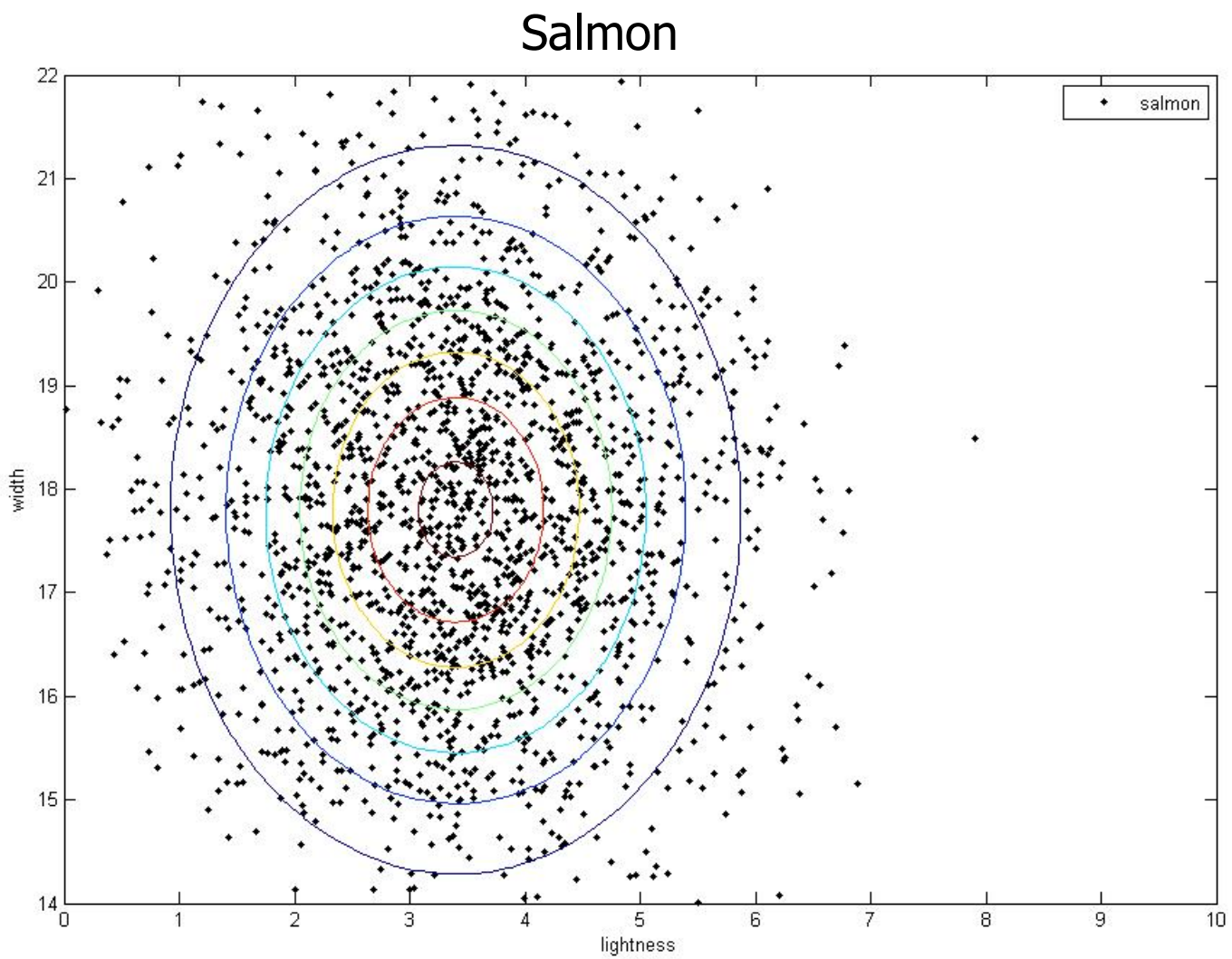
2c. Compute joint probability p(lightness, width) and plot contour lines

```
% contour lines

>> contour(pts1, pts2, p_x1x2_sal);
```

Example: Representation of fish samples using Gaussian Densities

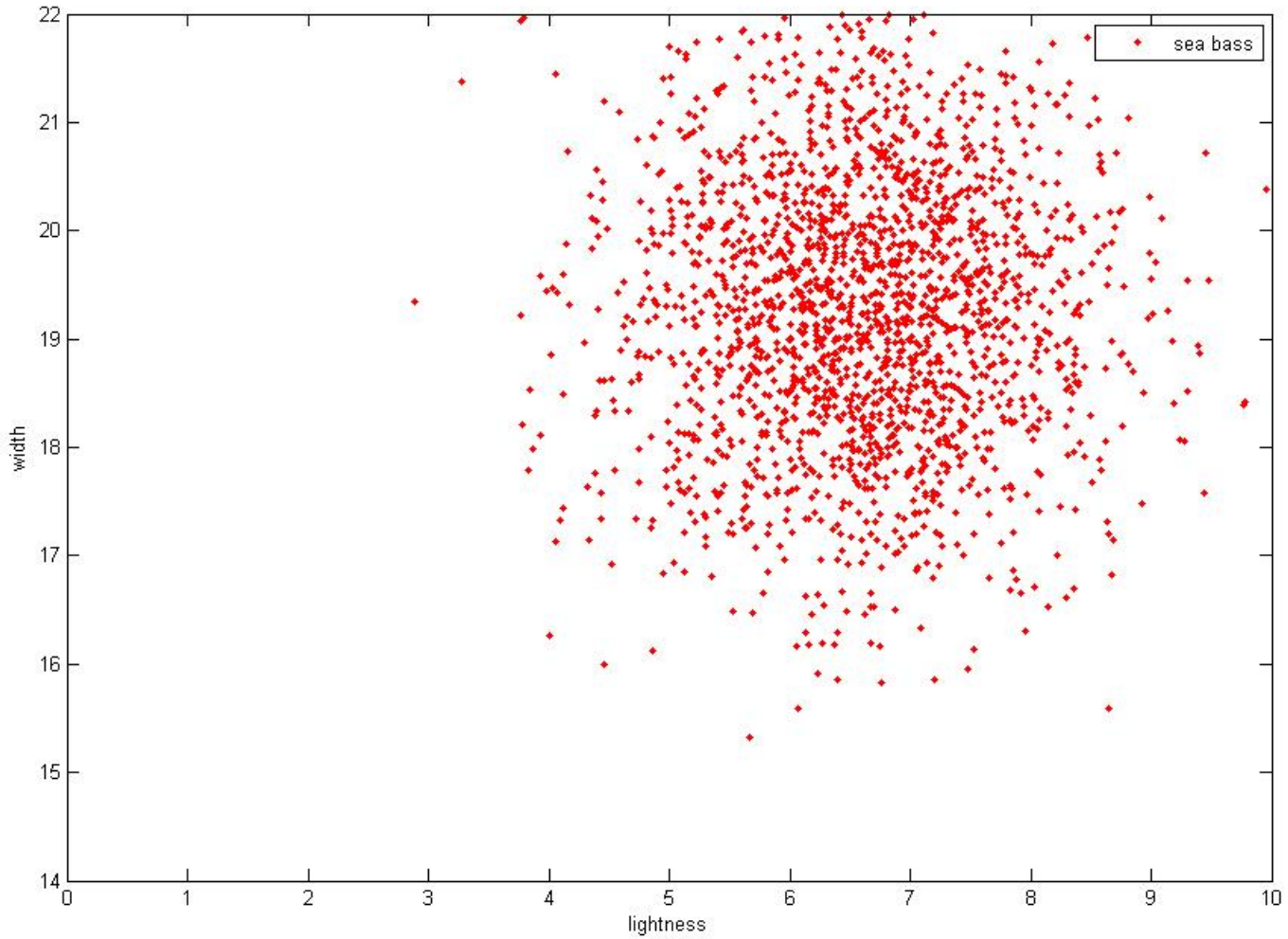


Example: Representation of fish samples using Gaussian Densities

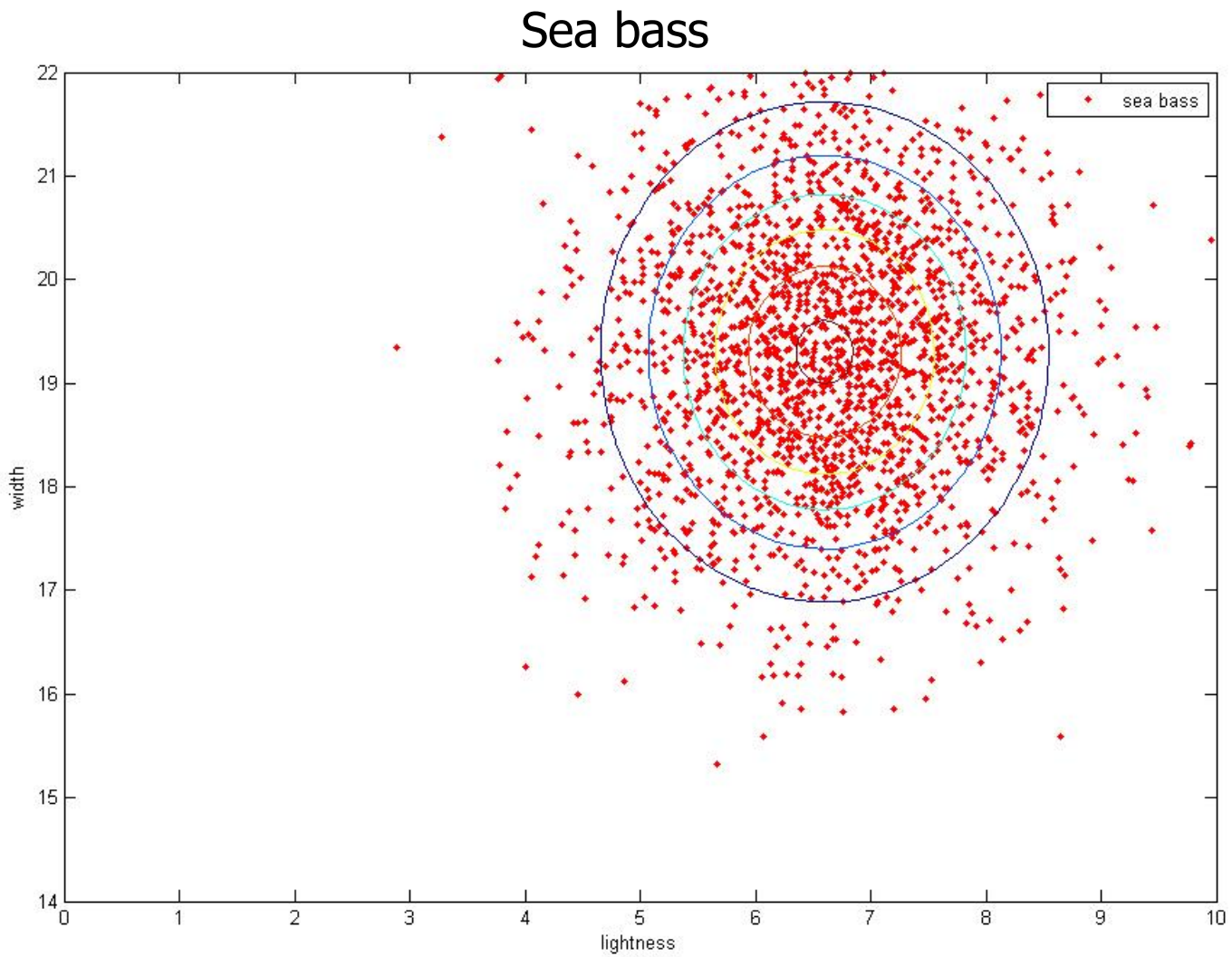


Example: Representation of fish samples using Gaussian Densities

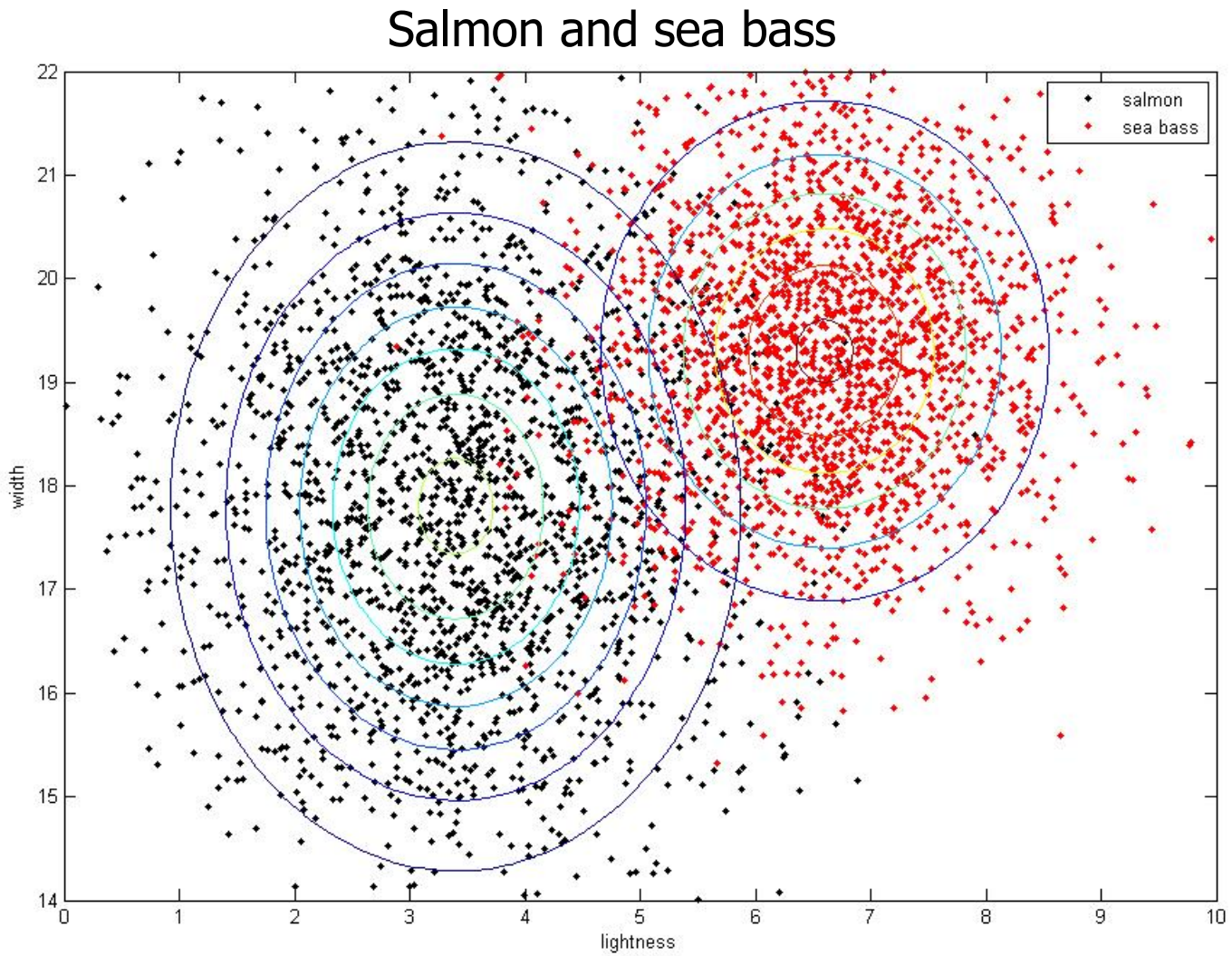
Repeat process for sea bass:



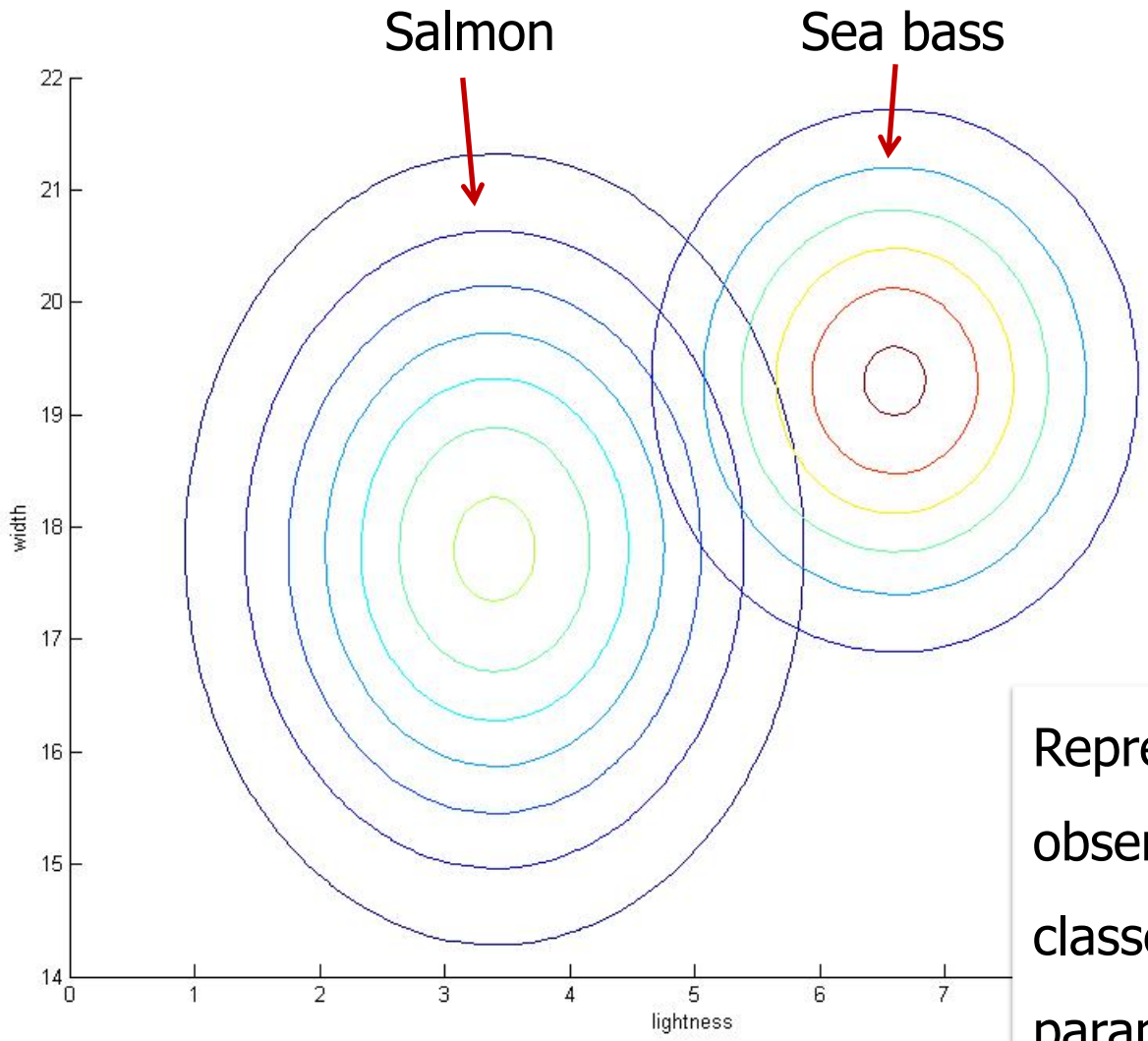
Example: Representation of fish samples using Gaussian Densities



Example: Representation of fish samples using Gaussian Densities



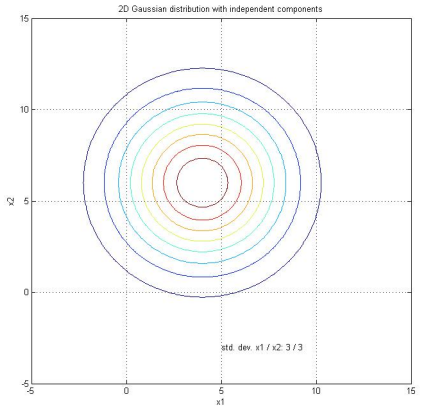
Example: Representation of fish samples using Gaussian Densities

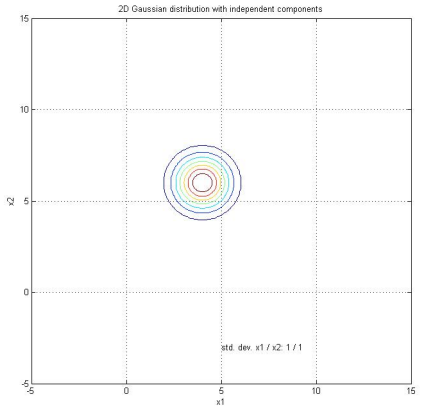


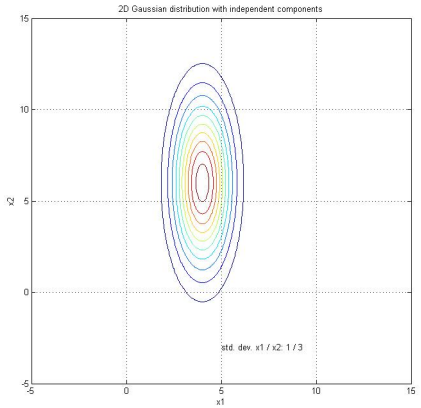
Representation of all observations of both classes with only 8 parameters!

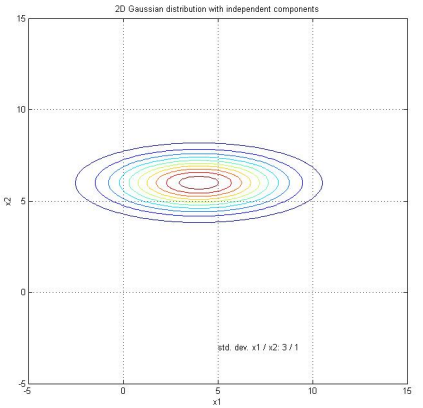
1.13 Multivariate Normal Densities

Multivariate Gaussian densities with **independent components** cover distributions of the following form:



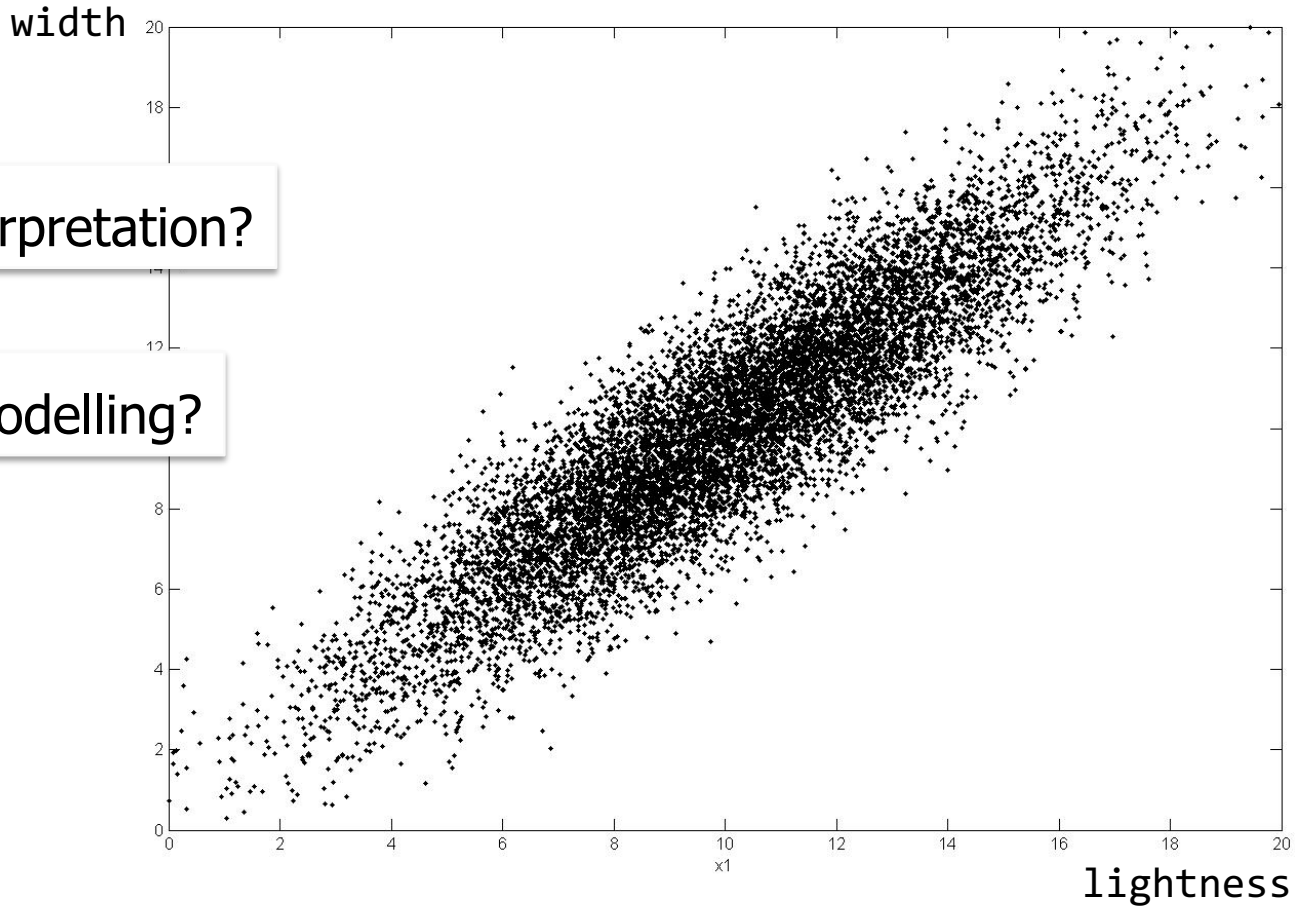






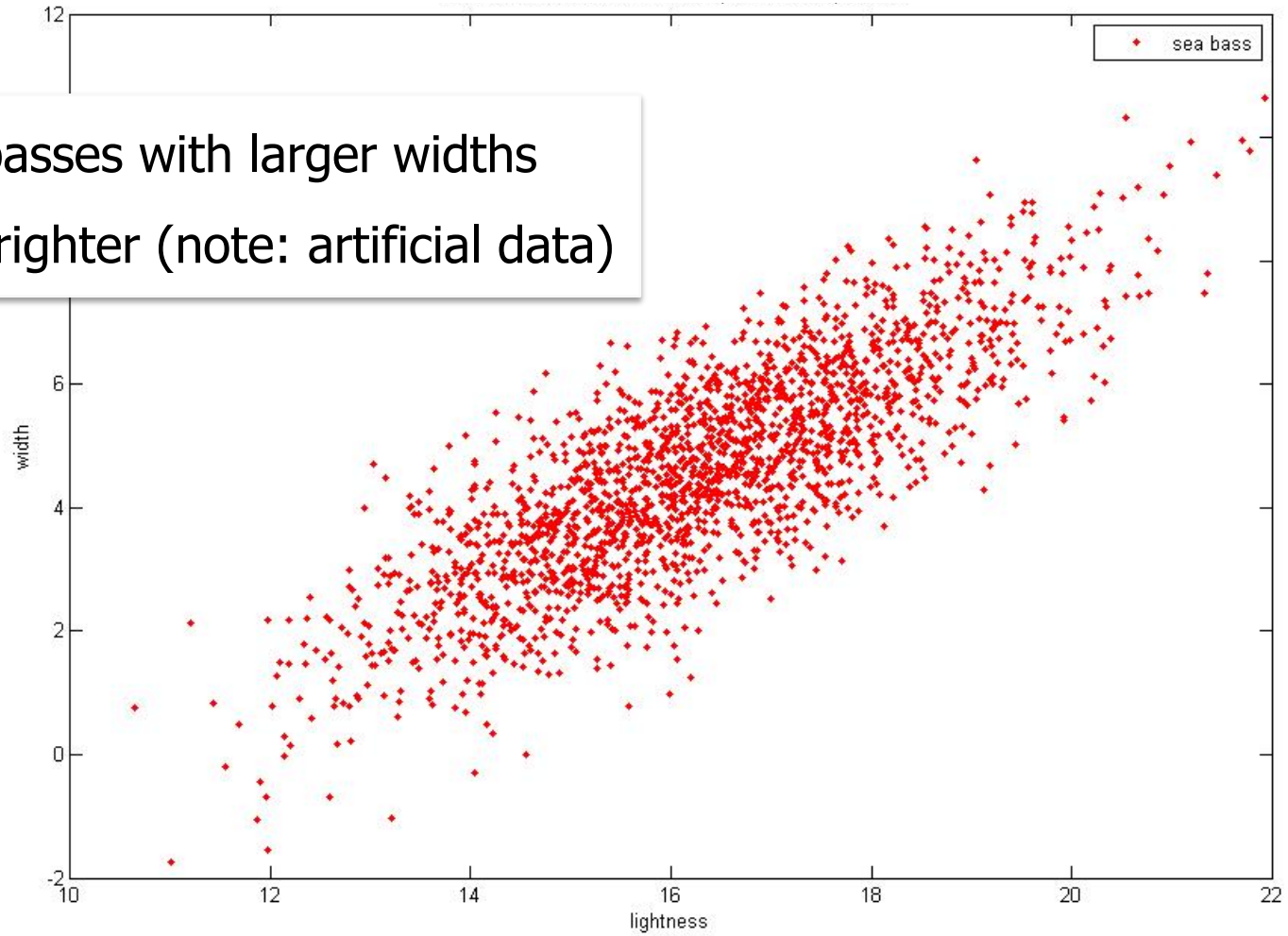
1.13 Multivariate Normal Densities

However, what about distributions like this?



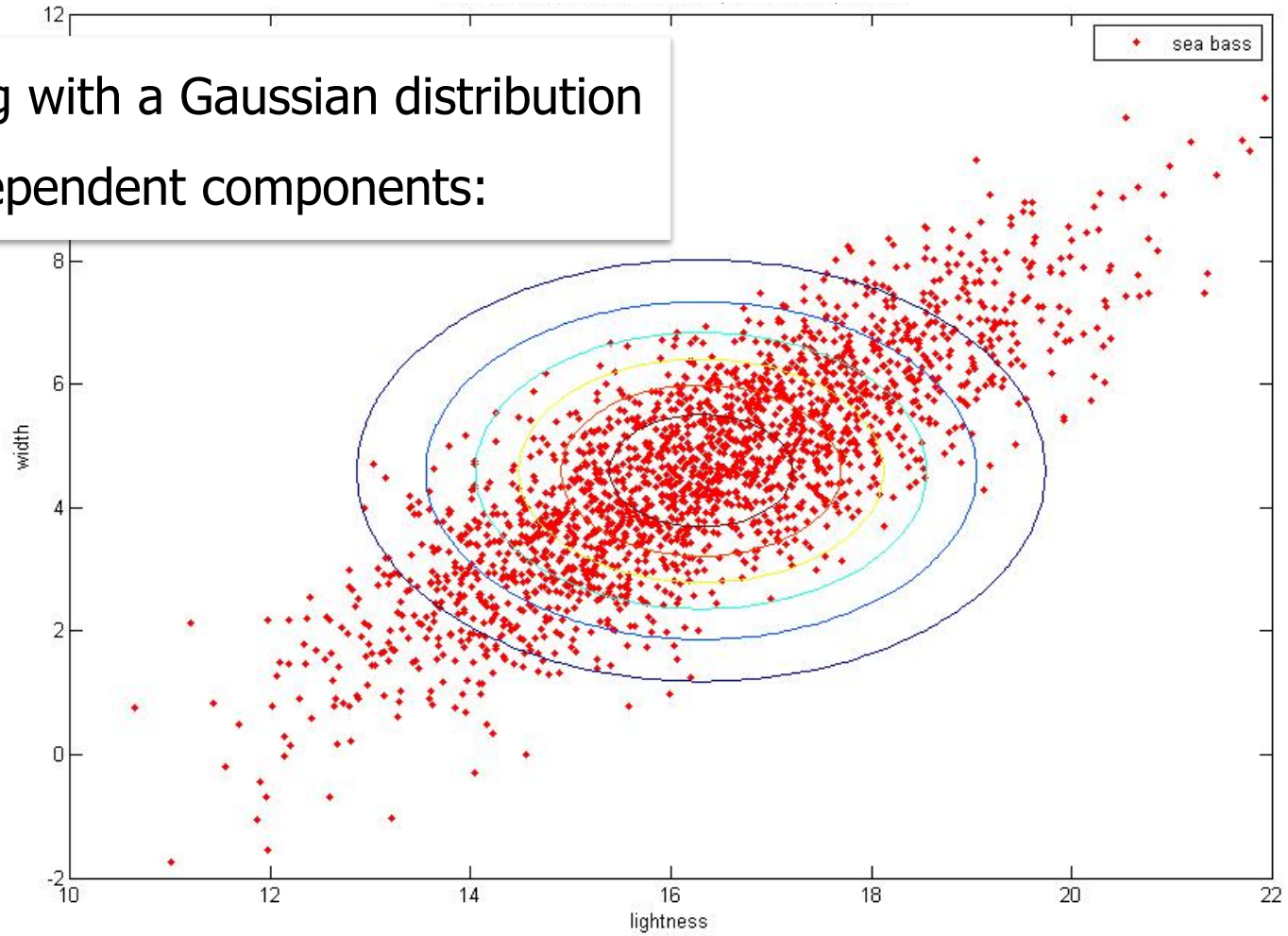
1.13 Multivariate Normal Densities

Sea basses with larger widths are brighter (note: artificial data)



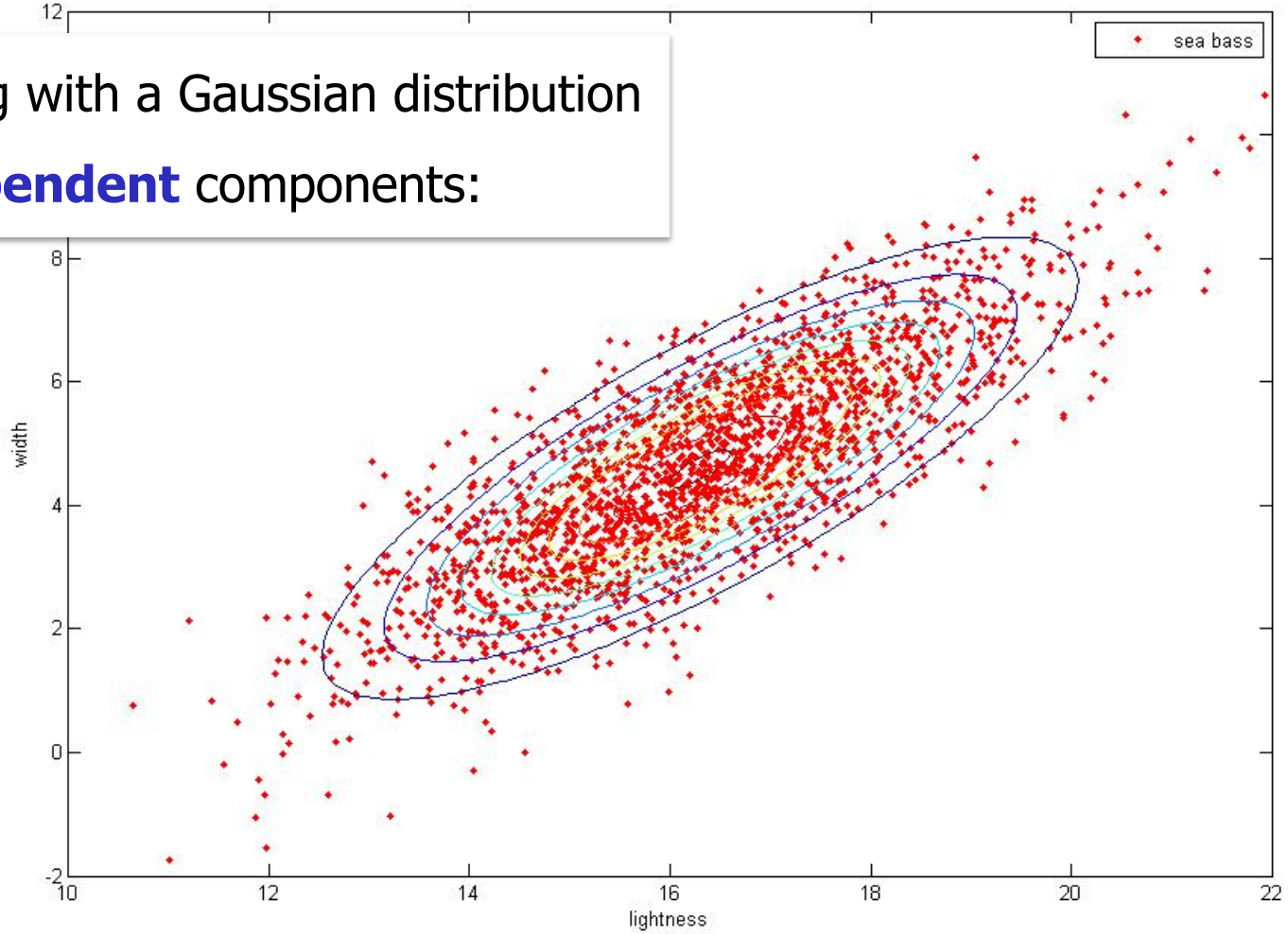
1.13 Multivariate Normal Densities

Modelling with a Gaussian distribution with independent components:



1.13 Multivariate Normal Densities

Modelling with a Gaussian distribution with **dependent** components:



1.13 Multivariate Normal Densities

Case 2: Statistically dependent variables

$$p(x_1, x_2, \ldots, x_D) = \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma)}} \cdot e^{-\frac{1}{2}\cdot(x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu)}$$

mean vector

determinant

covariance matrix

inverse matrix

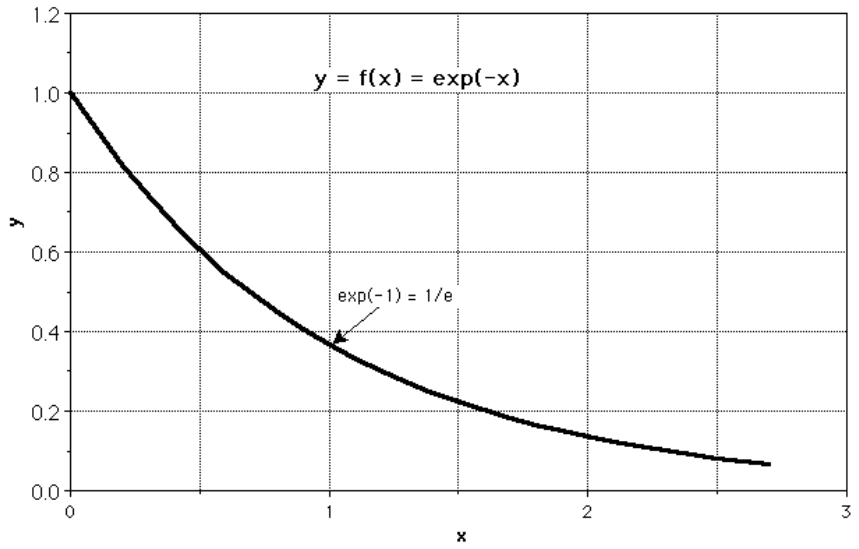
transpose → row vector

Multivariate Gaussian Density With Dependent Components

1.13 Multivariate Normal Densities

Case 2: Statistically dependent variables

$$p(x_1, x_2, \ldots, x_D) = \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma)}} \cdot e^{-\frac{1}{2} \cdot (x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu)}$$

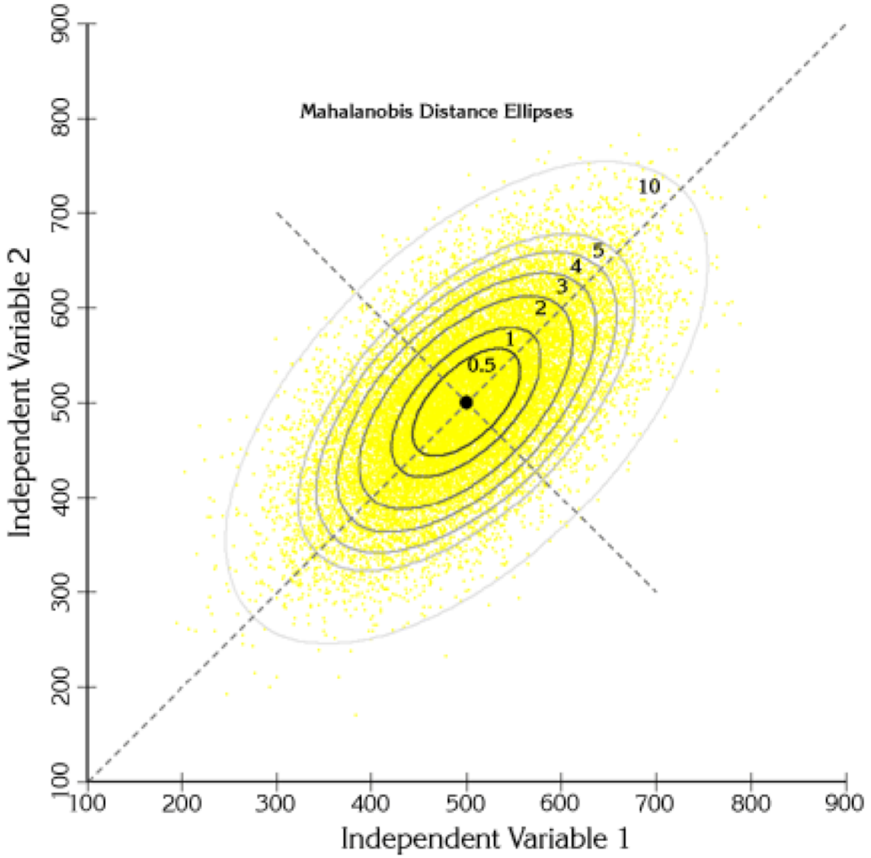


1.13 Multivariate Normal Densities

Mahalanobis Distance

$$d(\mathrm{x}, \mu) = \sqrt{(\mathrm{x}-\mu)^t \cdot \Sigma^{-1} \cdot (\mathrm{x}-\mu)}$$

Mahalanobis Distance between observation vector **x** and distribution with $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$



1.13 Multivariate Normal Densities

Practical realization with Octave:

```
mx=[7 9]';          % mean vector
Cx=[5 3; 3 4];      % covariance matrix
x1=0:0.1:14;
x2=0:0.1:14;
for i=1:length(x1)
    for j=1:length(x2)
        p(j,i)=(1/(2*pi*det(Cx)^1/2))*
        exp((-1/2)*([x1(i) x2(j)]-mx')*inv(Cx)*([x1(i);x2(j)]-mx));
    end
end
figure(1); mesh(x1,x2,p);
xlabel("x1");
ylabel("x2");
figure(2); contour(x1,x2,p);
```

1.13 Multivariate Normal Densities

Practical realization with Octave:

Remark:

D = 2 in our case (2D observation vector)

→ sqrt($2\pi^D$) = 2π

```
for i=1:length(x1)
    for j=1:length(x2)
        p(j,i)=(1/(2*pi*det(Cx)^1/2))*
        exp((-1/2)*([x1(i) x2(j)]-mx')*inv(Cx)*([x1(i);x2(j)]-mx));
    end
end
```

$$p(x_1, x_2, \ldots, x_D) = \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma)}} \cdot e^{-\frac{1}{2} \cdot (x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu)}$$

1.13 Multivariate Normal Densities

Practical realization with Octave:

```
for i=1:length(x1)
    for j=1:length(x2)
        p(j,i)=(1/(2*pi*det(Cx)^1/2))*
        exp((-1/2)*([x1(i) x2(j)]-mx')*inv(Cx)*([x1(i);x2(j)]-mx));
    end
end
```

$$p(x_1, x_2, \ldots, x_D) = \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma)}} \cdot e^{-\frac{1}{2}(x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu)}$$

1.13 Multivariate Normal Densities

Practical realization with Octave:

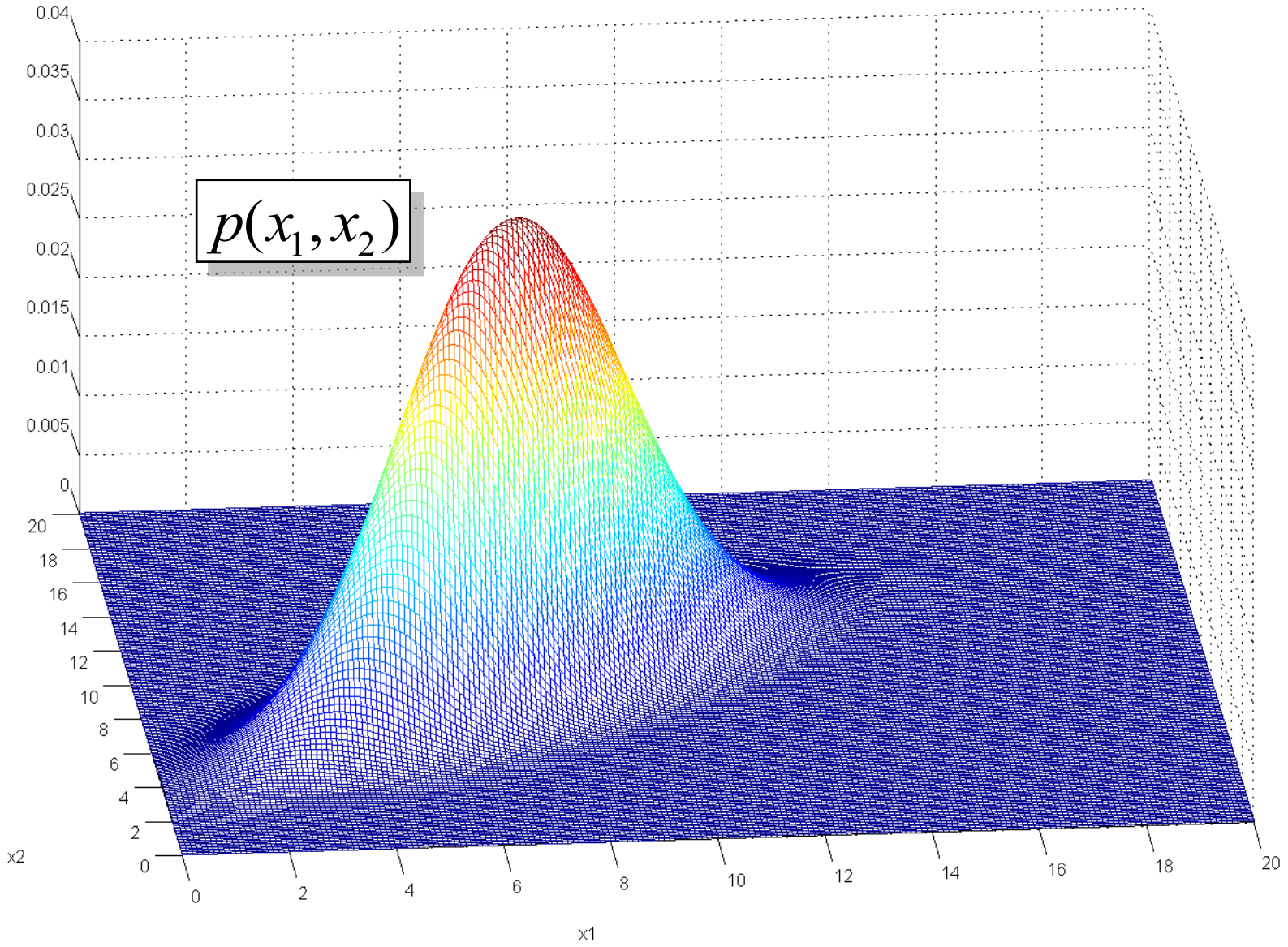
Example:

Covariance matrix:

$$\begin{pmatrix} 5 & 3 \\ 3 & 4 \end{pmatrix}$$

Mean vector:

$$\begin{pmatrix} 7 \\ 9 \end{pmatrix}$$



Multivariate Gaussian Distribution with dependent variables

1.13 Multivariate Normal Densities

$$p(x_1, x_2, \ldots, x_D) = \frac{1}{\sqrt{(2\pi)^D \cdot \det(\Sigma)}} \cdot e^{-\frac{1}{2}\cdot(x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu)}$$

Realization with function `mvnpdf` (statistics package)

```
mx=[1 -1];          % mean vector
Cx=[.9 .4; .4 .3];  % covariance matrix

% generate a grid to consider combinations of x1-x2 values
[X1,X2] = meshgrid(linspace(-1,3,25)', linspace(-3,1,25)');

% mvnpdf computes probability for each point i in X (i.e. X(i,:) )
% therefore: collect all x1-x2-value combinations in X
X = [X1(:) X2(:)];

% result: probability vector, each element contains prob. of one grid point
p = mvnpdf(X, mx, Cx);

% for display we must reshape p into a 25 x 25 plane
surf(X1,X2,reshape(p,25,25));
```

1.13 Multivariate Normal Densities

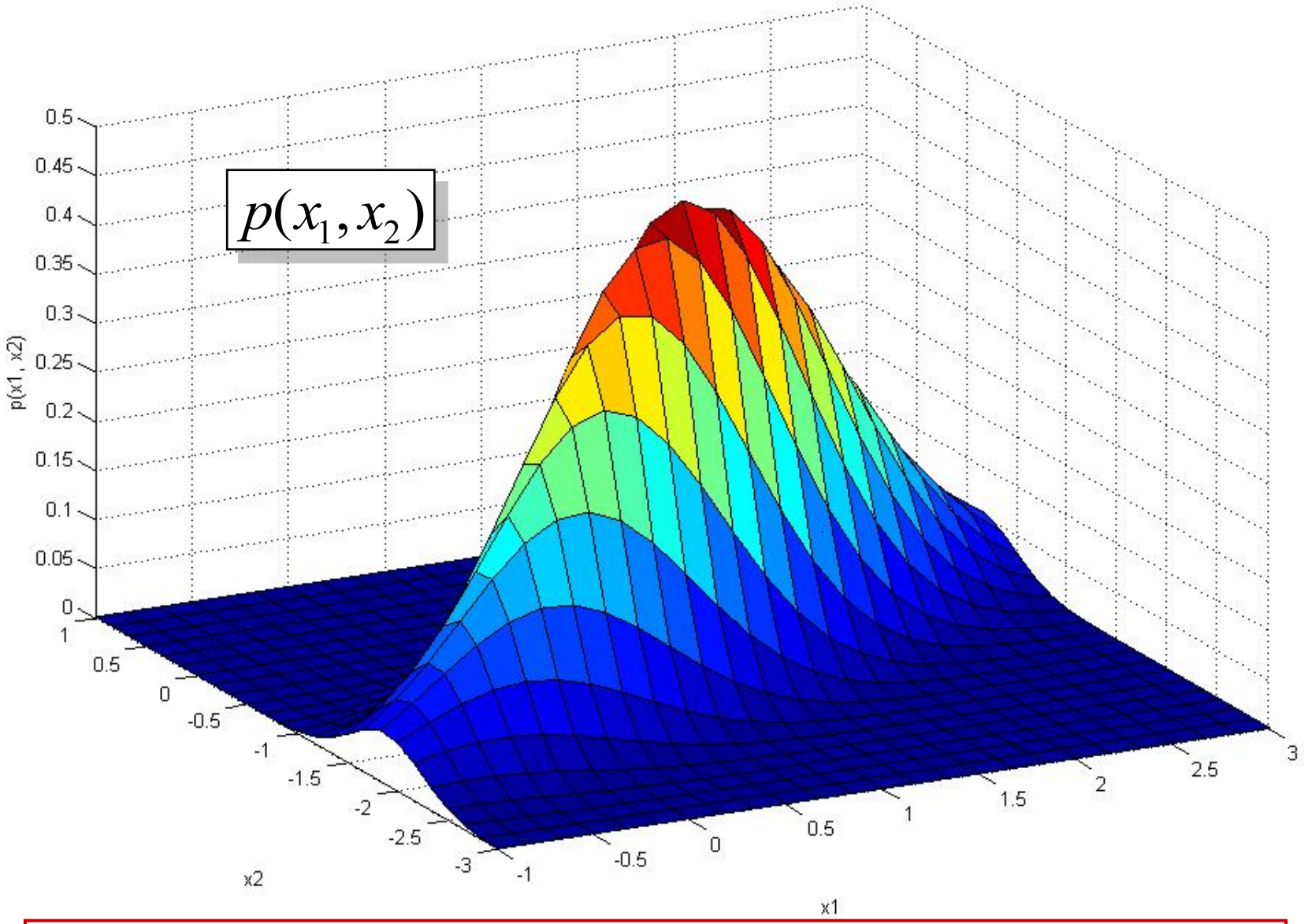
Simplified example for usage of meshgrid

```
[X1, X2] = meshgrid(linspace(-1,3,5)',linspace(-3,1,5)')

   X1 =

     -1     0     1     2     3
     -1     0     1     2     3
     -1     0     1     2     3
     -1     0     1     2     3
     -1     0     1     2     3

   X2 =

     -3    -3    -3    -3    -3
     -2    -2    -2    -2    -2
     -1    -1    -1    -1    -1
      0     0     0     0     0
      1     1     1     1     1
```

1.13 Multivariate Normal Densities

Simplified example for usage of meshgrid

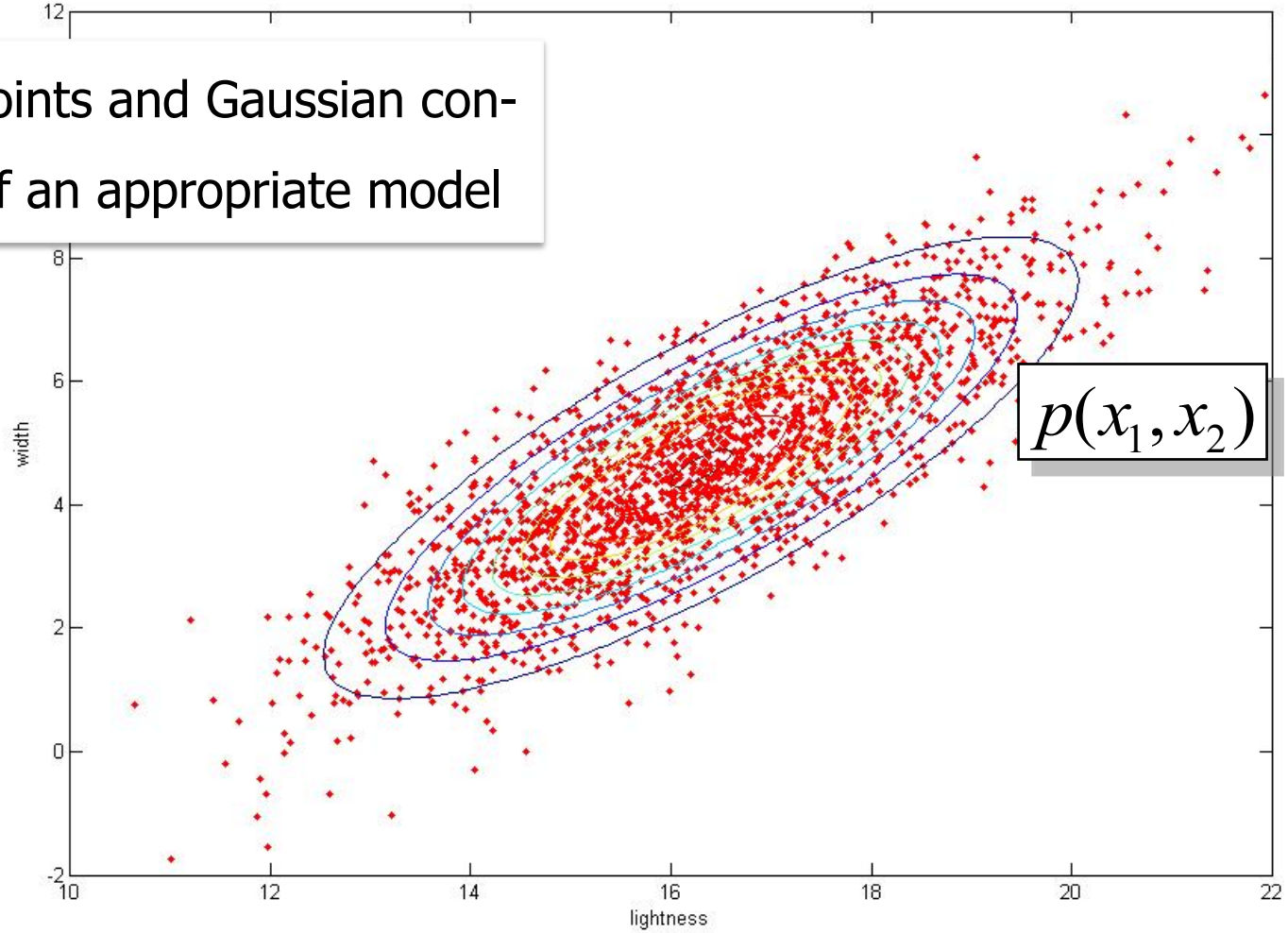
```
[X1, X2] = meshgrid(linspace(-1,3,5)',linspace(-3,1,5)');
X = [X1(:) X2(:)];

X =

    -1    -3
    -1    -2
    -1    -1
    -1     0
    -1     1
     0    -3
     0    -2
     0    -1
     0     0
     ...
```



Multivariate Gaussian Distribution with dependent variables

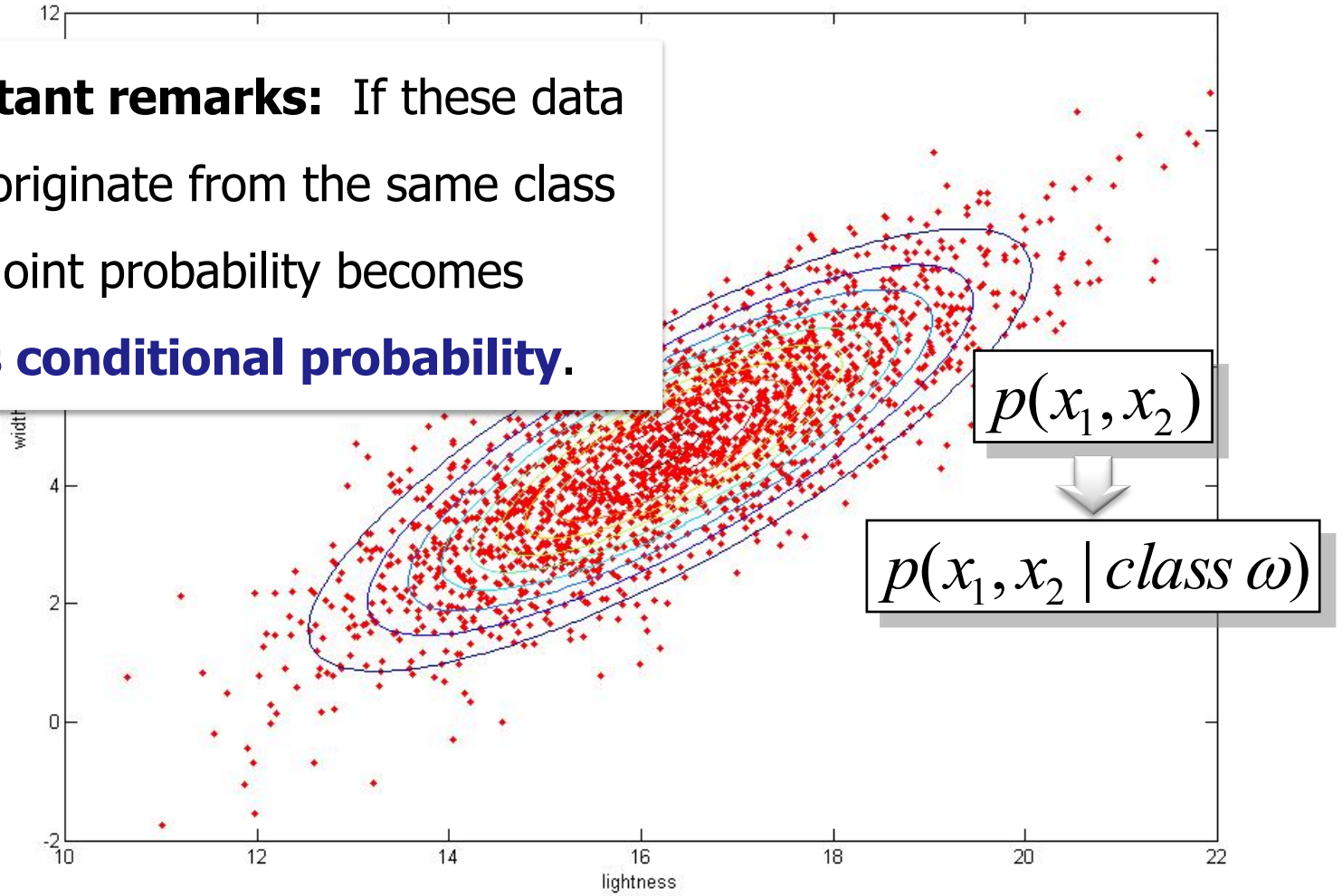
1.13 Multivariate Normal Densities

Data points and Gaussian contours of an appropriate model



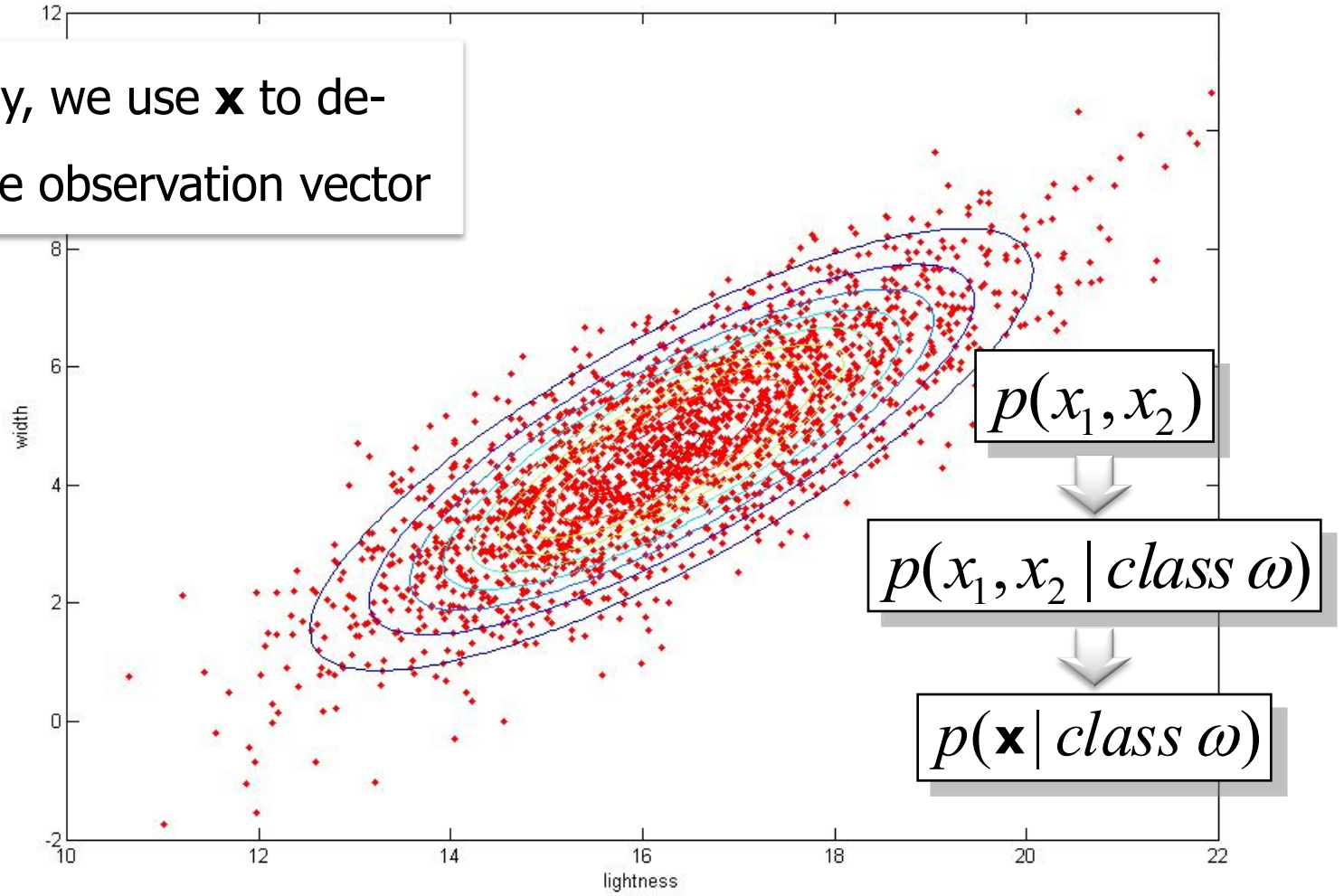
1.13 Multivariate Normal Densities

Important remarks: If these data points originate from the same class ω, the joint probability becomes a **class conditional probability**.



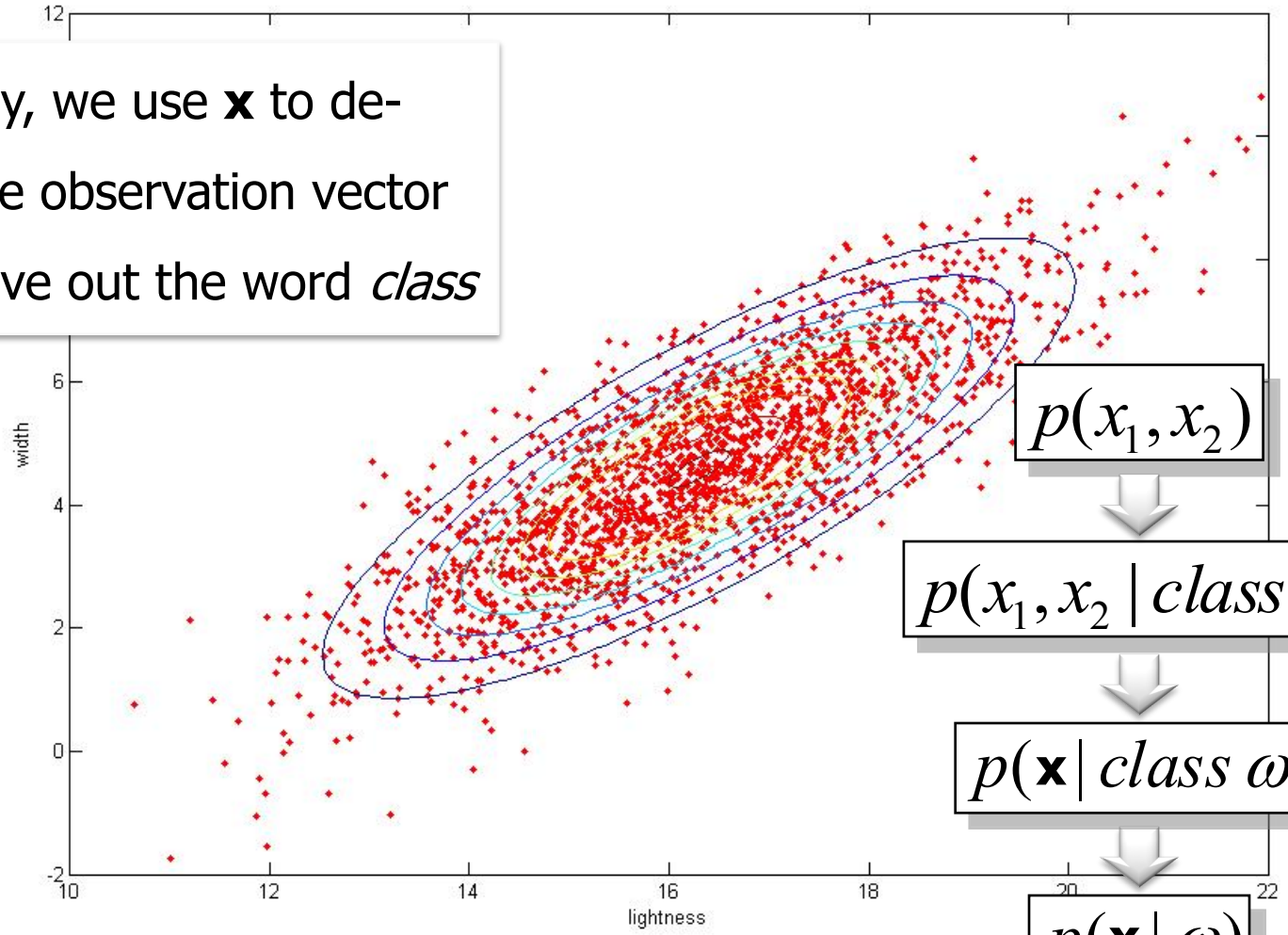
1.13 Multivariate Normal Densities

Typically, we use **x** to denote the observation vector



1.13 Multivariate Normal Densities

Typically, we use **x** to denote the observation vector and leave out the word *class*



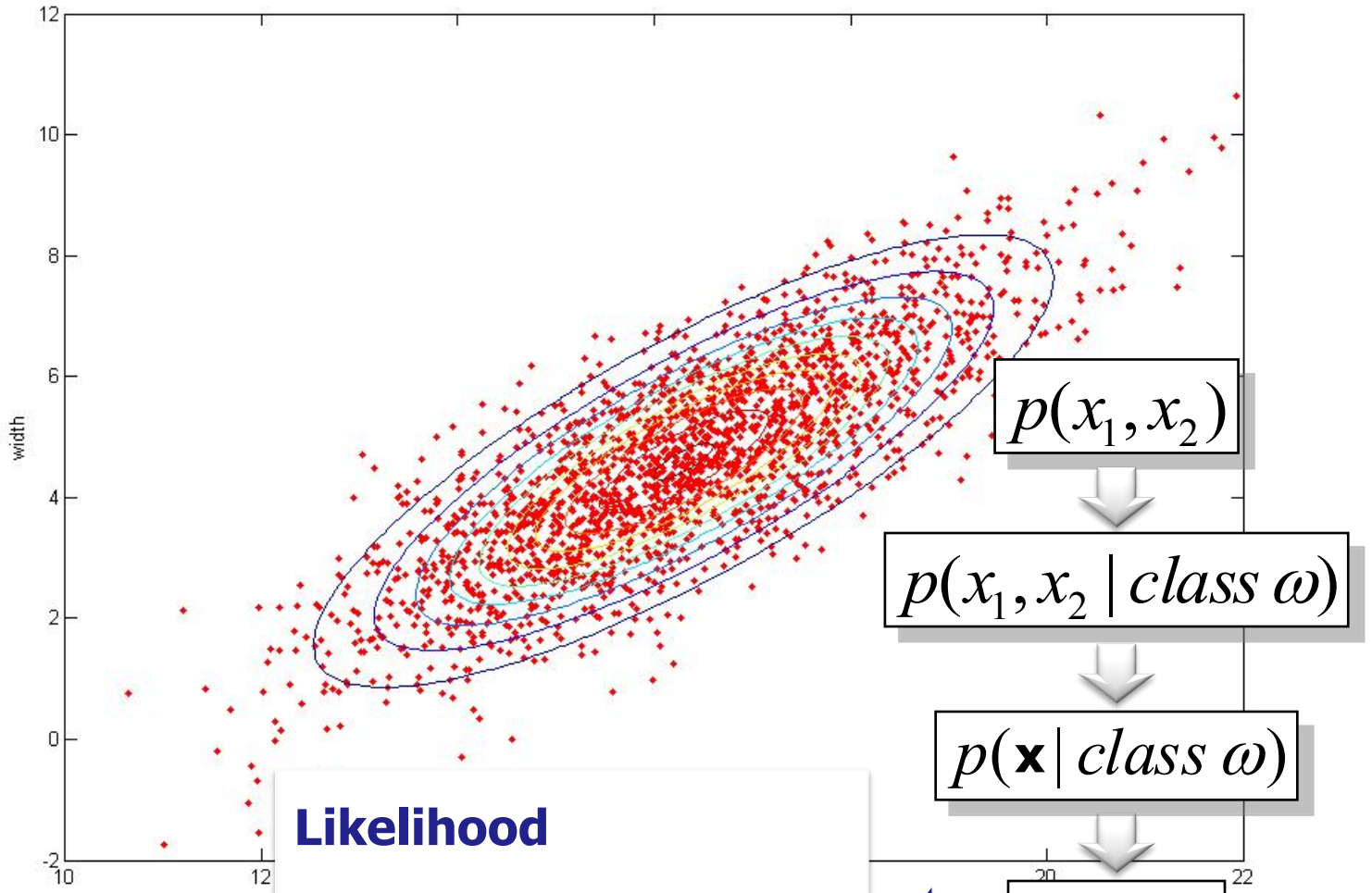
$$p(x_1, x_2)$$

$$p(x_1, x_2 \mid class\ \omega)$$

$$p(\mathbf{x} \mid class\ \omega)$$

$$p(\mathbf{x} \mid \omega)$$

1.13 Multivariate Normal Densities



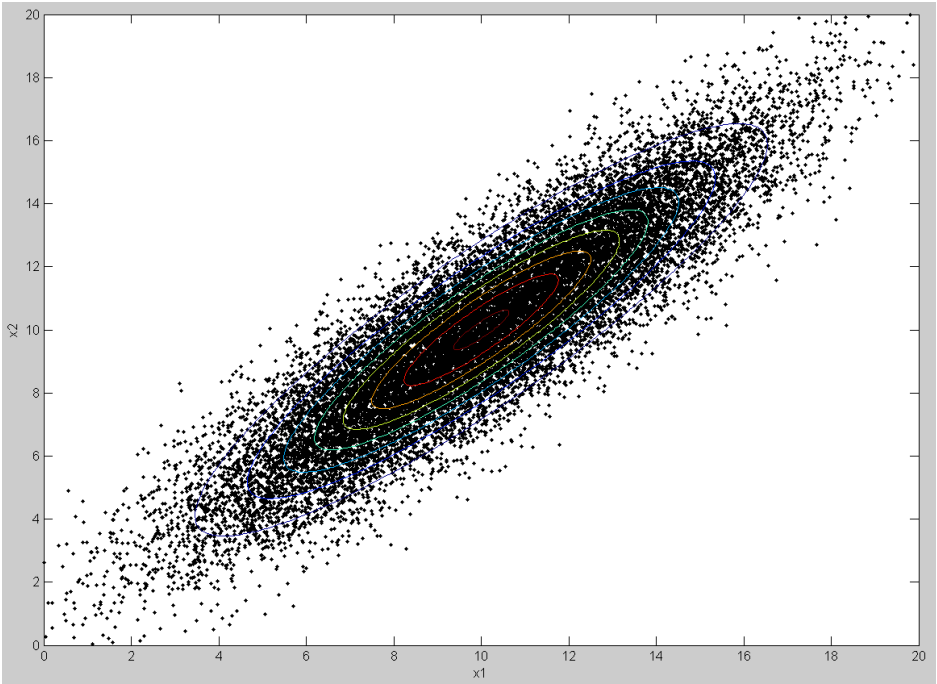
1.14 Mixture Densities

Explanation: next slides

Up to now:

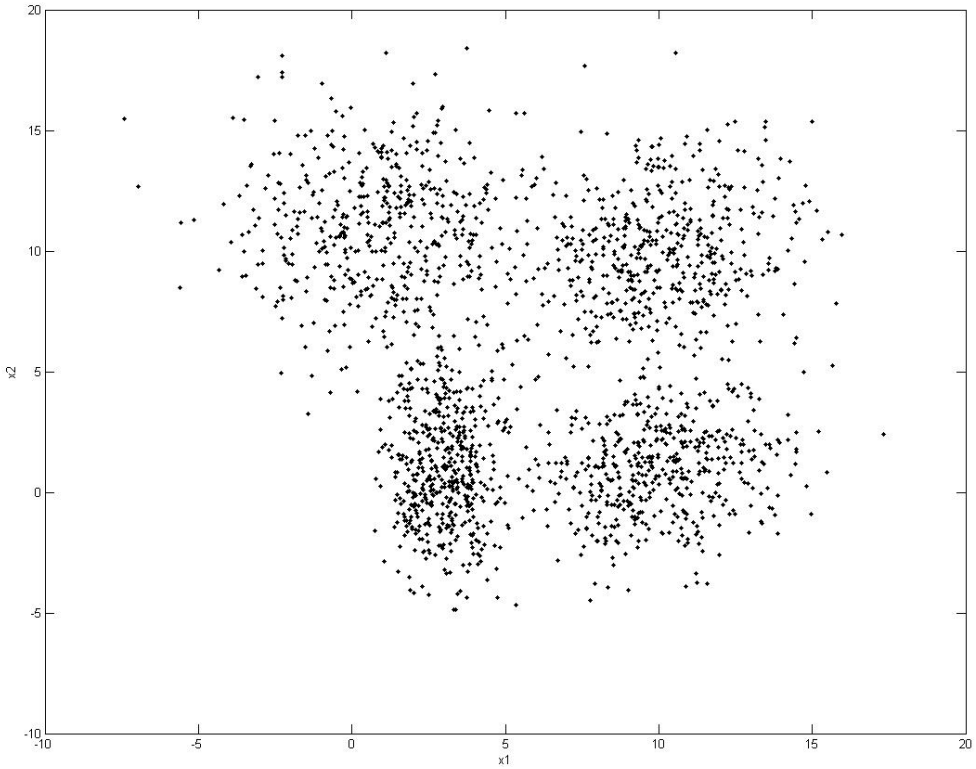
Arbitrary distributions which could be modeled by a unimodal but multivariate normal distribution

depending on a vector of random variables



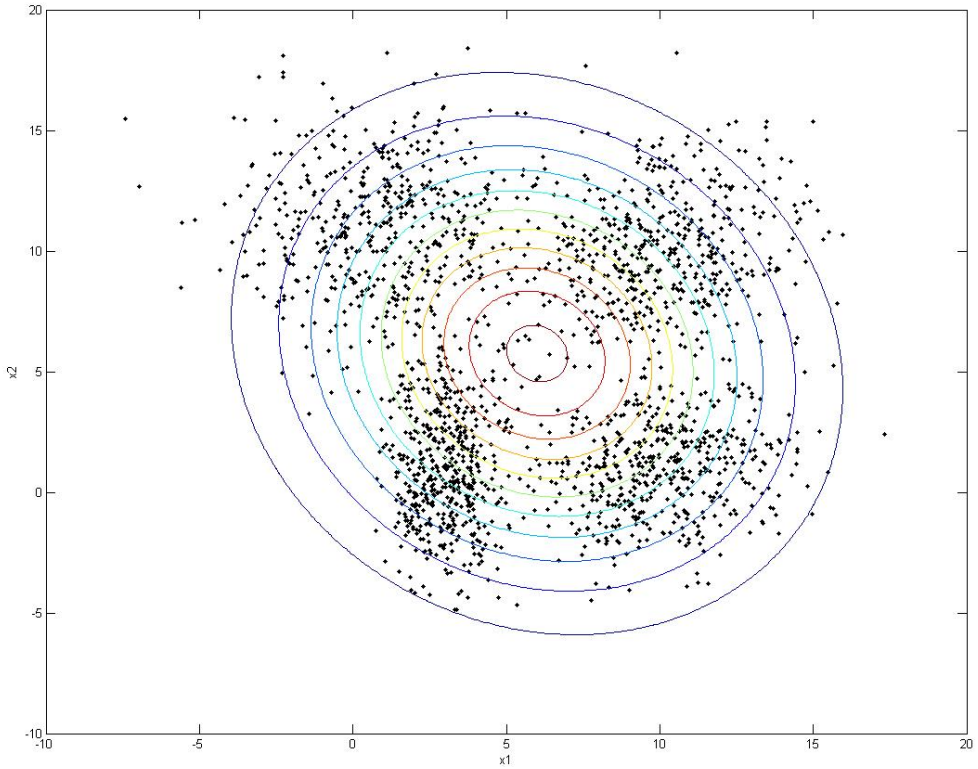
1.14 Mixture Densities

What can we do with the following type of distribution?



1.14 Mixture Densities

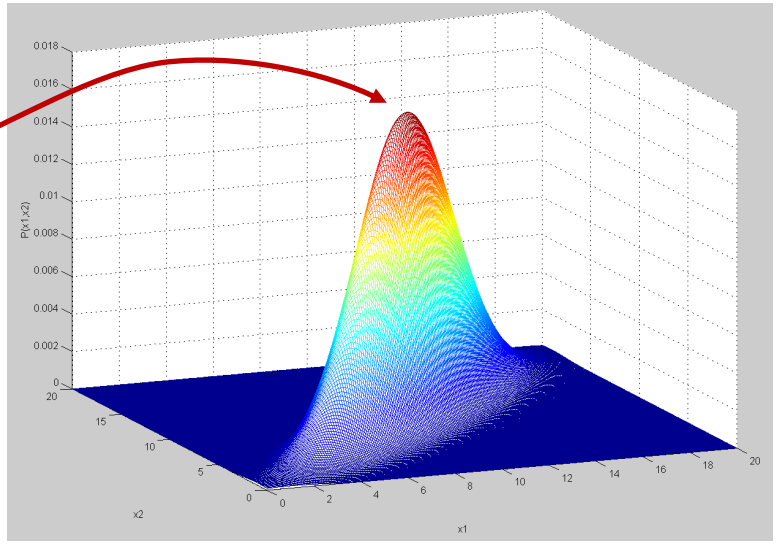
What can we do with the following type of distribution?



Definition

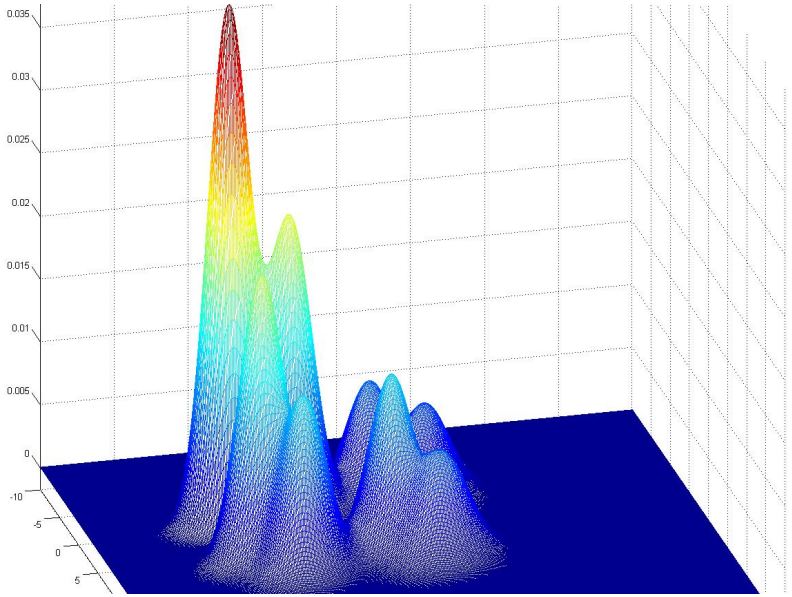
Unimodal distributions:

→ exhibit only a single local maximum



Multimodal distributions or **Mixture Densities**

→ exhibit several local maxima



1.14 Mixture Densities

Multimodal distributions can be generated by a combination of distributions.

Standard model:

Number of modes

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} p(\boldsymbol{x}, i)$$

Law of total probability

1.14 Mixture Densities

Multimodal distributions can be generated by a combination of distributions.

Standard model:

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} p(\boldsymbol{x}, i) = \sum_{i=1}^{I} p(i) \cdot p(\boldsymbol{x} \mid i)$$

Prior probability of mode i

Conditional probability of observation x given mode i

1.14 Mixture Densities

Multimodal distributions can be generated by a combination of distributions.

Standard model:

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} p(\boldsymbol{x}, i) = \sum_{i=1}^{I} p(i) \cdot p(\boldsymbol{x} \mid i) = \sum_{i=1}^{I} p(i) \cdot N(\boldsymbol{x} \mid \mu_i, \Sigma_i)$$

Modeling of conditional probability by unimodal normal distribution N( ).

1.14 Mixture Densities

Multimodal distributions can be generated by a combination of distributions.

Standard model:

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} p(c_i) \cdot N(\boldsymbol{x} \mid \mu_i, \Sigma_i)$$

Unimodal normal distribution

Interpretation:

A Mixture Density is a combination of weighted unimodal normal distributions.

1.14 Mixture Densities

Multimodal distributions can be generated by a combination of distributions.

Standard model:

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} p(c_i) \cdot N(\boldsymbol{x} \mid \mu_i, \Sigma_i)$$

Remark: mean vector and covariance matrix can be individually adjusted for each of the I modes.

Interpretation:

A Mixture Density is a combination of weighted unimodal normal distributions.

1.14 Mixture Densities

Multimodal distributions can be generated by a combination of distributions.

Standard model:

$$p(\boldsymbol{x}) = \sum_{i=1}^{I} p(c_i) \cdot N(\boldsymbol{x} \mid \mu_i, \Sigma_i)$$

Class-conditional probability (likelihood):

$$p(\boldsymbol{x} \mid k) = \sum_{i=1}^{I} p(\boldsymbol{x}, i \mid k) = \sum_{i=1}^{I} p(i \mid k) \cdot p(\boldsymbol{x} \mid i, k)$$

$$p(\boldsymbol{x} \mid k) = \sum_{i=1}^{I} p(i \mid k) \cdot N(\boldsymbol{x} \mid \mu_{ik}, \Sigma_{ik})$$

1.14 Mixture Densities

Class-conditional distribution:

$$p(\boldsymbol{x} \mid k) = \sum_{i=1}^{I} p(i \mid k) \cdot N(\boldsymbol{x} \mid \mu_{ik}, \Sigma_{ik})$$

class dependency

1.14 Mixture Densities

In praxis usually sufficient: **Maximum Approximation**

$$p(\boldsymbol{x} \mid k) = \sum_{i=1}^{I} p(i \mid k) \cdot N(\boldsymbol{x} \mid \mu_{ik}, \Sigma_{ik})$$

$$p(\boldsymbol{x} \mid k) \approx \max_{i} \left\{ p(i \mid k) \cdot N(\boldsymbol{x} \mid \mu_{ik}, \Sigma_{ik}) \right\}$$

Remarks:

- Sum is replaced by maximum operation

  → distribution with largest weighted probab. represents conditional prob.

- Motivation: exponential drop of normal distribution

  → sum is dominated by largest summand

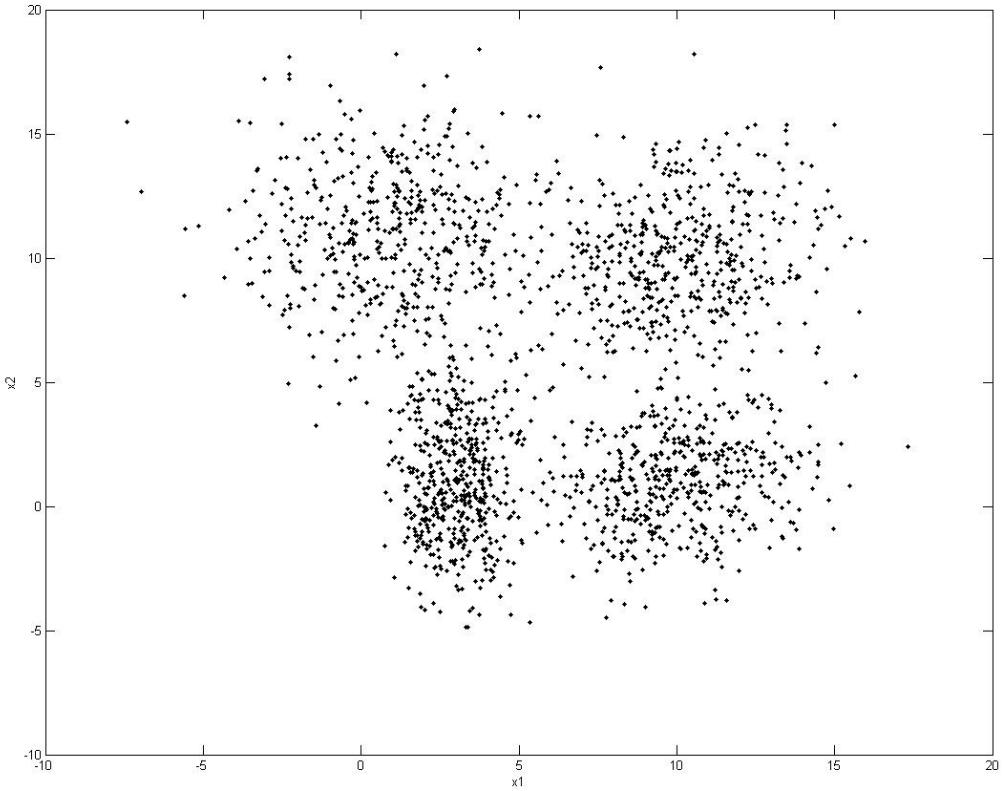
1.14 Mixture Densities

$$p(\boldsymbol{x} \mid k) = \sum_{i=1}^{I} p(i \mid k) \cdot N(\boldsymbol{x} \mid \mu_{ik}, \Sigma_{ik})$$

$$p(\boldsymbol{x} \mid k) \approx \max_{i} \{\, p(i \mid k) \cdot N(\boldsymbol{x} \mid \mu_{ik}, \Sigma_{ik}) \,\}$$

Question: How to estimate

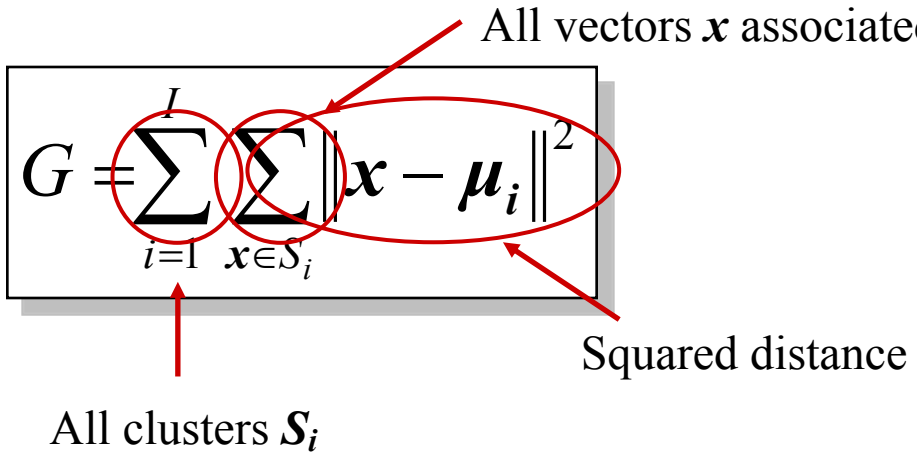
- number of modes I,
- priors $p(i \mid k)$,
- mean vectors,
- covariance matrices?



K-Means Clustering

Goal:

Split dataset into I partitions to minimize square error criterion



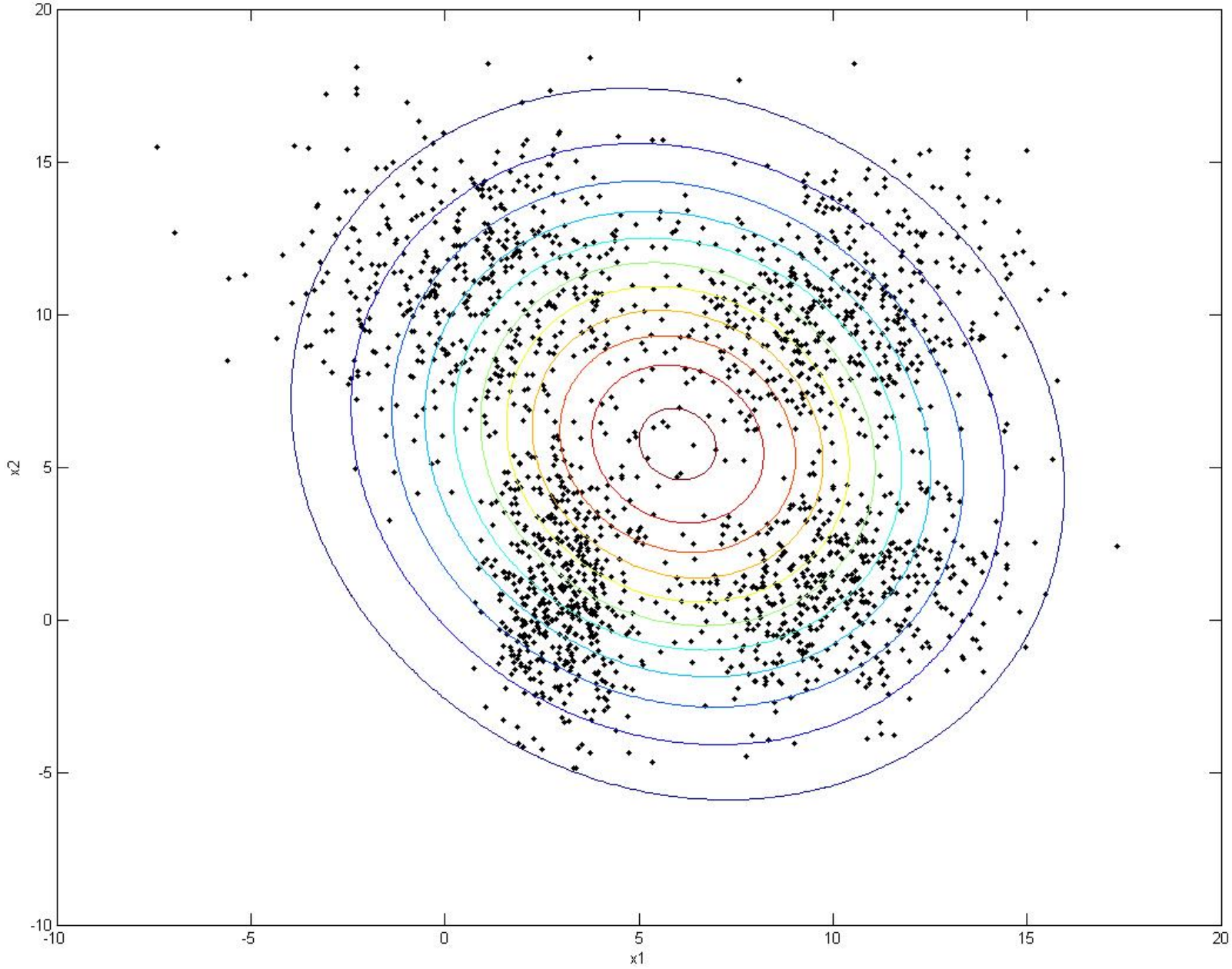
K-Means Clustering

Iterative Technique:

Given: Training observations $\boldsymbol{x_1}, \ldots, \boldsymbol{x_j}, \ldots, \boldsymbol{x_n}$ (with $\boldsymbol{x}_j = \begin{bmatrix} x_{j1} \\ \ldots \\ x_{jD} \end{bmatrix}$)

1. Determine parameters of unimodal multivariate distribution

   Compute empirical mean vector and covariance matrix from all training observations

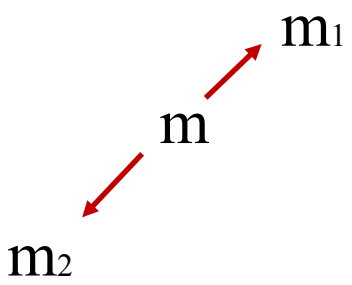

20
15
10
5
0
-5
-10
x2
-10
-5
0
5
10
15
20
x1

K-Means Clustering

2. Split distribution

   a. Split mean

```
eps = 0.001;
m1 = m + eps;
m2 = m – eps;
```



   b. Assign each training sample x_j to optimal cluster

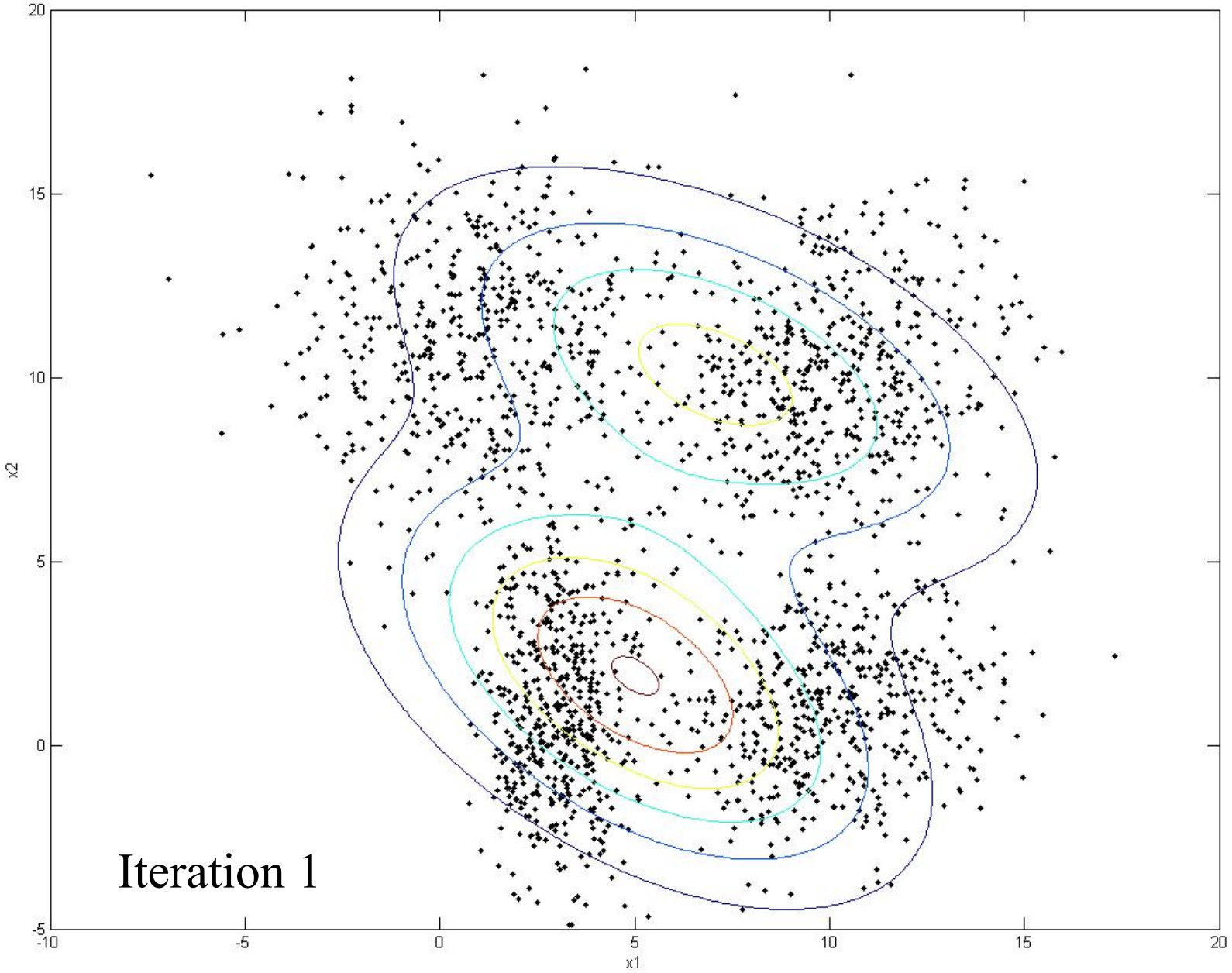
   optimal → minimum increase of cluster variance

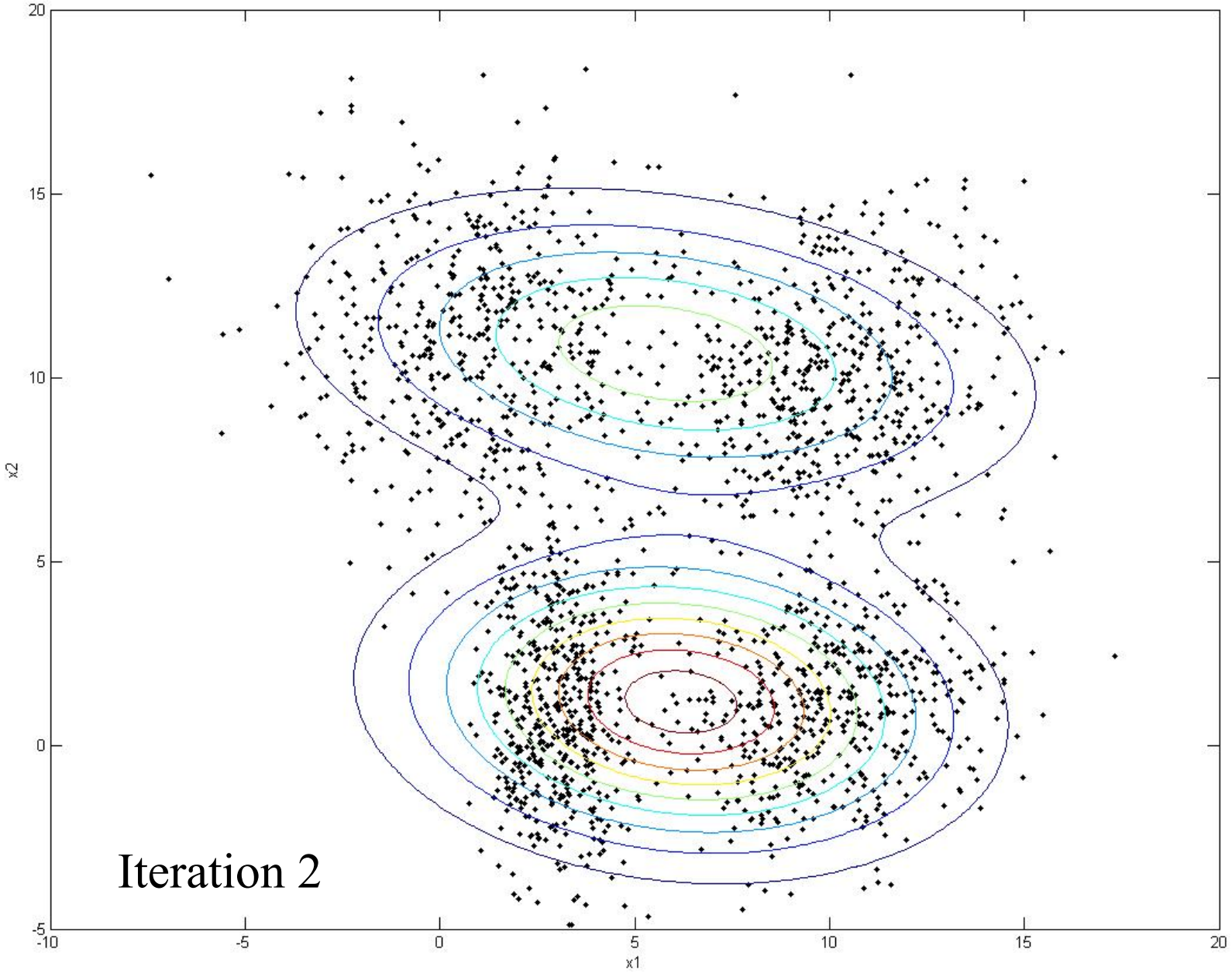
$$S_k^{(t)} = \left\{ \boldsymbol{x}_j : \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_k^{(t)} \right\|^2 \leq \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i^{(t)} \right\|^2 \text{ for all clusters } i = 1,\ldots,I, i \neq k \right\}$$

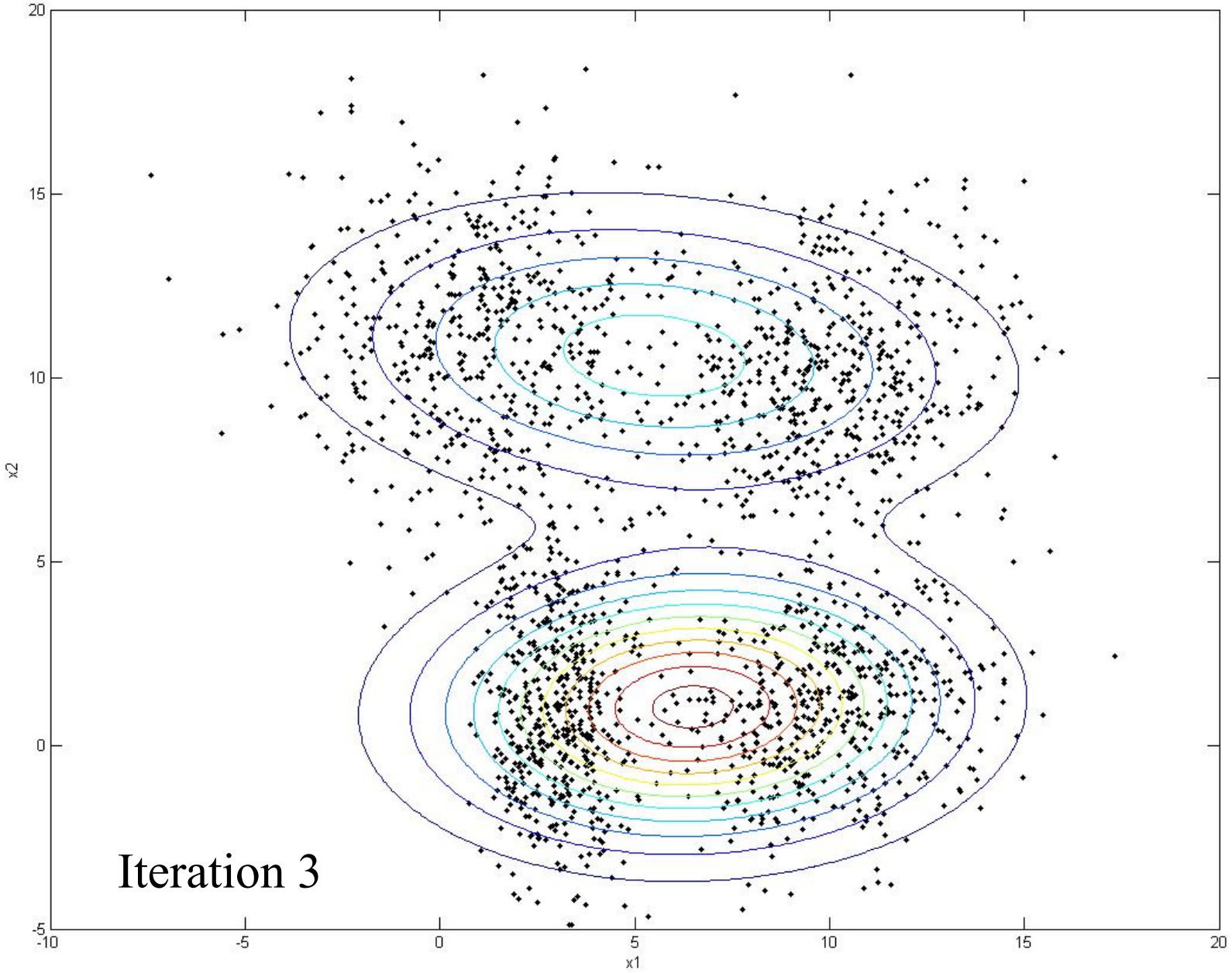
iterate

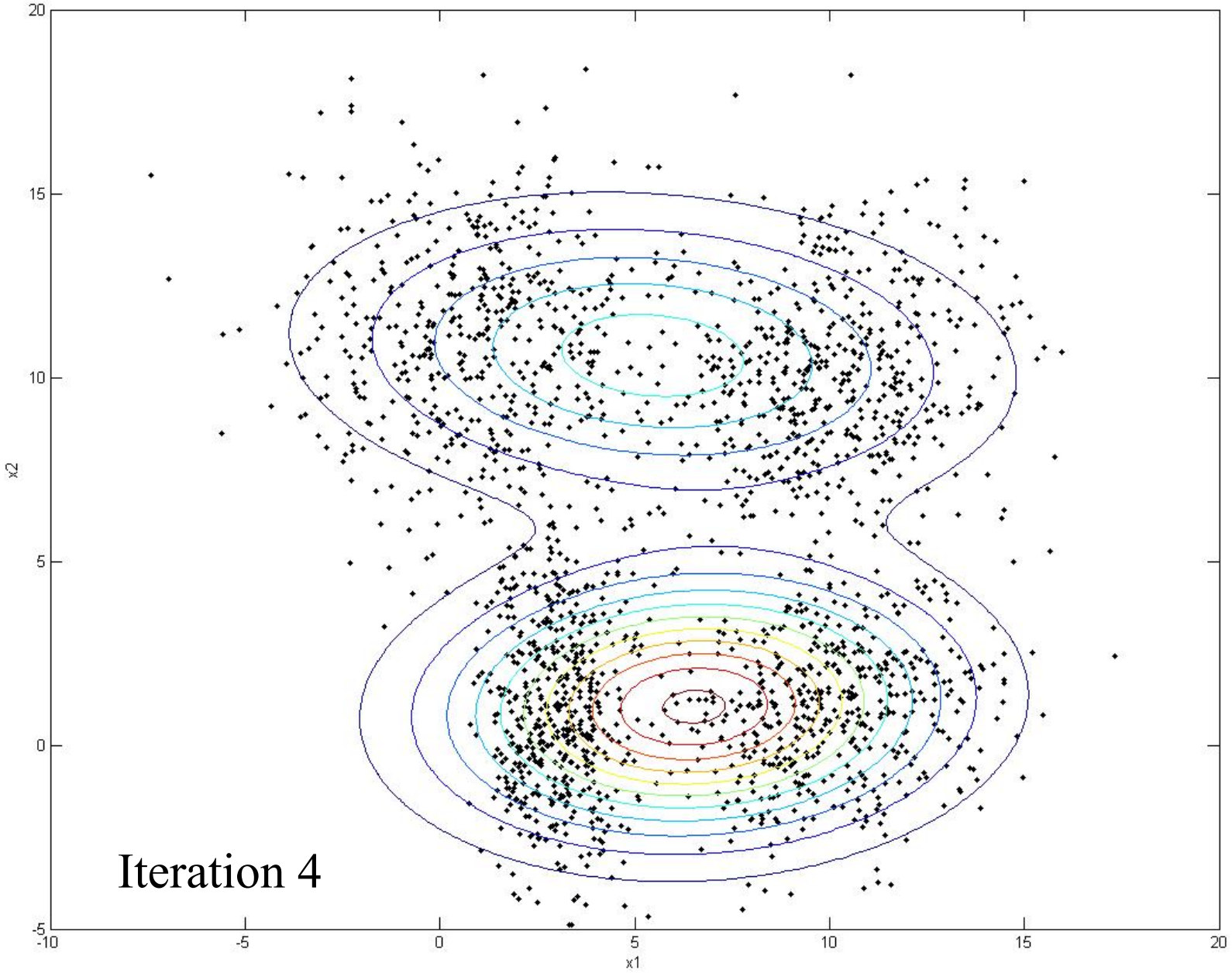
   c. Determine the new means for all clusters with new associated x_j

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{\left| S_k^{(t)} \right|} \sum_{\boldsymbol{x}_j \in S_k^{(t)}} \boldsymbol{x}_j$$


20
15
10
5
0
-5
-10
-5
0
5
10
15
20
x1
x2
Iteration 1


20
15
10
5
0
-5
x2
-10
-5
0
5
10
15
20
x1
Iteration 2


20
15
10
5
0
-5
-10
-5
0
5
10
15
20
x1
x2
Iteration 3


20
15
10
5
0
-5
-10
-5
0
5
10
15
20
x1
x2
Iteration 4

K-Means Clustering

2. Split distribution

   a. Split means

```
eps = 0.001;
m11 = m1 + eps;   m12 = m1 - eps;
m21 = m2 + eps;   m22 = m2 - eps;
```

   b. Assign each training sample x_j to optimal cluster

   optimal → minimum increase of cluster variance

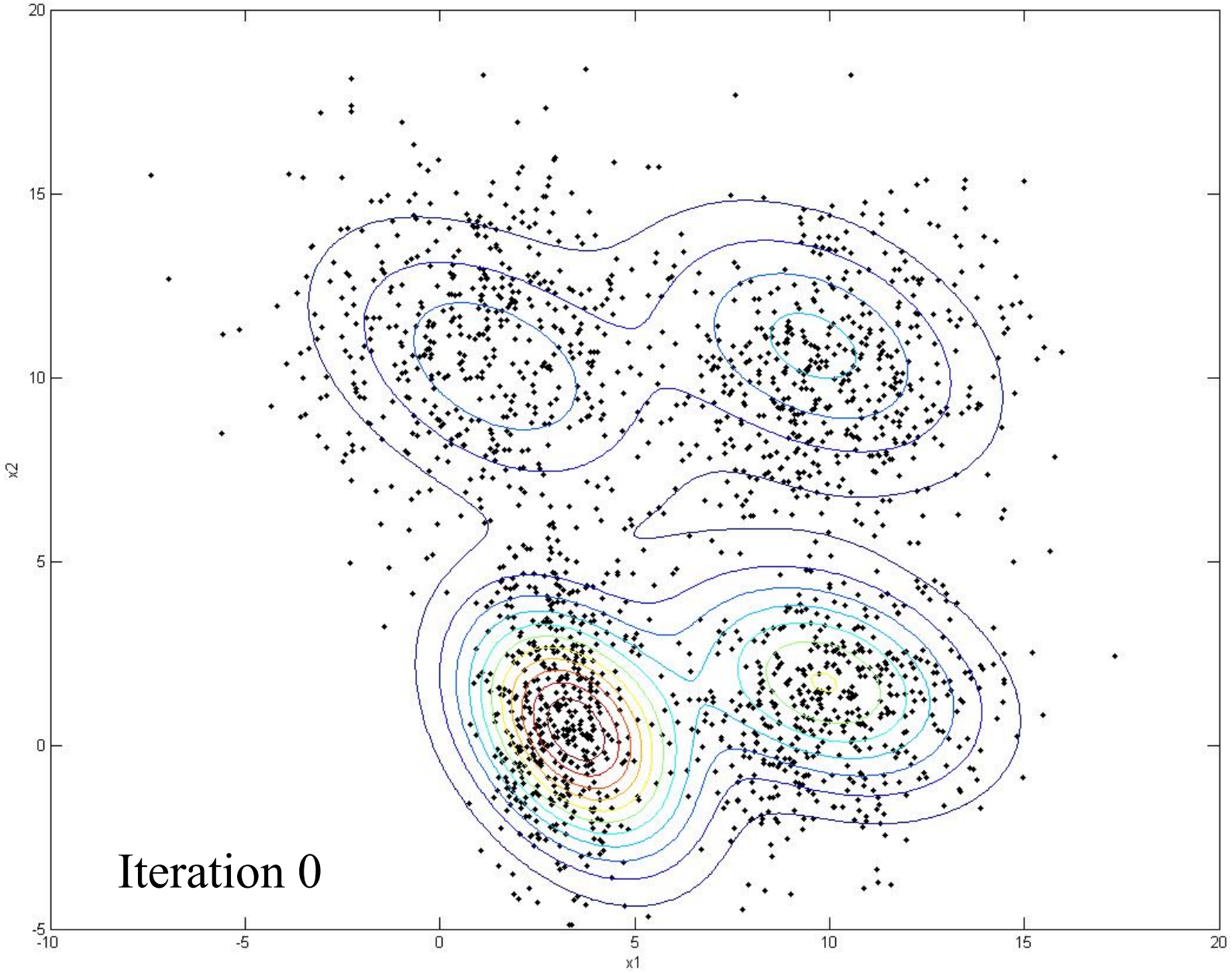
iterate

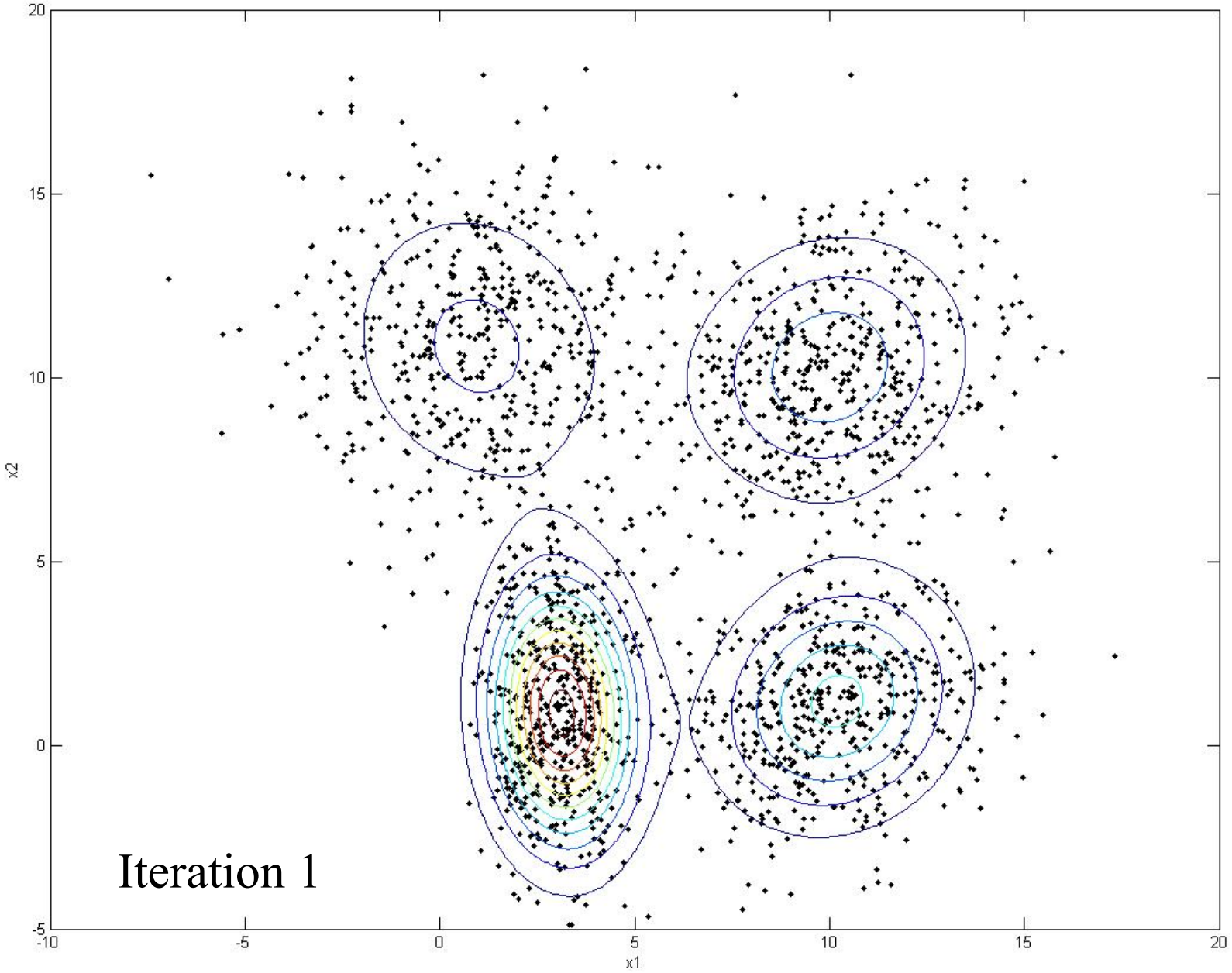
$$S_k^{(t)} = \left\{ \boldsymbol{x}_j : \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_k^{(t)} \right\|^2 \leq \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i^{(t)} \right\|^2 \text{ for all clusters } i = 1,\ldots,I \right\}$$

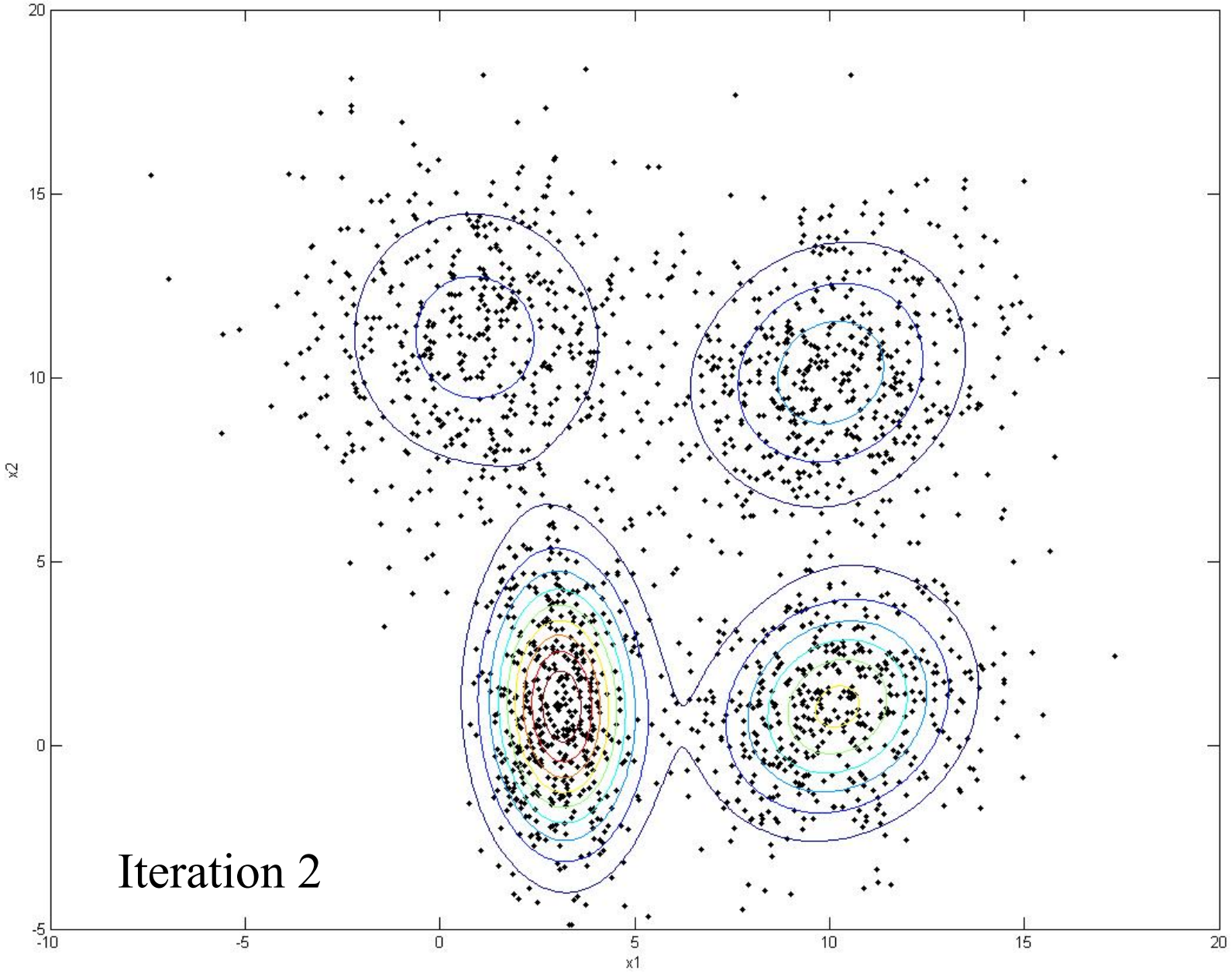
iterate

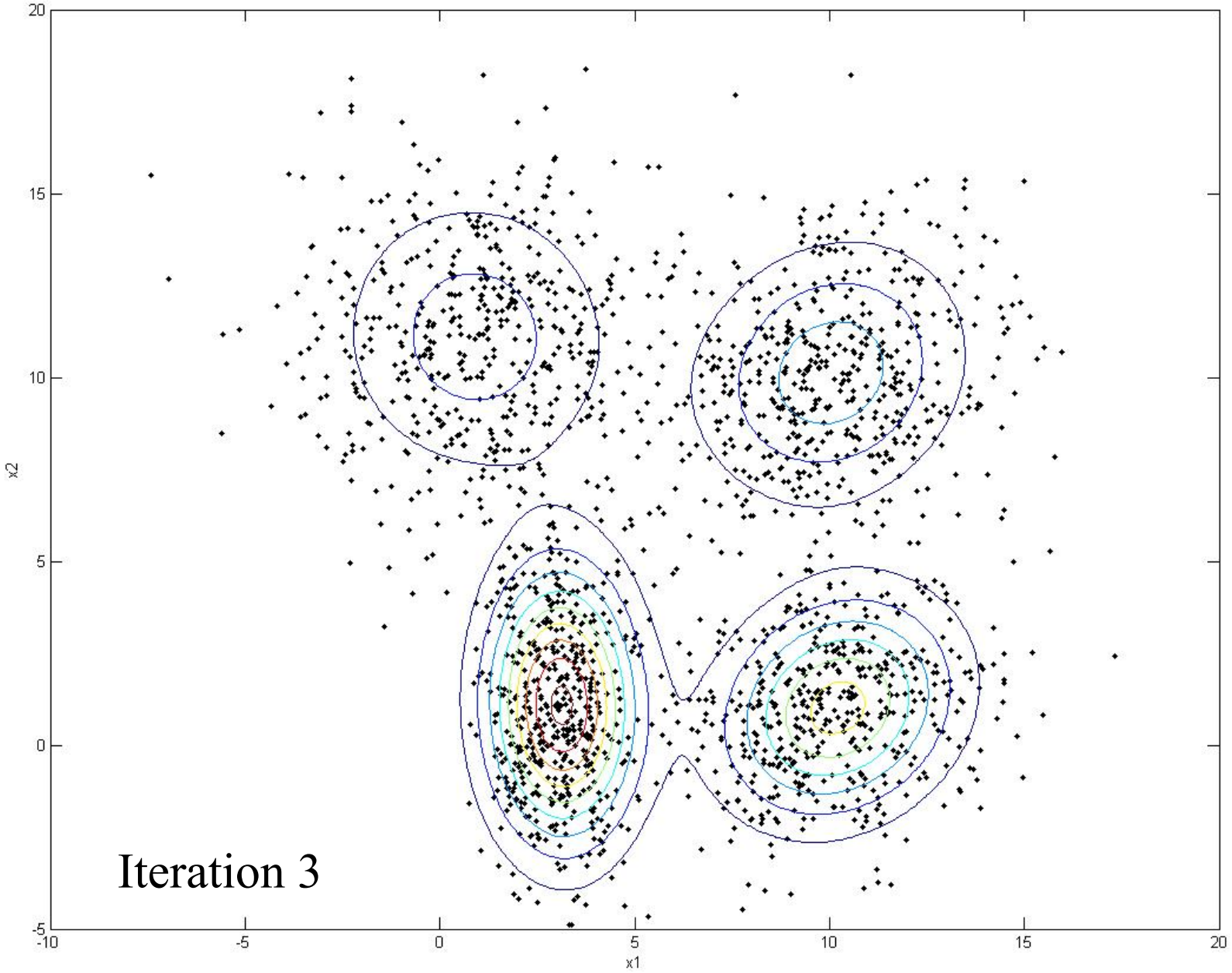
   c. Determine the new means for all clusters with new associated x_j

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{\boldsymbol{x}_j \in S_i^{(t)}} \boldsymbol{x}_j$$


Iteration 0
x1
x2
20
15
10
5
0
-5
-10
-5
0
5
10
15
20


20
15
10
5
0
-5
x2
-10
-5
0
5
10
15
20
x1
Iteration 1


20
15
10
5
0
-5
x2
-10
-5
0
5
10
15
20
x1
Iteration 2


20
15
10
5
0
-5
-10
-5
0
5
10
15
20
x1
x2
Iteration 3

K-Means Clustering

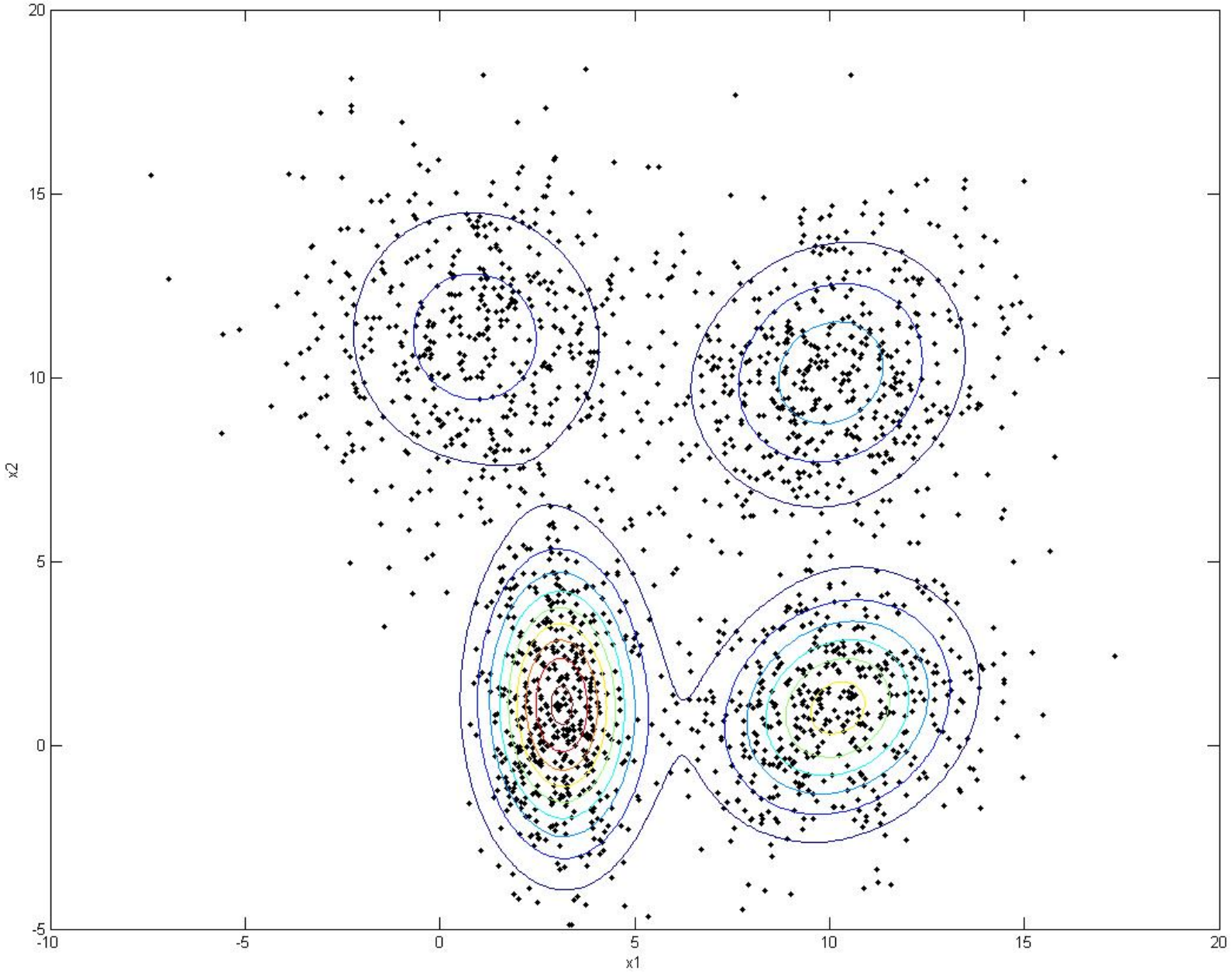
Repeat of split and reestimation until minimum square error achieved.

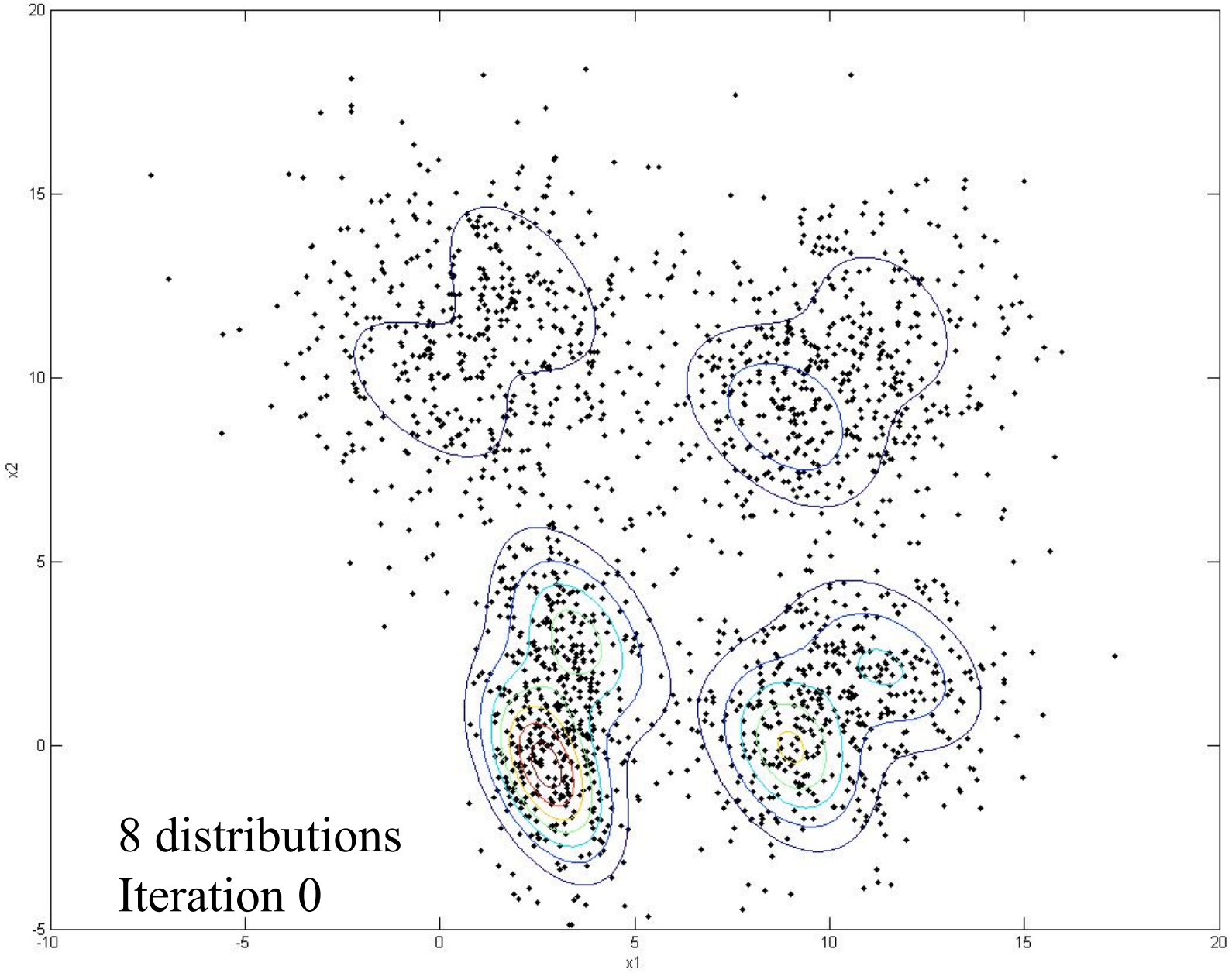
$$G = \sum_{i=1}^{I} \sum_{\boldsymbol{x} \in S_i} \left\| \boldsymbol{x} - \boldsymbol{\mu}_i \right\|^2$$

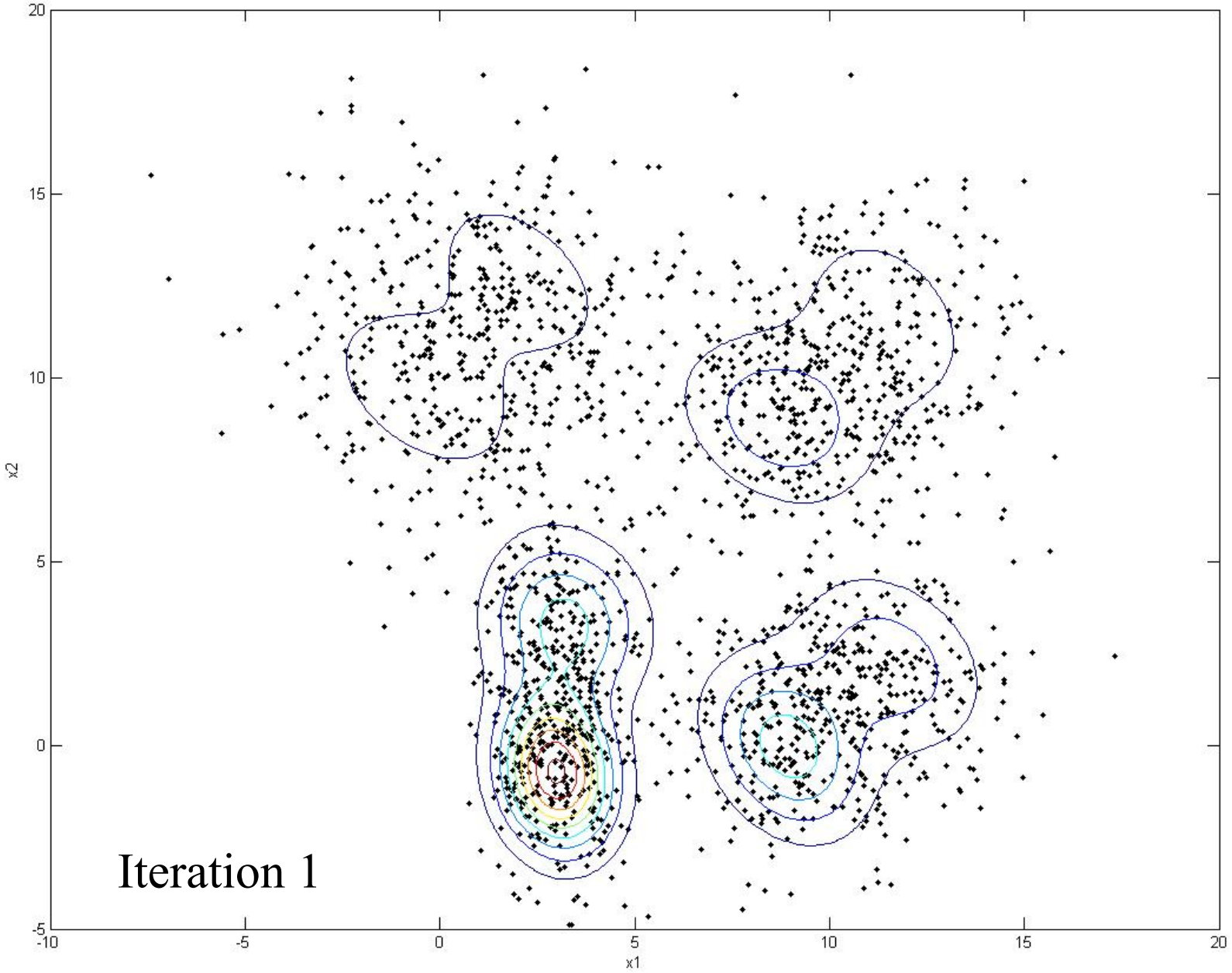
- Determination of optimal number of splits/iterations requires experience
- Possible other criteria:
  - Classification error rate on test data
  - Likelihood increase on training data below threshold
- Possible problem: **Overfitting**

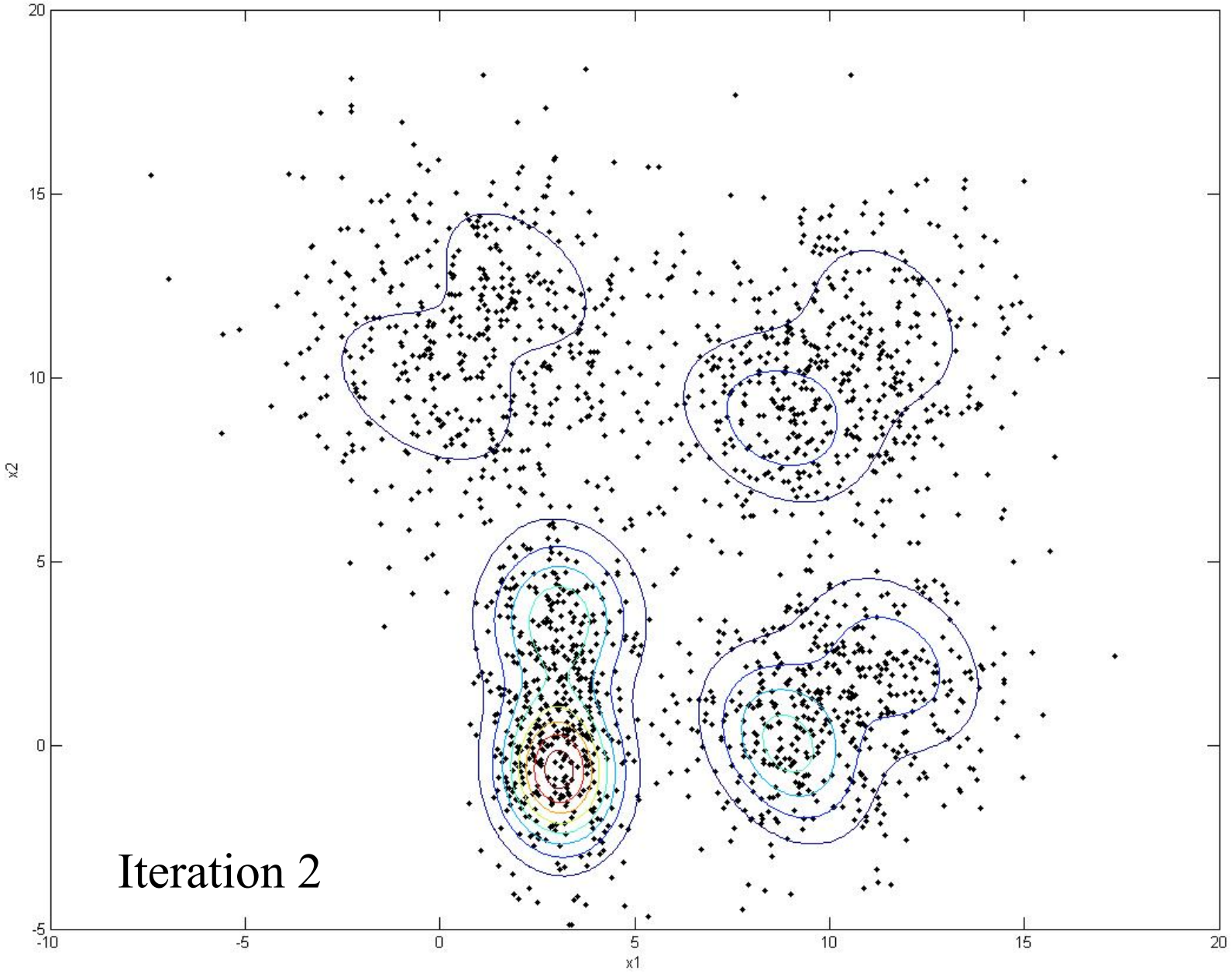
  → Too strong adaptation to training data

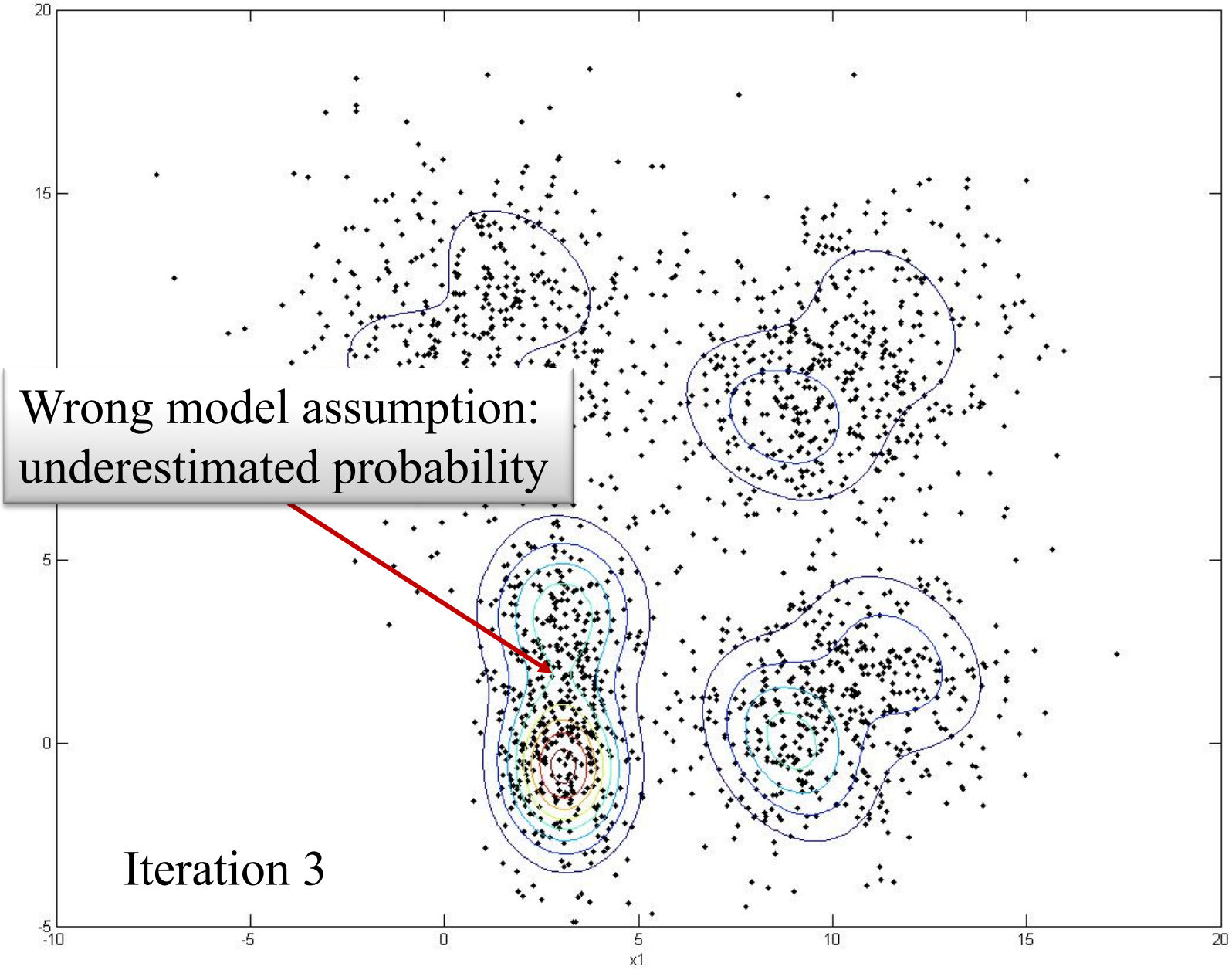
  → No generalization capability of the model to unknown data.


20
15
10
5
0
-5
x2
-10
-5
0
5
10
15
20
x1


20
15
10
5
0
-5
-10
-5
0
5
10
15
20
x1
x2
8 distributions
Iteration 0


Iteration 1
x1
x2
20
15
10
5
0
-5
-10
-5
0
5
10
15
20


20
15
10
5
0
-5
x2
-10
-5
0
5
10
15
20
x1
Iteration 2


Wrong model assumption:
underestimated probability
Iteration 3
20
15
5
0
-5
-10
-5
0
5
10
15
20
x1

1.15 Bayes rule

We already know: $P(x \mid y) = \frac{P(x,y)}{P(y)}$ and $P(y \mid x) = \frac{P(x,y)}{P(x)}$

1.15 Bayes rule

We already know: $P(x \mid y) = \frac{P(x,y)}{P(y)}$ and $P(y \mid x) = \frac{P(x,y)}{P(x)}$

$$P(x \mid y) = \frac{P(x,y)}{P(y)} = \frac{\frac{P(x,y)}{P(x)} \cdot P(x)}{P(y)} = \frac{P(y|x) \cdot P(x)}{P(y)}$$

insert

1.15 Bayes rule

We already know: $P(x \mid y) = \frac{P(x,y)}{P(y)}$ and $P(y \mid x) = \frac{P(x,y)}{P(x)}$

insert

$$P(x \mid y) = \frac{P(x,y)}{P(y)} = \frac{\frac{P(x,y)}{P(x)} \cdot P(x)}{P(y)} = \frac{P(y|x) \cdot P(x)}{P(y)}$$

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

Bayes Rule

1.15 Bayes rule

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

With

conditional probability (see above)

$$P(y) = \sum_{x \in X} P(x, y) = \sum_{x \in X} P(y, x) = \sum_{x \in X} P(y \mid x) \cdot P(x)$$

Law of Total Probability

(Marginal Distribution)

1.15 Bayes rule

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

With

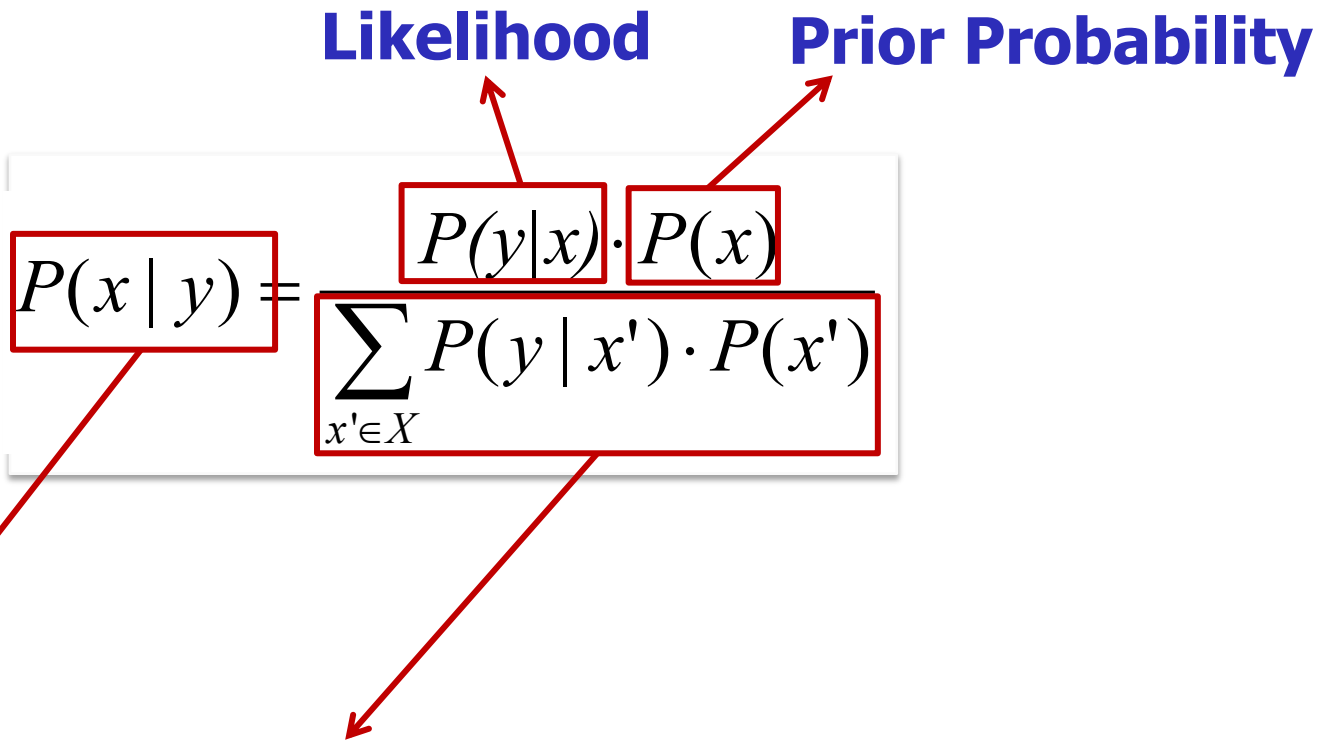
$$P(y) = \sum_{x \in X} P(x, y) = \sum_{x \in X} P(y, x) = \sum_{x \in X} P(y \mid x) \cdot P(x)$$

we get another important representation of Bayes rule:

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)} = \frac{P(y|x) \cdot P(x)}{\sum_{x' \in X} P(y \mid x') \cdot P(x')}$$

1.15 Bayes rule

Terminology:



- does not depend on x
- y is known in the posterior probability P(x | y)

→ does not influence the shape of the distribution P(x|y)

→ just a normalizing term, ensures that P(x|y) sums to 1

1.15 Bayes rule

Relevance and interpretation:

- "Inverts" statistical connections
- If P(y|x) is known but P(x|y) is required

1.15 Bayes rule

Example: Spam-Filter

Given: 10.000 different mails, 5.000 spam, 5.000 ham (not spam)

It is easy to learn the probability of words, given that we know it is spam, that is:

P(Words | Spam) → We know the class (spam) and estimate the probability of words.

However: A spam filter must decide on Spam / Not-Spam based on **a given mail**

→ We know the words (from mail) and want to get the probability of the class

→ We need a model for: P(Spam | Words)

→ How can we invert the conditional probability?

1.15 Bayes rule

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)} = \frac{P(y|x) \cdot P(x)}{\sum_{x \in X} P(y \mid x) \cdot P(x)}$$

Example: Spam-Filter

Answer: with Bayes Rule

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words}|\text{Spam}) \cdot P(\text{Spam})}{\sum_{\substack{\text{all classes} \\ \text{C} \in [\text{Spam, Ham}]}} P(\text{Words} \mid \text{C}) \cdot P(\text{C})}$$

1.15 Bayes rule

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)} = \frac{P(y|x) \cdot P(x)}{\sum_{x \in X} P(y \mid x) \cdot P(x)}$$

Example: Spam-Filter

Answer: with Bayes Rule

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words}|\text{Spam}) \cdot P(\text{Spam})}{\sum_{\substack{\text{all classes} \\ \text{C} \in [\text{Spam, Ham}]}} P(\text{Words} \mid \text{C}) \cdot P(\text{C})}$$

1.15 Bayes rule

Example: Spam-Filter

Answer: with Bayes Rule

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words}|\text{Spam}) \cdot P(\text{Spam})}{\sum_{\substack{\text{all classes} \\ \text{C} \in [\text{Spam, Ham}]}} P(\text{Words} \mid \text{C}) \cdot P(\text{C})}$$

$$= \frac{P(\text{Words}|\text{Spam}) \cdot P(\text{Spam})}{P(\text{Words} \mid \text{Spam}) \cdot P(\text{Spam}) + P(\text{Words} \mid \text{Ham}) \cdot P(\text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

Answer: with Bayes Rule

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{Spam}) \cdot P(\text{Spam})}{\sum_{\text{all classes}} P(\text{Words} \mid \text{C}) \cdot P(\text{C})}$$

"afterwards"

all classes
C ∈ [Spam, Ham]

Posterior Probability

Probability of class spam

after we observed the words.

1.15 Bayes rule

Example: Spam-Filter

Answer: with Bayes Rule

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{Spam}) \cdot P(\text{Spam})}{\sum_{\substack{\text{all classes} \\ C \in [\text{Spam, Ham}]}} P(\text{Words} \mid C) \cdot P(C)}$$

Likelihood or class dependent probability

Probability of words if we know that it is spam.

1.15 Bayes rule

Example: Spam-Filter

Answer: with Bayes Rule

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{Spam}) \cdot P(\text{Spam})}{\sum_{\substack{\text{all classes} \\ \text{C} \in [\text{Spam, Ham}]}} P(\text{Words} \mid \text{C}) \cdot P(\text{C})}$$

Prior Spam Probability

Probability of class spam

before we made any observation

(i.e. without observation)

1.15 Bayes rule

Example: Spam-Filter

$$P(\text{Spam} \mid \text{Words})$$

$$= \frac{P(\text{Words} \mid \text{Spam}) \cdot P(\text{Spam})}{P(\text{Words} \mid \text{Spam}) \cdot P(\text{Spam}) + P(\text{Words} \mid \text{Ham}) \cdot P(\text{Ham})}$$

How do we get the required probabilities?

→ from training data!

1.15 Bayes rule

Example: Spam-Filter

P(Spam), P(Ham) ?

→ Realistic value: P(Spam) = 0.8, P(Ham) = 0.2

→ Most spam filters, however, use P(Spam)=0.5, P(Ham)=0.5

(give each incoming mail a chance)

1.15 Bayes rule

Example: Spam-Filter

Simplification: Same prior probability of spam and ham

$$P(\text{Spam} \mid \text{Words}) = \frac{P(\text{Words} \mid \text{Spam})}{P(\text{Words} \mid \text{Spam}) + P(\text{Words} \mid \text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

P(Words | Spam), P(Words | Ham) ?

→ Probability of a word in spam mails can be estimated from training spams

P(Word | Spam) = Frequency of word in spam / Total number of spam-words

But this is the estimation of P(Word | Spam) instead of P(Words | Spam)

Assumption: word probabilities are statistically independent

→ we can factorize as follows:

$$P(W_1, W_2, ..., W_N \mid Spam) = P(W_1 \mid Spam) \cdot P(W_2 \mid Spam) \cdot ... \cdot P(W_N \mid Spam)$$

1.15 Bayes rule

Example: Spam-Filter

Details:

$$
\begin{aligned}
&P(\text{Words} \mid \text{Spam}) = P(\text{Word1}, \text{Word2}, \ldots, \text{WordN} \mid \text{Spam}) \\
&= P(\text{Word1} \mid \text{Word2}, \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2}, \text{Word3}, \ldots, \text{WordN} \mid \text{Spam}) \\
&= P(\text{Word1} \mid \text{Word2}, \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2} \mid \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word3}, \ldots, \text{WordN} \mid \text{Spam}) \\
&= \ldots \\
&= P(\text{Word1} \mid \text{Word2}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2} \mid \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot \ldots \cdot P(\text{WordN-1} \mid \text{WordN}, \text{Spam}) \cdot P(\text{WordN} \mid \text{Spam})
\end{aligned}
$$

1.15 Bayes rule

Example: Spam-Filter

Details:

$$P(\text{Words} \mid \text{Spam}) = P(\text{Word1}, \text{Word2}, \ldots, \text{WordN} \mid \text{Spam})$$
$$= P(\text{Word1} \mid \text{Word2}, \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2}, \text{Word3}, \ldots, \text{WordN} \mid \text{Spam})$$
$$= P(\text{Word1} \mid \text{Word2}, \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2} \mid \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word3}, \ldots, \text{WordN} \mid \text{Spam})$$
$$= \ldots$$
$$= P(\text{Word1} \mid \text{Word2}, \ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2} \mid \text{Word3}, \ldots, \text{WordN}, \text{Spam}) \cdot \ldots \cdot P(\text{WordN-1} \mid \text{WordN}, \text{Spam}) \cdot P(\text{WordN} \mid \text{Spam})$$

Naive assumption which often works well in practice:

Probability of one word is not influenced by other words in the mail.

1.15 Bayes rule

Example: Spam-Filter

Details:

$$P(\text{Words}\,|\,\text{Spam}) = P(\text{Word1}, \text{Word2},\ldots, \text{WordN}\,|\,\text{Spam})$$
$$= P(\text{Word1}\,|\,\text{Word2}, \text{Word3},\ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2}, \text{Word3},\ldots, \text{WordN}\,|\,\text{Spam})$$
$$= P(\text{Word1}\,|\,\text{Word2}, \text{Word3},\ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2}\,|\,\text{Word3},\ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word3},\ldots, \text{WordN}\,|\,\text{Spam})$$
$$= \ldots$$
$$= P(\text{Word1}\,|\,\text{Word2},\ldots, \text{WordN}, \text{Spam}) \cdot P(\text{Word2}\,|\,\text{Word3},\ldots, \text{WordN}, \text{Spam}) \cdot \ldots \cdot P(\text{WordN-1}\,|\,\text{WordN}, \text{Spam}) \cdot P(\text{WordN}\,|\,\text{Spam})$$

Naive assumption which often works well in practice:

Probability of one word is not influenced by other words in the mail.

$$P(\text{Word1}, \text{Word2},\ldots, \text{WordN}\,|\,\text{Spam})$$
$$= P(\text{Word1}\,|\,\text{Spam}) \cdot P(\text{Word2}\,|\,\text{Spam}) \cdot \ldots \cdot P(\text{WordN}\,|\,\text{Spam})$$

1.15 Bayes rule

Example: Spam-Filter

Using this so-called Naive Bayes Rule leads to

$$P(\text{Spam} \mid w_1, w_2, ..., w_N)$$

$$= \frac{P(w_1, w_2, ..., w_N \mid \text{Spam})}{P(w_1, w_2, ..., w_N \mid \text{Spam}) + P(w_1, w_2, ..., w_N \mid \text{Ham})}$$

$$= \frac{P(w_1 \mid \text{Spam}) \cdot P(w_2 \mid \text{Spam}) \cdot ... \cdot P(w_N \mid \text{Spam})}{P(w_1 \mid \text{Spam}) \cdot ... \cdot P(w_N \mid \text{Spam}) + P(w_1 \mid \text{Ham}) \cdot ... \cdot P(w_N \mid \text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

Similar we obtain

$$P(\text{Ham} \mid w_1, w_2, ..., w_N)$$

$$= \frac{P(w_1 \mid \text{Ham}) \cdot P(w_2 \mid \text{Ham}) \cdot ... \cdot P(w_N \mid \text{Ham})}{P(w_1 \mid \text{Spam}) \cdot ... \cdot P(w_N \mid \text{Spam}) + P(w_1 \mid \text{Ham}) \cdot ... \cdot P(w_N \mid \text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

Decision (Spam/Ham) may be based on **Likelihood-Ratio**:

$$\frac{P(\text{Spam} \mid w_1, ..., w_N)}{P(\text{Ham} \mid w_1, ..., w_N)} = \frac{P(w_1 \mid \text{Spam}) \cdot ... \cdot P(w_N \mid \text{Spam})}{P(w_1 \mid \text{Ham}) \cdot ... \cdot P(w_N \mid \text{Ham})}$$

$$\frac{P(\text{Spam} \mid w_1, ..., w_N)}{P(\text{Ham} \mid w_1, ..., w_N)} = \frac{\prod_i P(w_i \mid \text{Spam})}{\prod_i P(w_i \mid \text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

Likelihood-Ratio:

$$\frac{P(\text{Spam} \mid w_1, ..., w_N)}{P(\text{Ham} \mid w_1, ..., w_N)} = \prod_i \frac{P(w_i \mid \text{Spam})}{P(w_i \mid \text{Ham})}$$

Log-Likelihood-Ratio:

$$\log \frac{P(\text{Spam} \mid w_1, ..., w_N)}{P(\text{Ham} \mid w_1, ..., w_N)} = \sum_i \log \frac{P(w_i \mid \text{Spam})}{P(w_i \mid \text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

Decision:

$$\sum_i \log \frac{P(\mathrm{w_i} \mid \mathrm{Spam})}{P(\mathrm{w_i} \mid \mathrm{Ham})} > 0 \quad \Rightarrow \quad \textbf{Spam}$$

$$\sum_i \log \frac{P(\mathrm{w_i} \mid \mathrm{Spam})}{P(\mathrm{w_i} \mid \mathrm{Ham})} < 0 \quad \Rightarrow \quad \textbf{Ham}$$

1.15 Bayes rule

Example: Spam-Filter

Example Spam Mail

Let us just take the first 4 words for the decision

```
Order Original Viagra directly from Pfizer

Here: http://www.dummynameshop.com/

All prices are tax/vat free and same-day free worldwide
shipping also included.
```

1.15 Bayes rule

Example: Spam-Filter

$$\text{Log}-\text{Lik}-\text{Ratio} = \log\frac{P(\text{order}\,|\,\text{Spam})}{P(\text{order}\,|\,\text{Ham})} + \log\frac{P(\text{original}\,|\,\text{Spam})}{P(\text{original}\,|\,\text{Ham})}$$
$$+ \log\frac{P(\text{viagra}\,|\,\text{Spam})}{P(\text{viagra}\,|\,\text{Ham})} + \log\frac{P(\text{directly}\,|\,\text{Spam})}{P(\text{directly}\,|\,\text{Ham})}$$

1.15 Bayes rule

Example: Spam-Filter

Estimate P(word|Spam) and P(word|Ham) from training data:

Spam corpus: 467748 words

word	frequency in spam corpus	probability P(word|Spam)
order	355	$7.59 \cdot 10^{-4}$
original	179	$3.83 \cdot 10^{-4}$
viagra	422	$9.02 \cdot 10^{-4}$
directly	114	$2.44 \cdot 10^{-4}$

1.15 Bayes rule

Example: Spam-Filter

Estimate P(word|Spam) and P(word|Ham) from training data:

We use Hamlet as Ham corpus: 32630 words

(in praxis: corpora should have similar sizes)

word	frequency in ham corpus	probability P(word|Ham)
order	1	$3.06 \cdot 10^{-5}$
original	0	0
viagra	0	0
directly	1	$3.06 \cdot 10^{-5}$

Problem: Division by zero

1.15 Bayes rule

Example: Spam-Filter

Practical solution to the problem of unseen words

Take probability mass from 'singletons' (words with only one observation) and distribute it equally to unseen words.

word	frequency in ham corpus	probability P(word|Spam)
order	1	$2.36 \cdot 10^{-5}$
original	0	$0.13 \cdot 10^{-5}$
viagra	0	$0.13 \cdot 10^{-5}$
directly	1	$2.36 \cdot 10^{-5}$

reduced probability (hypothetical numbers!)

1.15 Bayes rule

Example: Spam-Filter

Practical solution to the problem of unseen words

Take probability mass from 'singletons' (words with only one observation) and distribute it equally to unseen words.

word	frequency in ham corpus	probability P(word|Spam)
order	1	$2.36 \cdot 10^{-5}$
original	0	$0.13 \cdot 10^{-5}$
viagra	0	$0.13 \cdot 10^{-5}$
directly	1	$2.36 \cdot 10^{-5}$

increased probability (hypothetical numbers!)

1.15 Bayes rule

$$\text{Log}-\text{Lik}-\text{Ratio} = \log\frac{P(\text{order}\,|\,\text{Spam})}{P(\text{order}\,|\,\text{Ham})} + \log\frac{P(\text{original}\,|\,\text{Spam})}{P(\text{original}\,|\,\text{Ham})} + \log\frac{P(\text{viagra}\,|\,\text{Spam})}{P(\text{viagra}\,|\,\text{Ham})} + \log\frac{P(\text{directly}\,|\,\text{Spam})}{P(\text{directly}\,|\,\text{Ham})}$$

Example: Spam-Filter

$$\begin{aligned}\text{Log}-\text{Lik}-\text{Ratio} &= \log\frac{7.59\cdot10^{-4}}{2.36\cdot10^{-5}} + \log\frac{3.83\cdot10^{-4}}{0.13\cdot10^{-5}} \\ &\quad + \log\frac{9.02\cdot10^{-4}}{0.13\cdot10^{-5}} + \log\frac{2.44\cdot10^{-4}}{2.36\cdot10^{-5}} \\ &= 18.04\end{aligned}$$

word	frequency in spam corpus	probability P(word|Spam)
order	355	$7.59\cdot10^{-4}$
original	179	$3.83\cdot10^{-4}$
viagra	422	$9.02\cdot10^{-4}$
directly	114	$2.44\cdot10^{-4}$

word	frequency in ham corpus	probability P(word|Spam)
order	1	$2.36\cdot10^{-5}$
original	0	$0.13\cdot10^{-5}$
viagra	0	$0.13\cdot10^{-5}$
directly	1	$2.36\cdot10^{-5}$

1.15 Bayes rule

$$\text{Log}-\text{Lik}-\text{Ratio} = \log\frac{P(\text{order}\,|\,\text{Spam})}{P(\text{order}\,|\,\text{Ham})} + \log\frac{P(\text{original}\,|\,\text{Spam})}{P(\text{original}\,|\,\text{Ham})} + \log\frac{P(\text{viagra}\,|\,\text{Spam})}{P(\text{viagra}\,|\,\text{Ham})} + \log\frac{P(\text{directly}\,|\,\text{Spam})}{P(\text{directly}\,|\,\text{Ham})}$$

Example: Spam-Filter

$$\text{Log}-\text{Lik}-\text{Ratio} = \log\frac{7.59\cdot10^{-4}}{2.36\cdot10^{-5}} + \log\frac{3.83\cdot10^{-4}}{0.13\cdot10^{-5}} + \log\frac{9.02\cdot10^{-4}}{0.13\cdot10^{-5}} + \log\frac{2.44\cdot10^{-4}}{2.36\cdot10^{-5}} = 18.04 \Rightarrow \textbf{Spam}$$

word	frequency in spam corpus	probability P(word|Spam)
order	355	$7.59\cdot10^{-4}$
original	179	$3.83\cdot10^{-4}$
viagra	422	$9.02\cdot10^{-4}$
directly	114	$2.44\cdot10^{-4}$

word	frequency in ham corpus	probability P(word|Spam)
order	1	$2.36\cdot10^{-5}$
original	0	$0.13\cdot10^{-5}$
viagra	0	$0.13\cdot10^{-5}$
directly	1	$2.36\cdot10^{-5}$