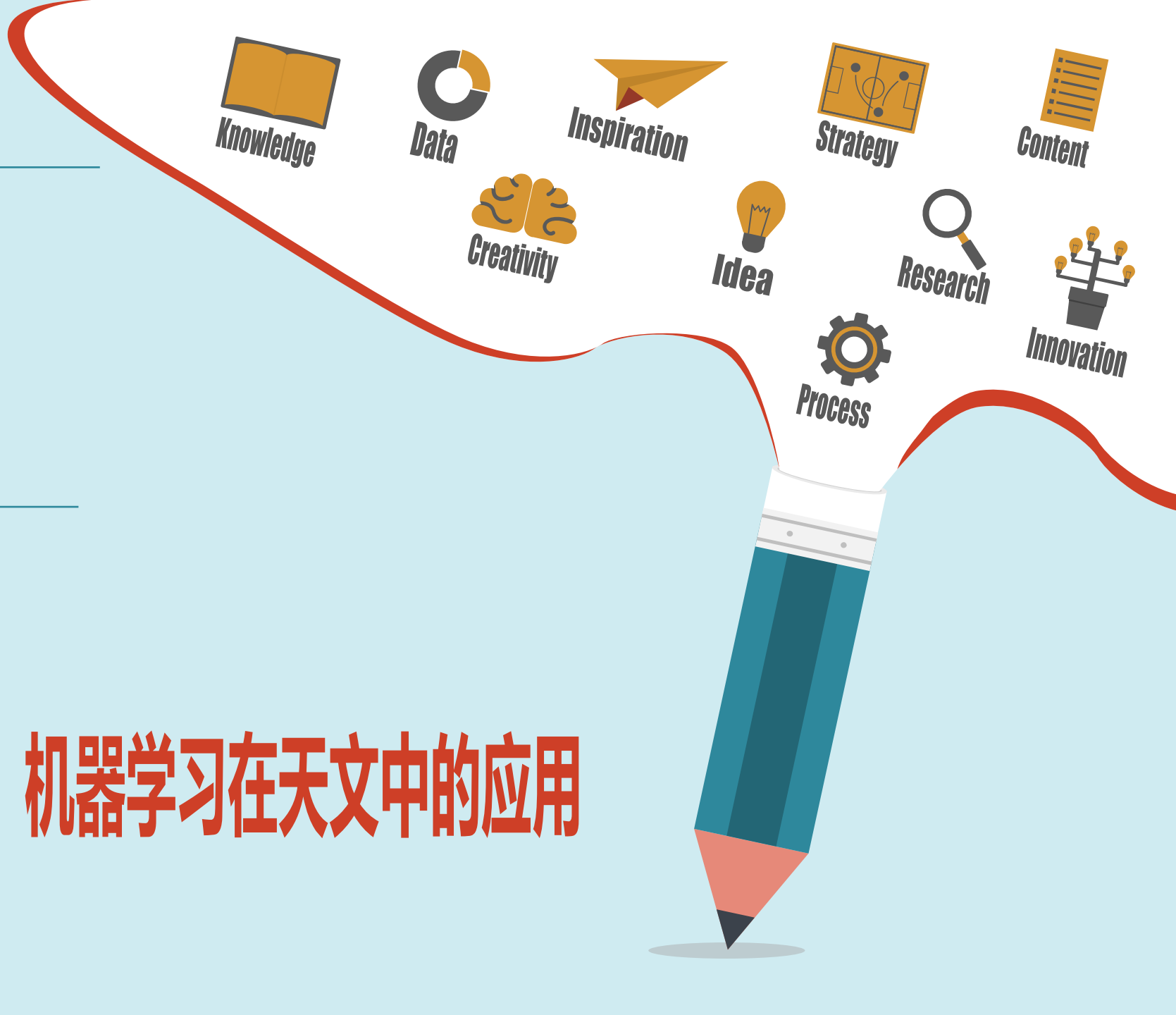
概率生成模型

Generative Models

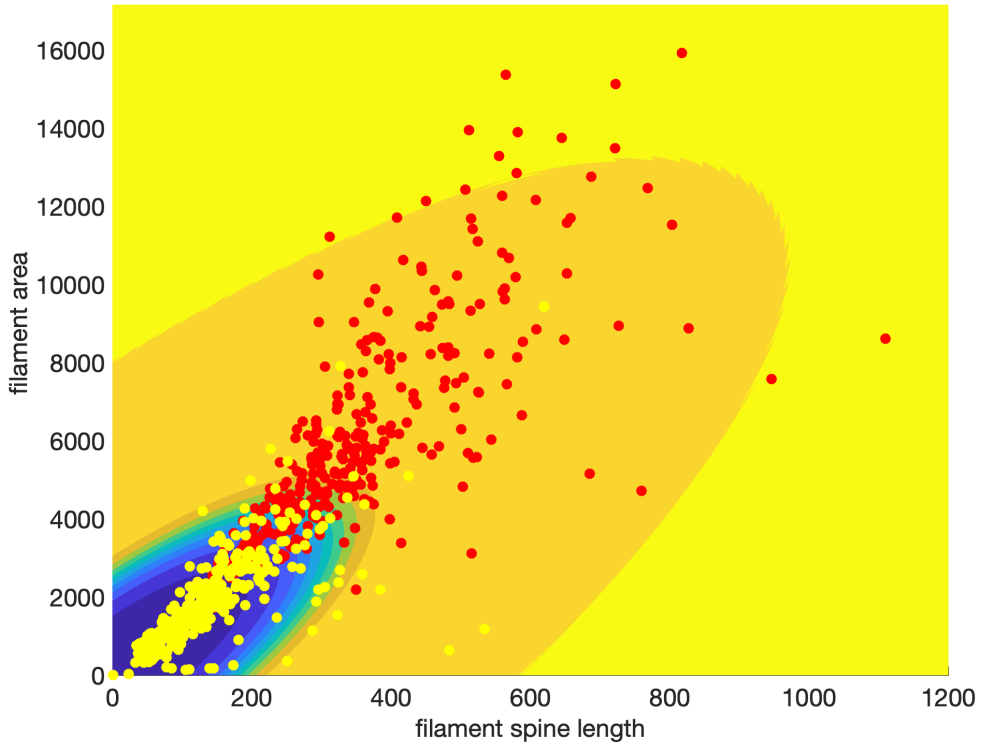
郝　奇
南京大学 天文与空间科学学院

Application of Machine Learning in Astronomy

机器学习在天文中的应用



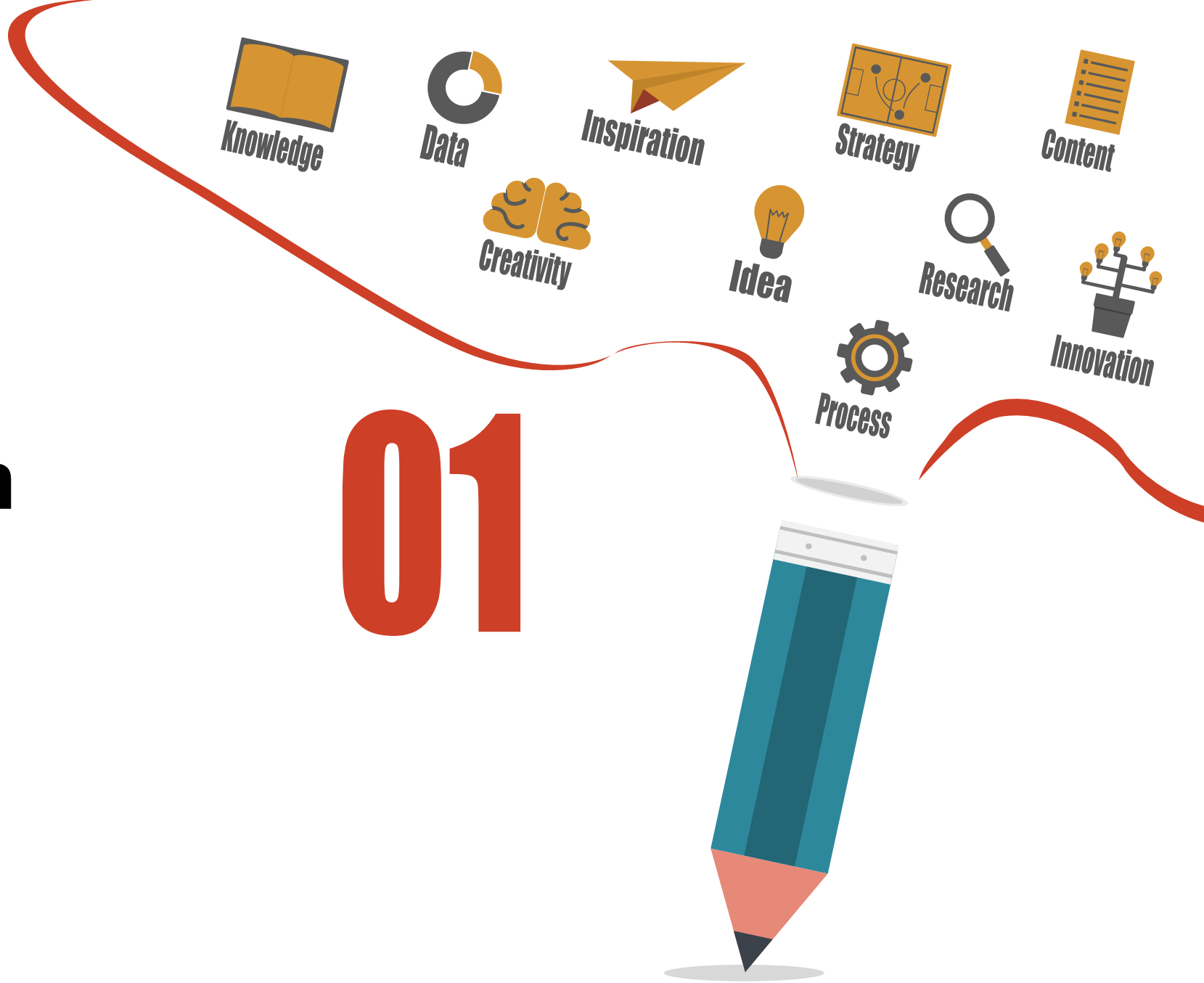
Contents



Cite: Hung-yi Lee, Machine Learning

Probability distribution

01



1. Probability distribution

1.1 Classification vs Regression

Assume a binary classification task, learn a classifier $f(x)$:

$$f(\boldsymbol{X}_i)\begin{cases} \geq 0 & y_i = 1 \\ < 0 & y_i = -1 \end{cases} \quad (\boldsymbol{X}_i, y_i), i = 1, 2, \dots, n, \boldsymbol{X}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$$

Training:

- Class 1 means the target is 1;
- Class 2 means the target is -1;

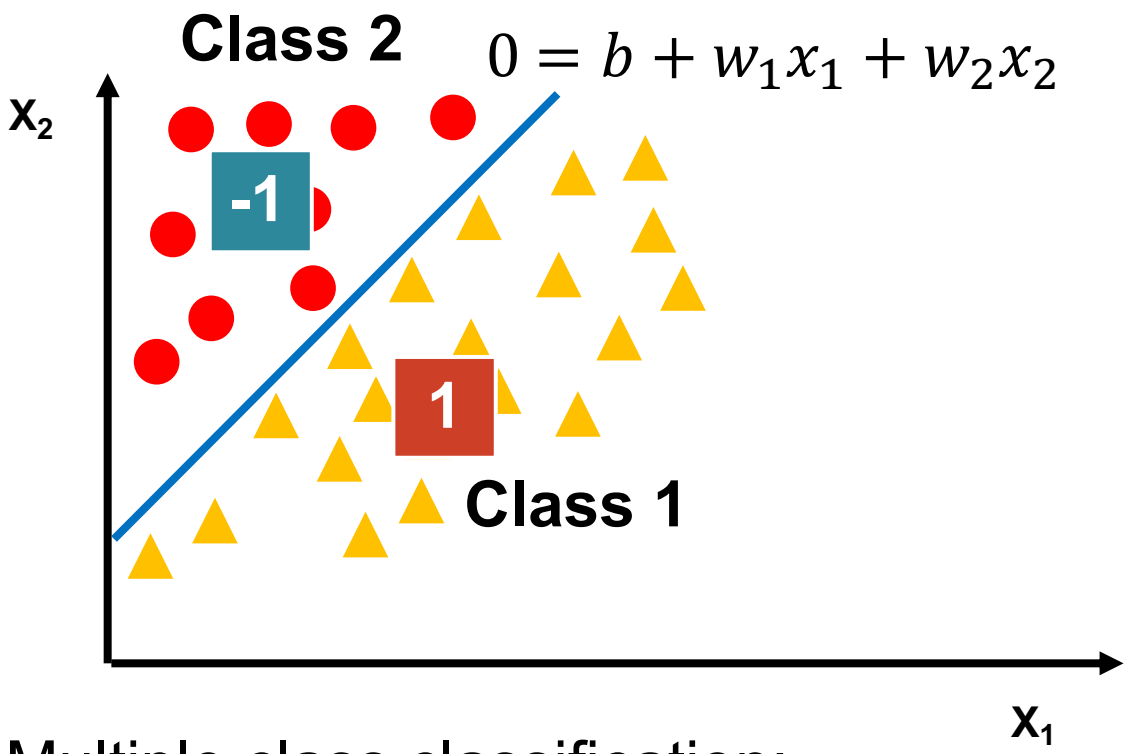
Testing:

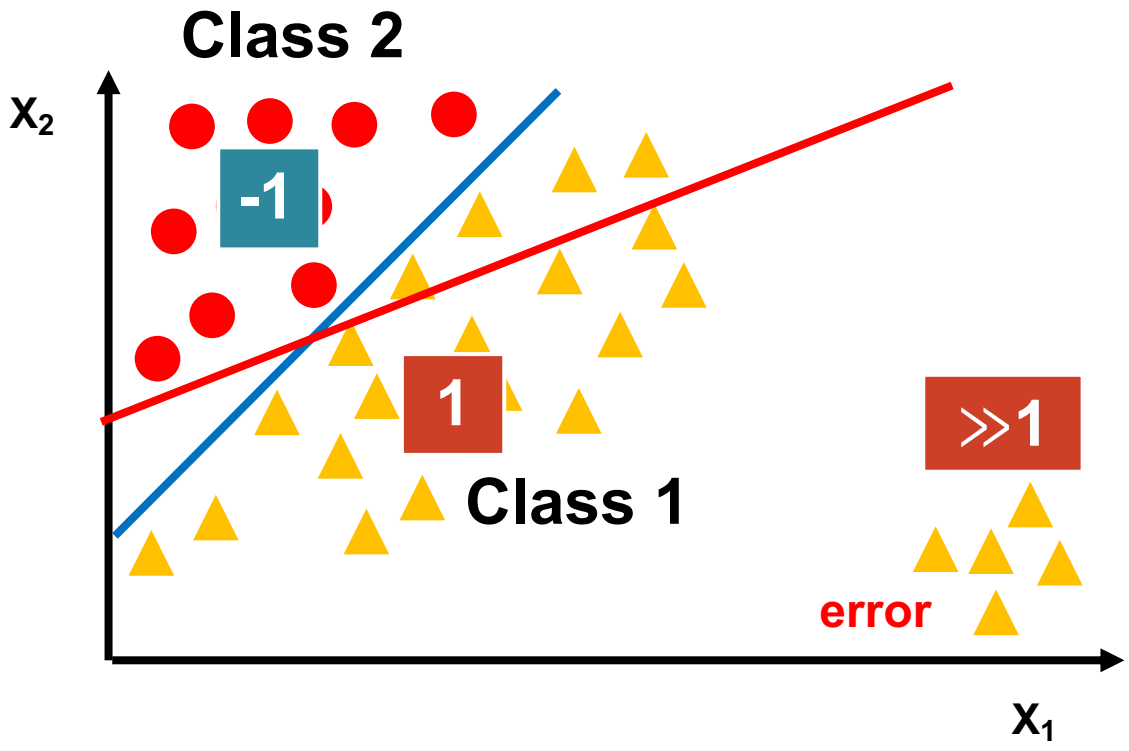
- Closer to 1 → class 1;
- Closer to -1 → class 2.

1. Probability distribution

1.1 Classification vs Regression

$$y = b + w_1 x_1 + w_2 x_2$$





Multiple class classification:

- Class 1 means the target is 1; Class 2 means the target is 2; Class 3 means the target is 3 ?
- Is it correct?

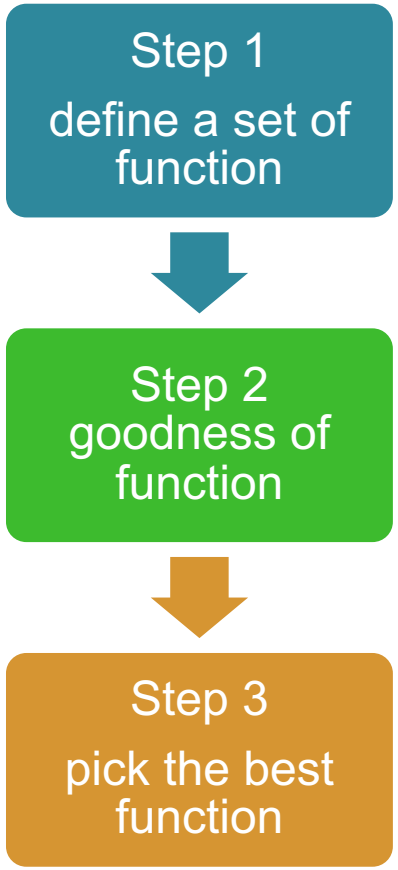
1. Probability distribution

1.1 Classification vs Regression

Previous Lecture

1. Probability distribution

1.1 Classification vs Regression



$$\boldsymbol{X} \rightarrow f(x) = \begin{cases} g(x) > 0 & \text{Output} = \text{class 1} \\ else & \text{Output} = \text{class 2} \end{cases}$$

Loss function: $\mathrm{L}(w, b) = \sum_n \delta\left(f(x^n) \neq \hat{y}^n\right)$

The number of times f get incorrect results on training data

Perceptron, SVM, and Generative model

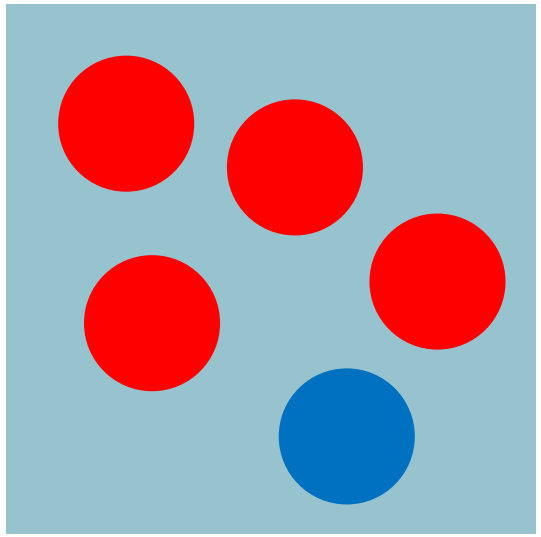
1. Probability distribution

1.2 Probability

Two boxes

Box 1

$P(B_1) = 2/3$

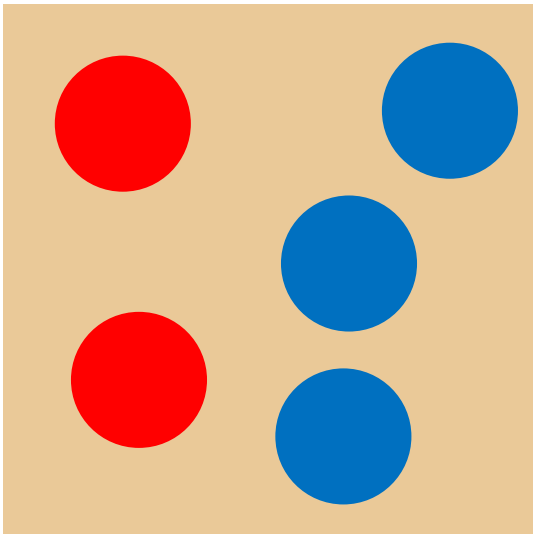


$P(Red|B_1) = 4/5$

$P(Blue|B_1) = 1/5$

Box 2

$P(B_2) = 1/3$



$P(Red|B_2) = 2/5$

$P(Blue|B_2) = 3/5$

● from box 1 :

$$P(B_1|Red) = \frac{P(Red|B_1)P(B_1)}{P(Red|B_1)P(B_1) + P(Red|B_2)P(B_2)}$$

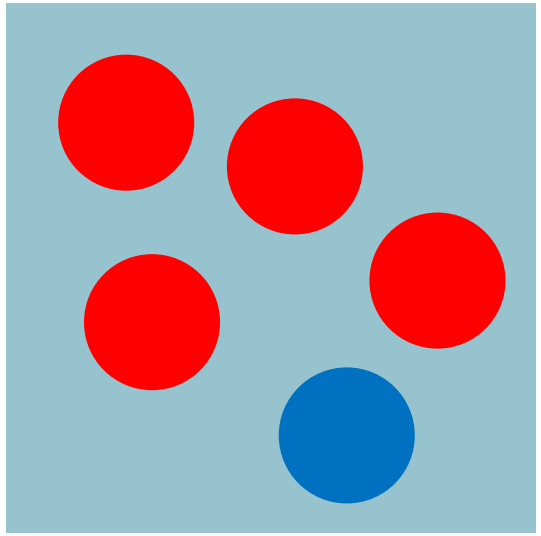
1. Probability distribution

1.3 Generative model **Estimating the probabilities from the training data**

Two classes

Class 1

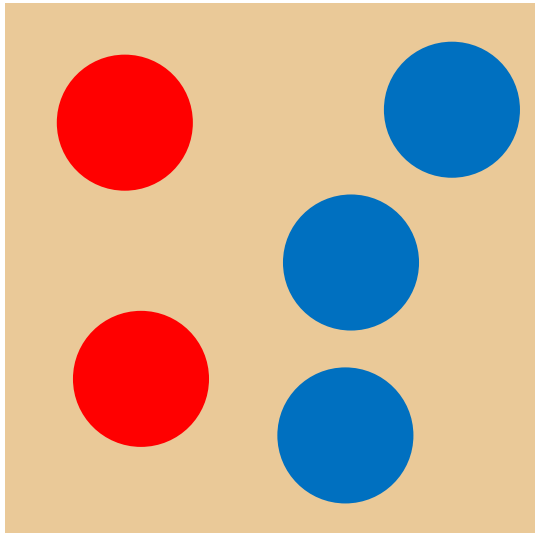
$P(C_1)$



$P(x|C_1)$

Class 2

$P(C_2)$



$P(x|C_2)$

Given an x, which class does it belong to:

$$\boldsymbol{P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}}$$

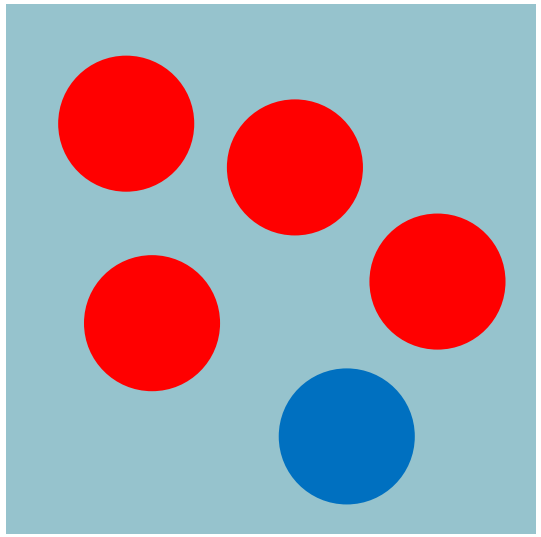
1. Probability distribution

1.3 Generative model

Two classes

Class 1

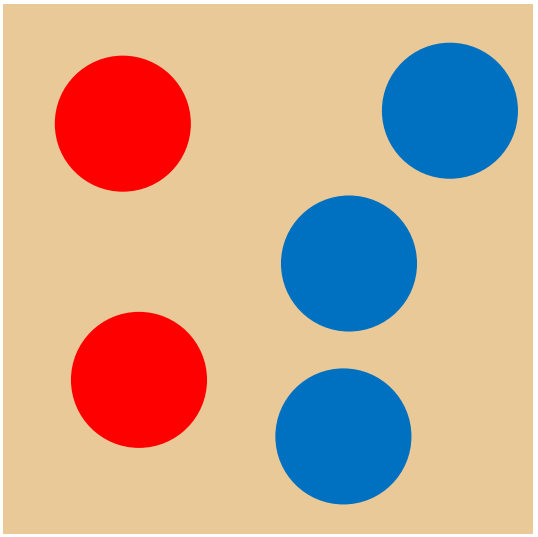
$P(C_1)$



$P(x|C_1)$

Class 2

$P(C_2)$



$P(x|C_2)$

Generative model: $\boldsymbol{P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$

1. Probability distribution

1.3 Generative model

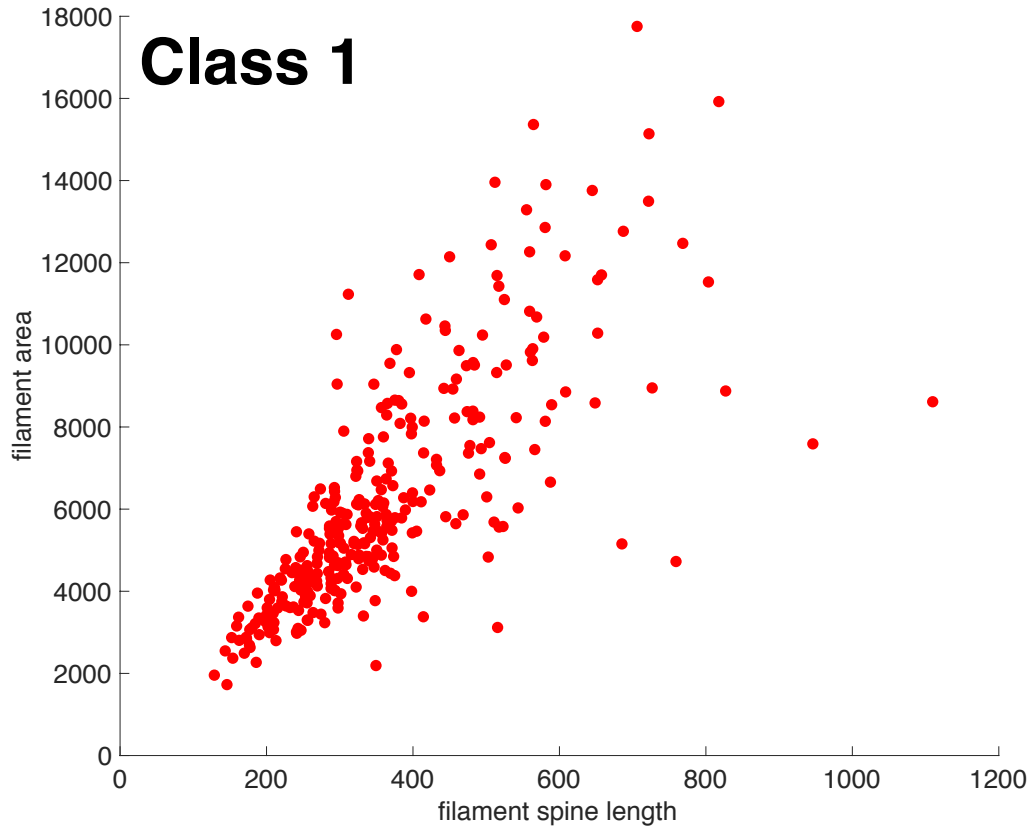
Data

Date	Latitude	Longitude	Area	Spine length	Barb Number	Tilt angle	Quiescent filament
2011-07-01	83.42	338.75	5160.42	319.65	11	-28.47	1
2013-07-14	-218.13	-356.48	10677.93	568.90	25	23.46	1
2014-08-04	-97.91	-218.33	8642.63	380.44	18	-34.36	1
2016-06-22	-424.80	417.06	3115.01	159.19	11	-29.67	1
2011-11-16	221.32	-611.63	1159.69	81.15	3	23.29	0
2014-12-29	-431.07	-362.29	2762.41	134.87	6	11.12	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

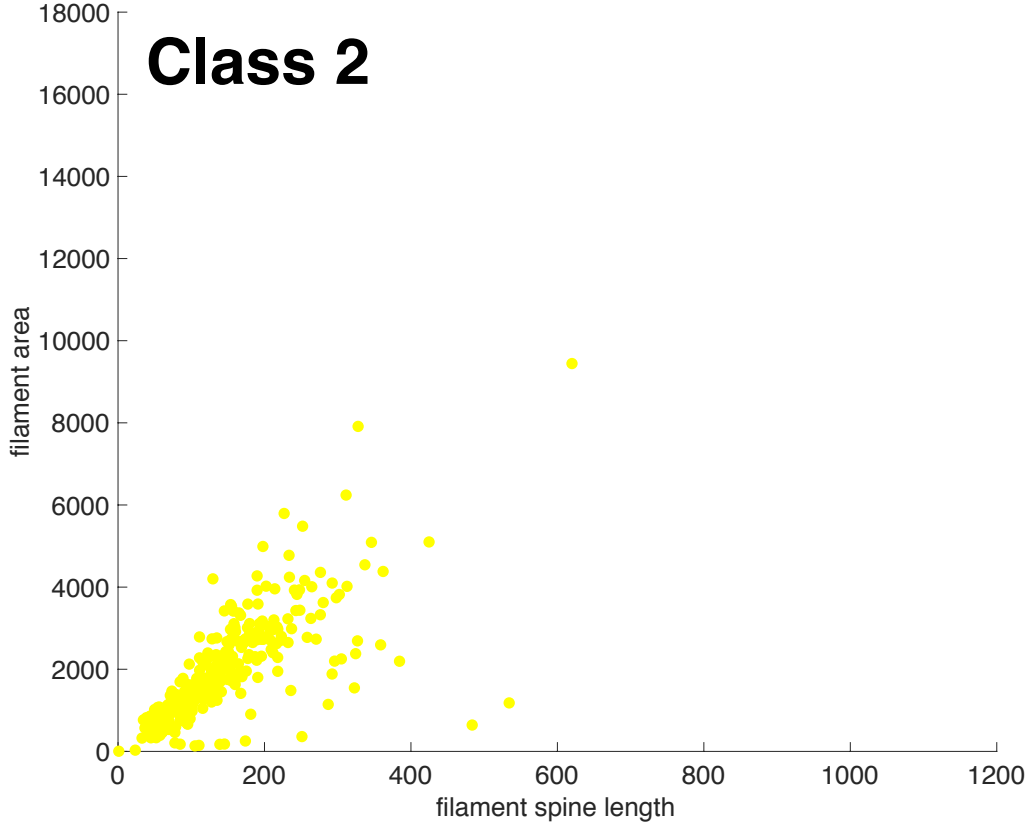
1. Probability distribution

1.3 Generative model

Training Data



Quiescent filament

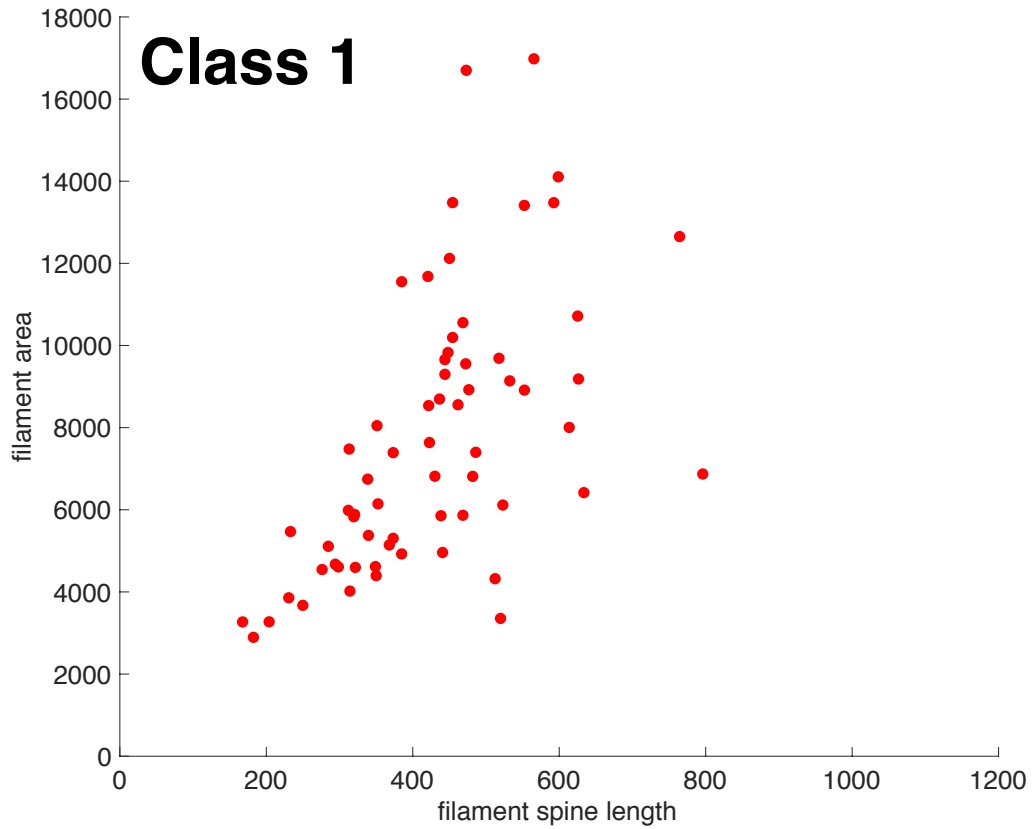


Non-quiescent filament

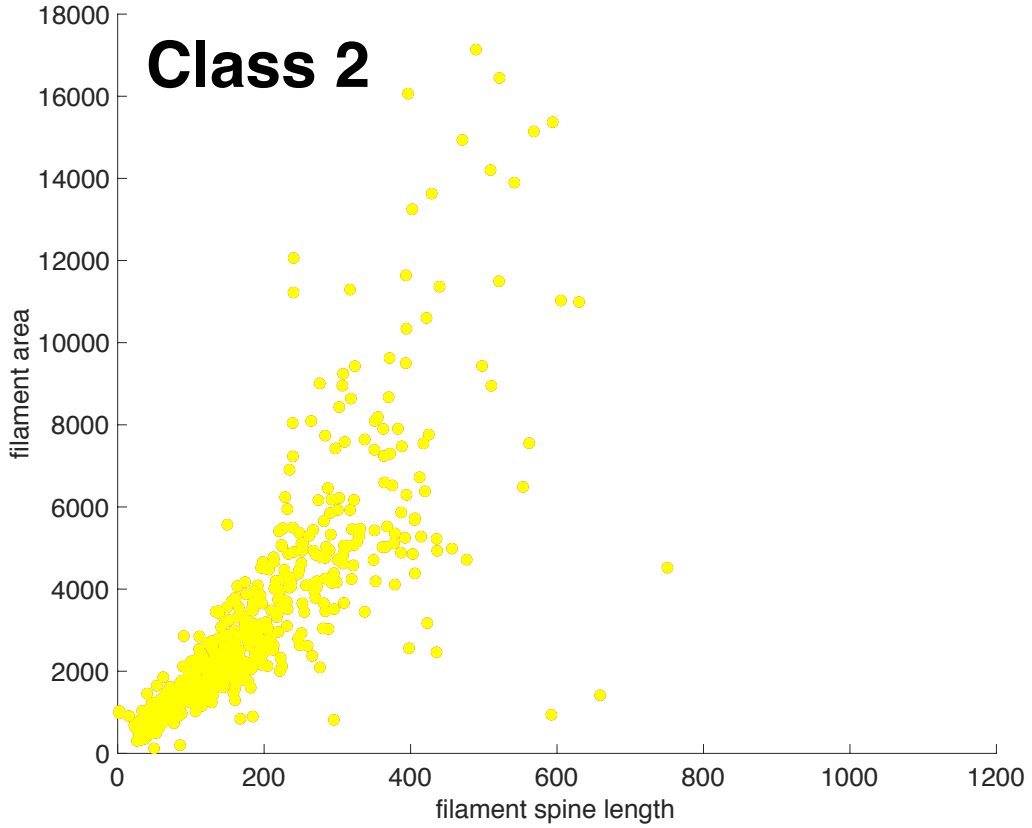
1. Probability distribution

1.3 Generative model

Testing Data



Quiescent filament



Non-quiescent filament

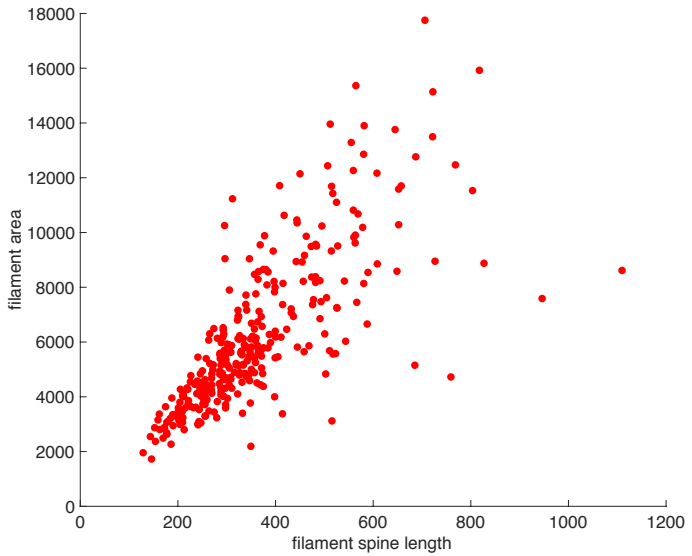
1. Probability distribution

1.3 Generative model

Prior

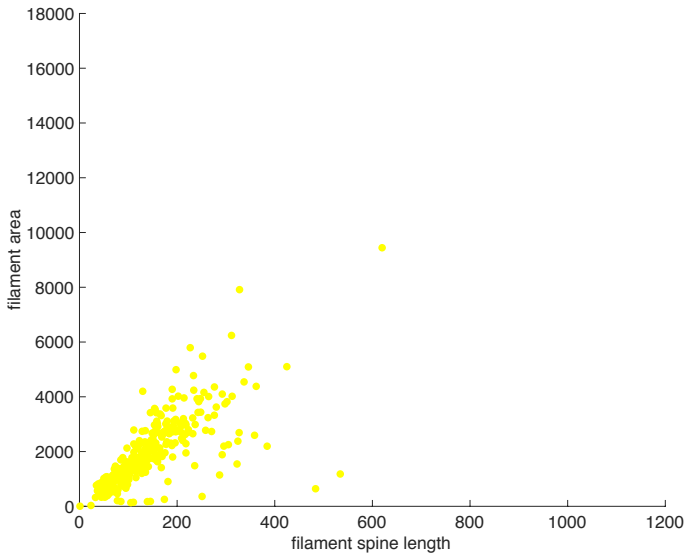
Class 1

$P(C_1)$



Class 2

$P(C_2)$



Information		Quiescent filament	Non-quiescent filament
Solar cycle 24	Training	324	387
Solar cycle 23	Testing	67	630

$$P(C_1) = \frac{324}{(324 + 387)} = 0.4557$$

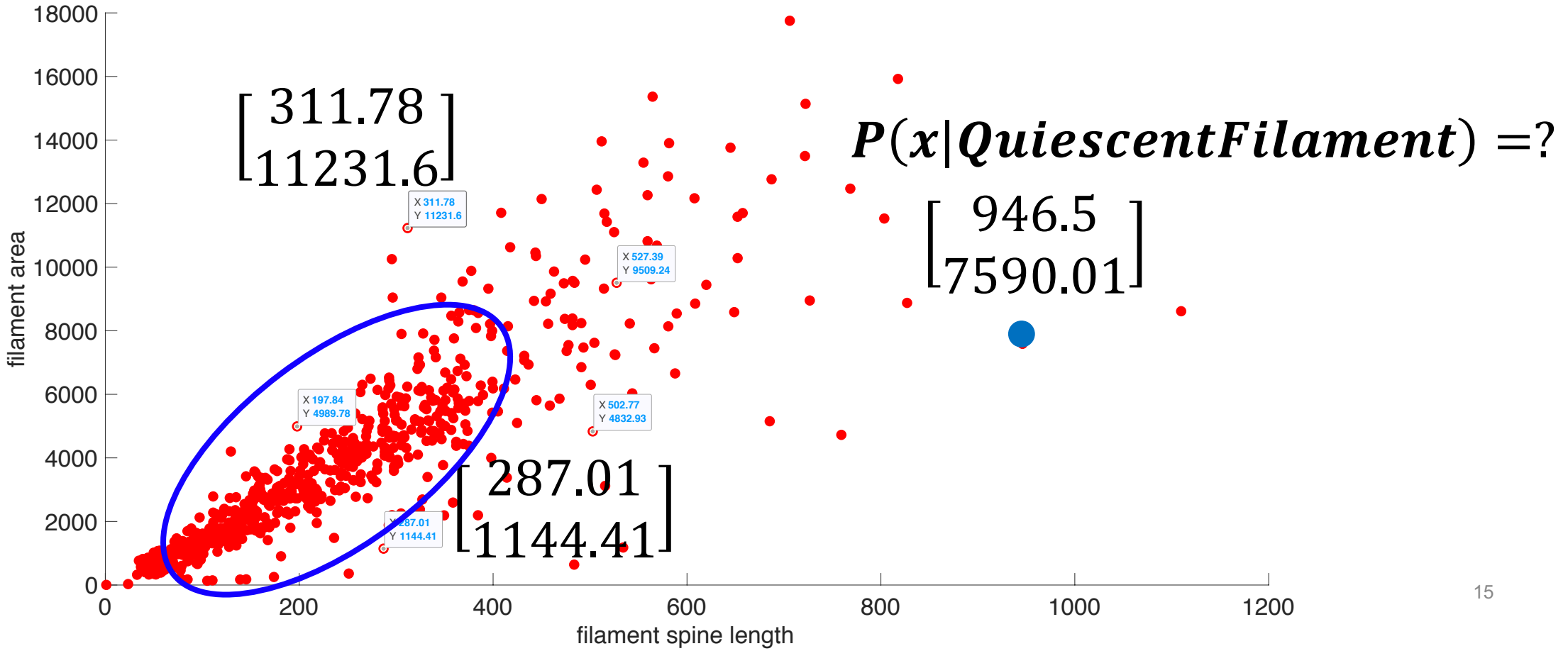
$$P(C_2) = \frac{387}{(324 + 387)} = 0.5443$$

1. Probability distribution

1.3 Generative model

Probability from class $P(x|C_1)$

Assume the points are sampled from a **Gaussian distribution**



1. Probability distribution

1.3 Generative model

Gaussian distribution

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

- Input: vector $\boldsymbol{x}$;
- Output: probability (density) of sampling $\boldsymbol{x}$;
- The shape of the function is determined by **mean $\boldsymbol{\mu}$** and **covariance matrix $\boldsymbol{\Sigma}$**

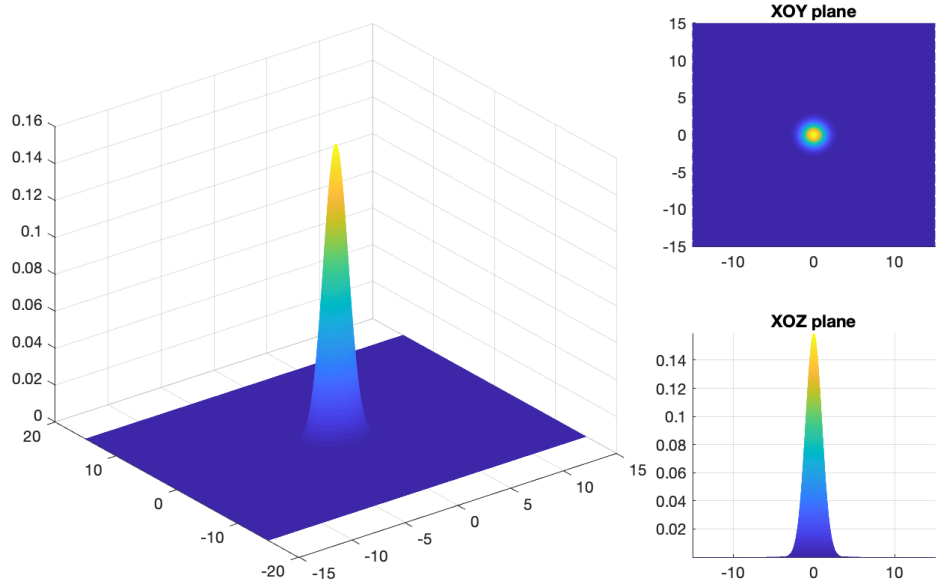
1. Probability distribution

1.3 Generative model

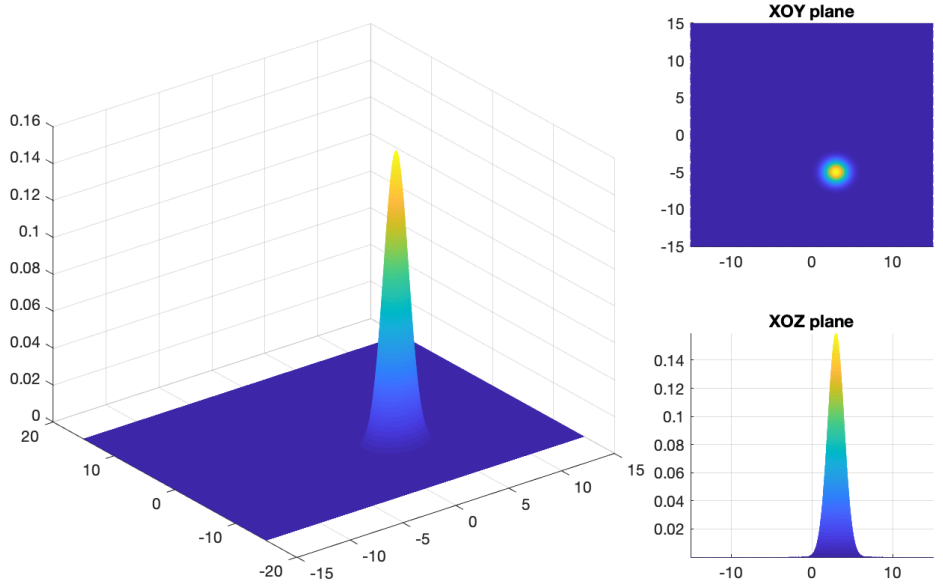
Gaussian distribution

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 3 \\ -5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



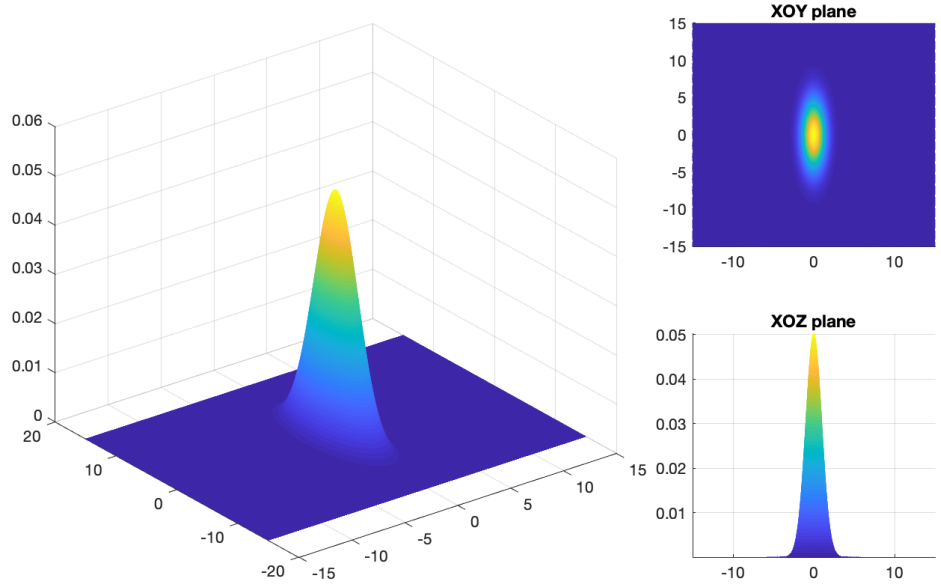
1. Probability distribution

1.3 Generative model

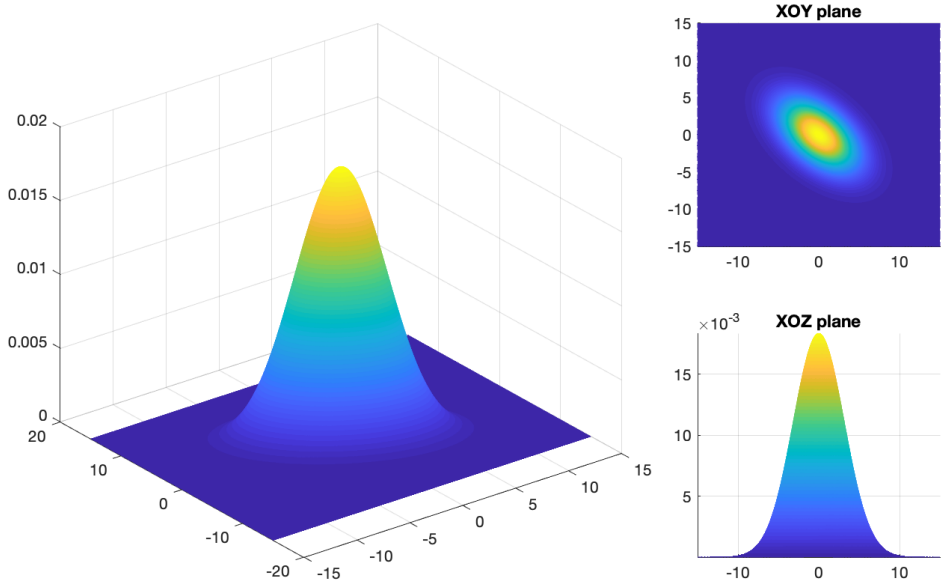
Gaussian distribution

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$$

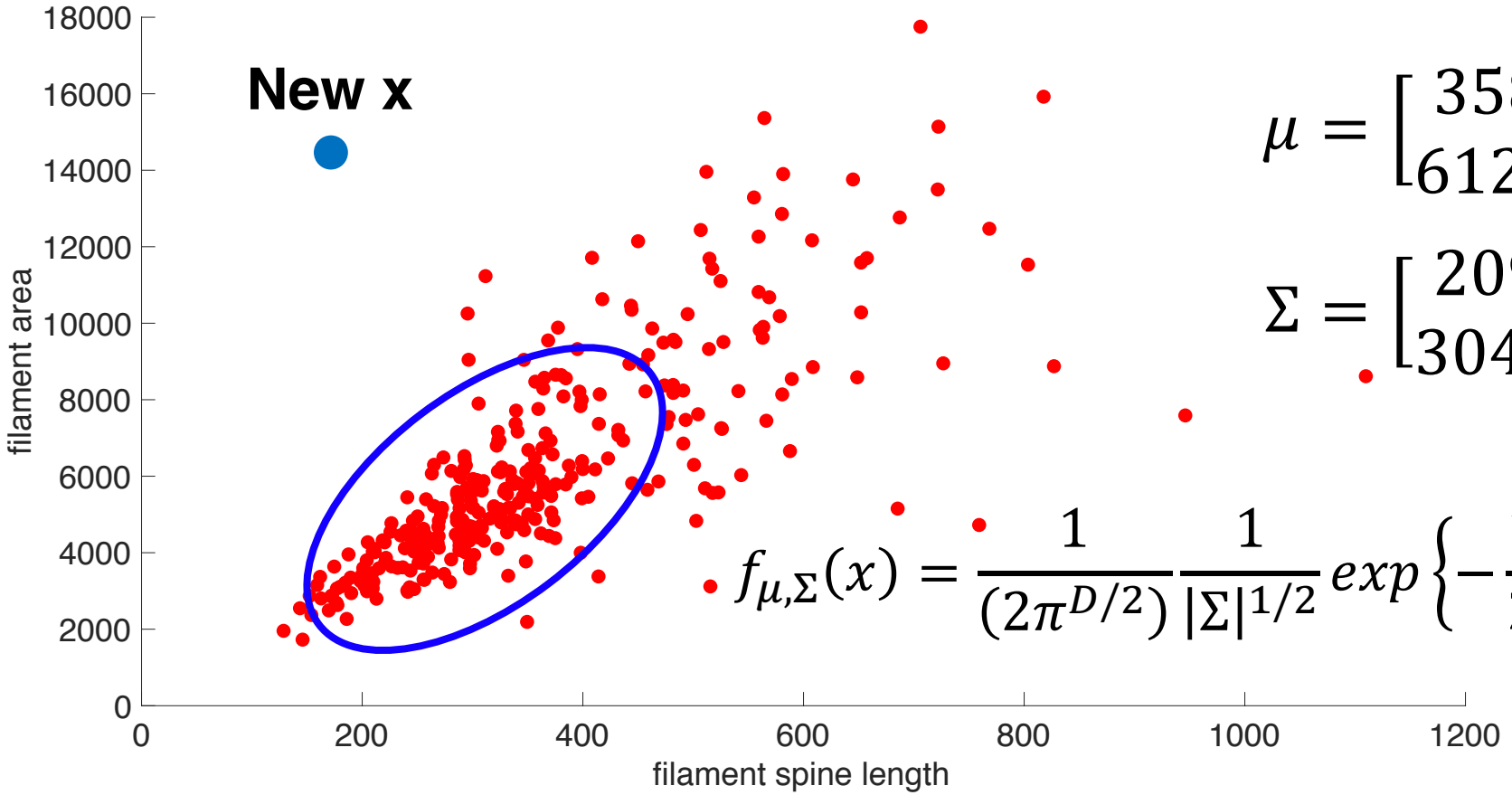


1. Probability distribution

1.3 Generative model

Probability from class

Assume the points are sampled from a **Gaussian distribution**



$$\mu = \begin{bmatrix} 358.9 \\ 6128.6 \end{bmatrix}$$

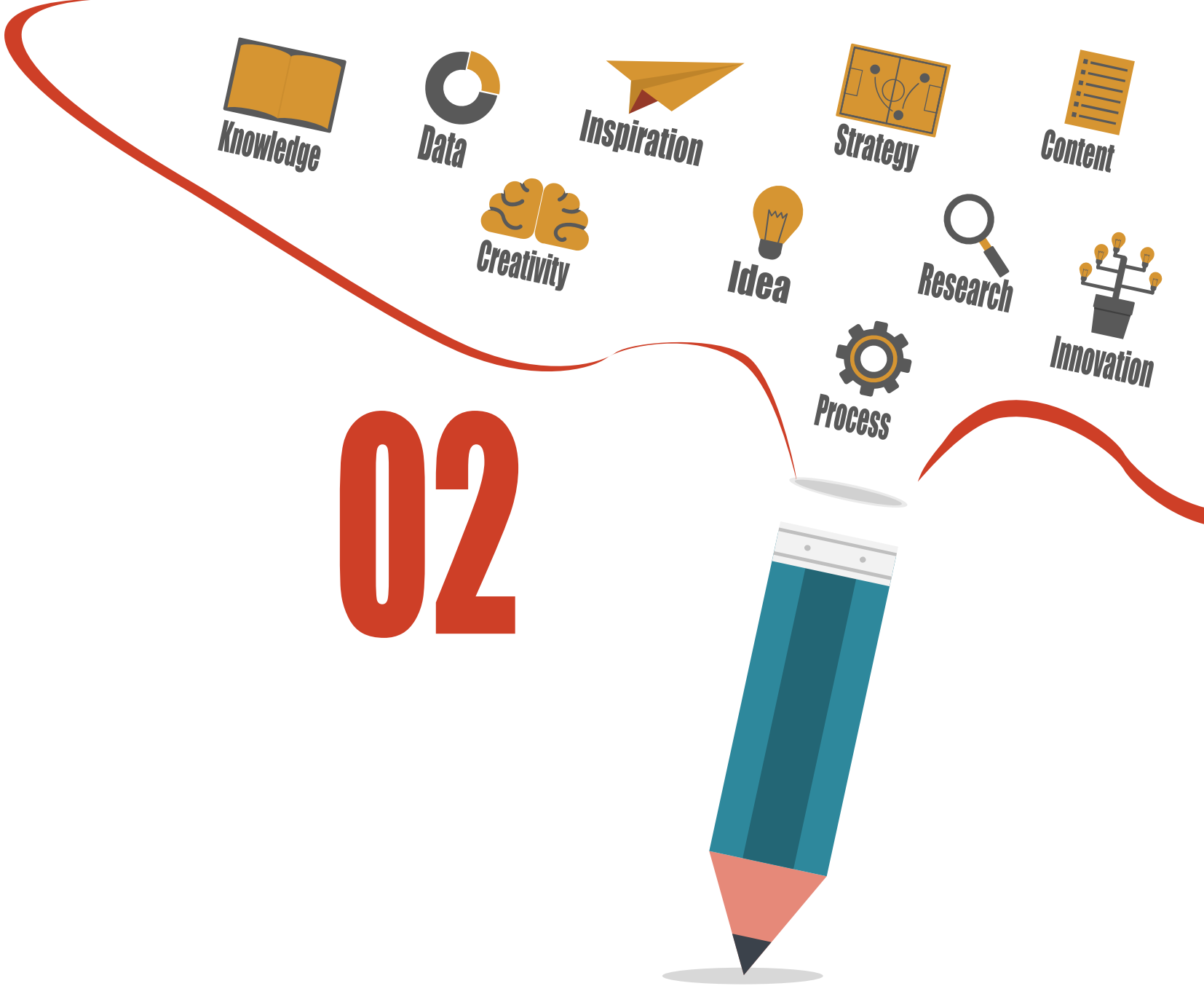
$$\Sigma = \begin{bmatrix} 20900 & 304700 \\ 304700 & 7745300 \end{bmatrix}$$

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

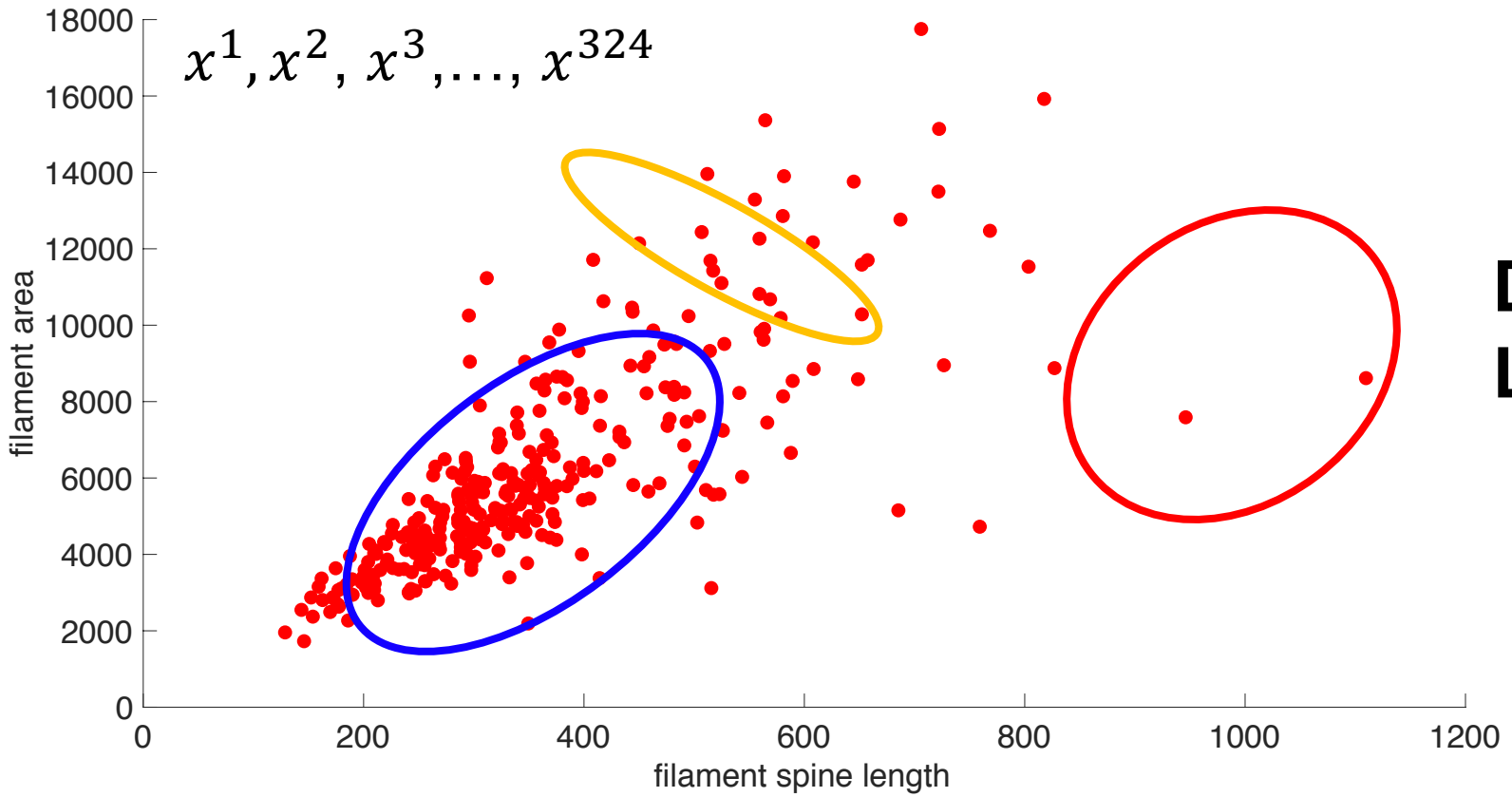
How to find the Gaussian distribution behind the data?

Maximum Likelihood

02



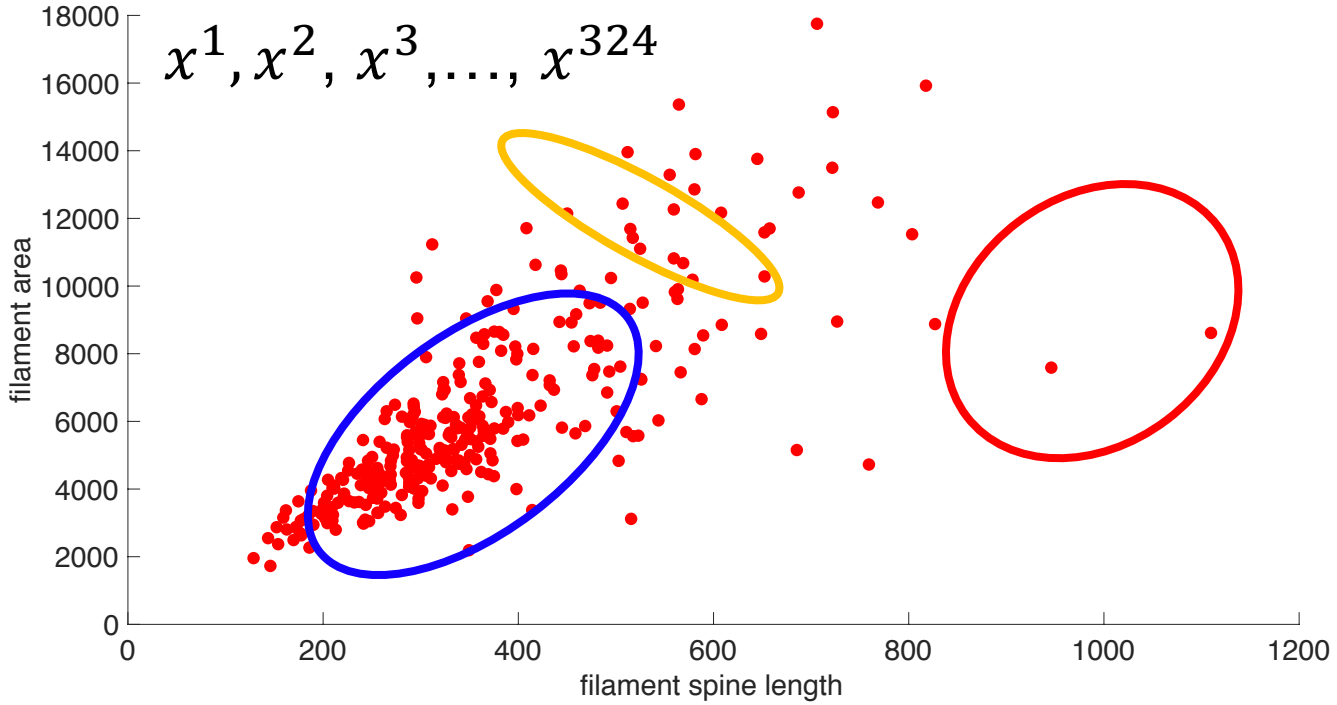
2. Maximum Likelihood



Different Likelihood

The Gaussian with any **mean μ** and **covariance matrix Σ** can generate these 324 points.

2. Maximum Likelihood



Likelihood of a Gaussian with mean μ and covariance matrix Σ = the probability of the Gaussian samples $x^1, x^2, x^3, \ldots\ldots, x^{324}$

$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x^1)\ f_{\mu,\Sigma}(x^3)\ f_{\mu,\Sigma}(x^3) \ldots\ldots\ f_{\mu,\Sigma}(x^{324})$$

2. Maximum Likelihood

We have the Quiescent filament samples: $x^1, x^2, x^3, \ldots\ldots, x^{324}$;

Assume these points are generate from a **Gaussian distribution** (μ^*, Σ^*)

with the **maximum likelihood :**

$$L(\mu, \Sigma) = f_{\mu,\Sigma}(x^1)\ f_{\mu,\Sigma}(x^2)\ f_{\mu,\Sigma}(x^3)\ldots\ldots\ f_{\mu,\Sigma}(x^{324})$$

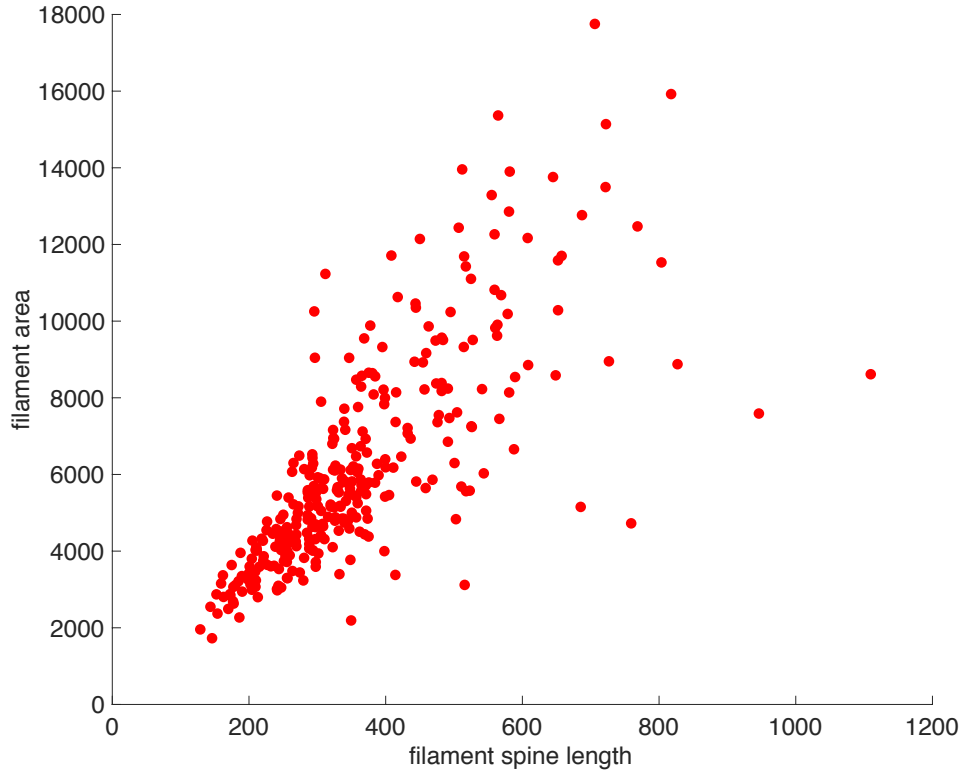
$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi^{D/2})}\frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\mu^*, \Sigma^* = \arg\max_{\mu,\Sigma} L(\mu, \Sigma)$$

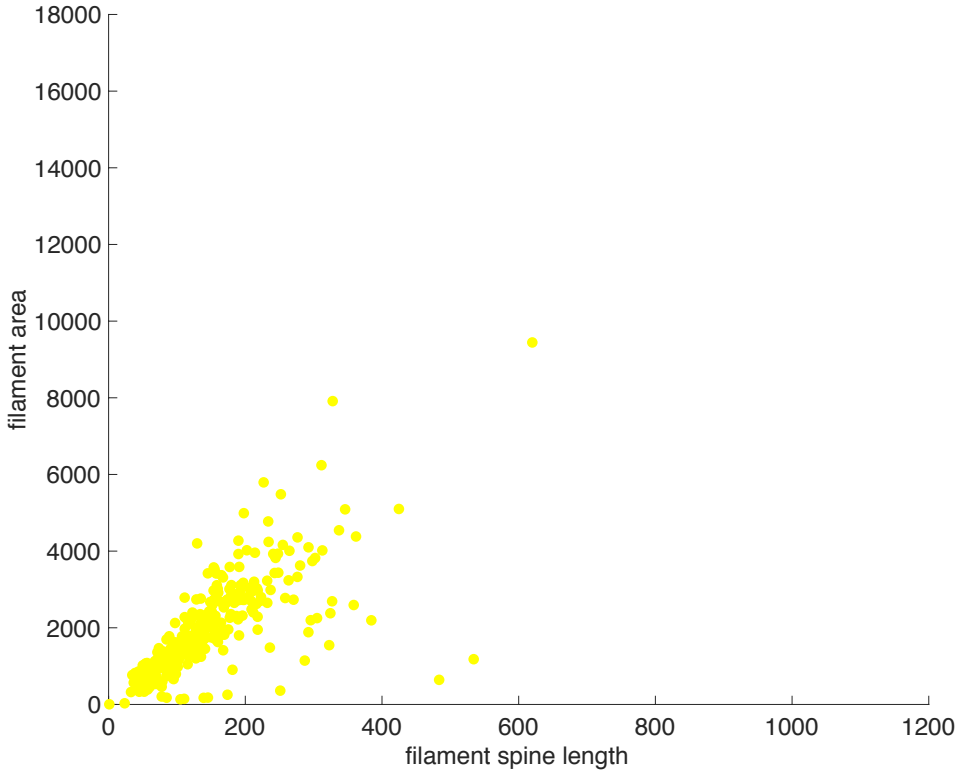
$$\mu^* = \frac{1}{324}\sum_{i=1}^{324} x^i \qquad \Sigma^* = \frac{1}{324}\sum_{i=1}^{324}\left(x^i - \mu^*\right)\left(x^i - \mu^*\right)^T$$

2. Maximum Likelihood

Class 1: Quiescent filament



Class 2: Non-quiescent filament



$$\mu^1 = \begin{bmatrix} 358.9 \\ 6128.6 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 20900 & 304700 \\ 304700 & 7745300 \end{bmatrix} \quad \mu^2 = \begin{bmatrix} 129 \\ 1716.9 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 7100 & 81100 \\ 81100 & 1558500 \end{bmatrix}$$

2. Maximum Likelihood

$$f_{\mu^1,\Sigma^1}(x) = \frac{1}{(2\pi^{D/2})}\frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}$$

$$\mu^1 = \begin{bmatrix} 358.9 \\ 6128.6 \end{bmatrix} \qquad \Sigma^1 = \begin{bmatrix} 20900 & 304700 \\ 304700 & 7745300 \end{bmatrix}$$

$$P(C_1) = \frac{324}{(324+387)} = 0.4557$$

$$\boldsymbol{P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}}$$

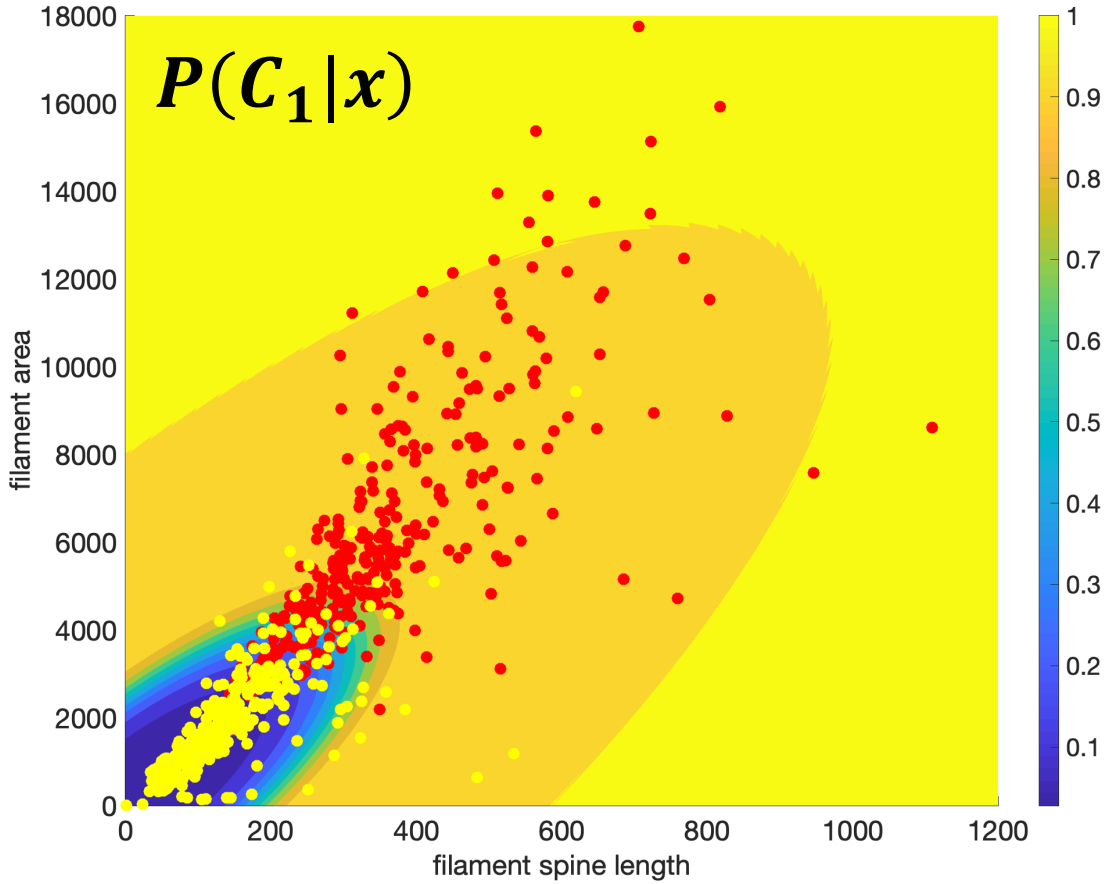
$$f_{\mu^2,\Sigma^2}(x) = \frac{1}{(2\pi^{D/2})}\frac{1}{|\Sigma^2|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

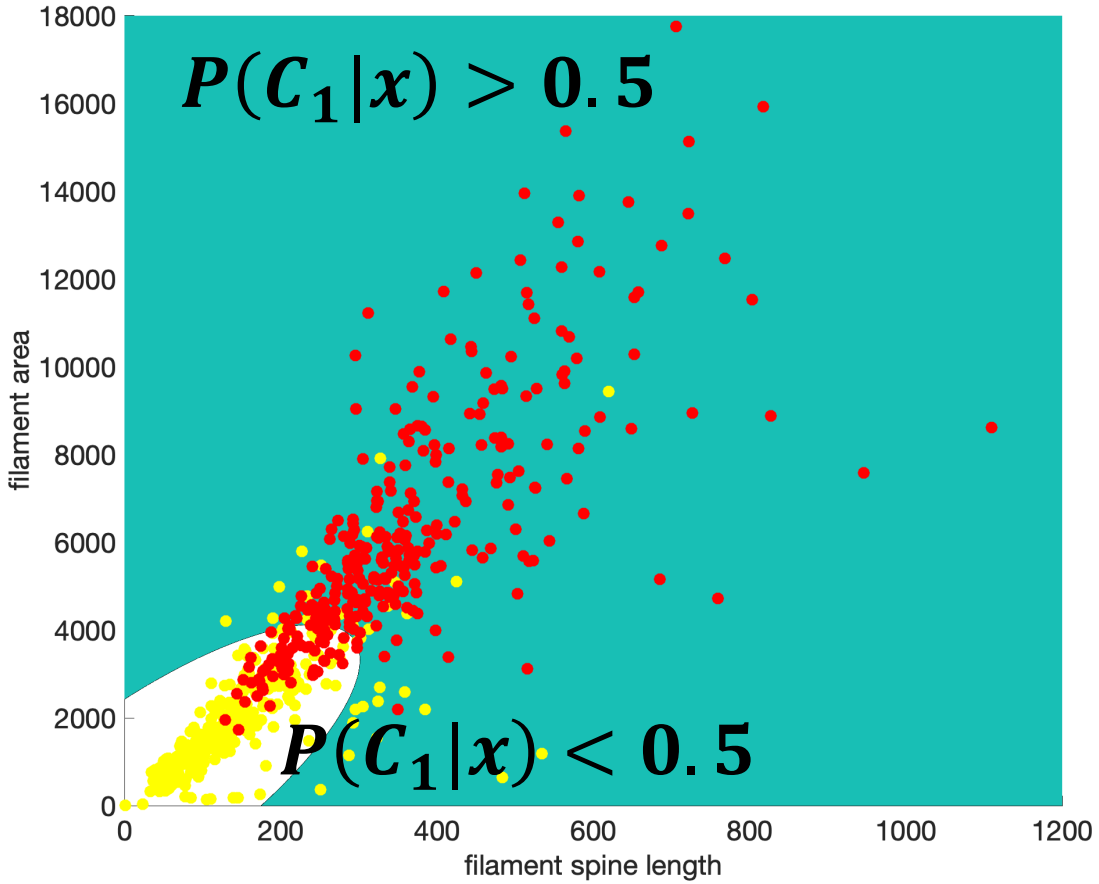
$$\mu^2 = \begin{bmatrix} 129 \\ 1716.9 \end{bmatrix} \qquad \Sigma^2 = \begin{bmatrix} 7100 & 81100 \\ 81100 & 1558500 \end{bmatrix}$$

$$P(C_2) = \frac{387}{(324+387)} = 0.5443$$

If $\boldsymbol{P(C_1|x)} > 0.5$, then $\boldsymbol{x}$ belongs to class 1 (quiescent filament.)

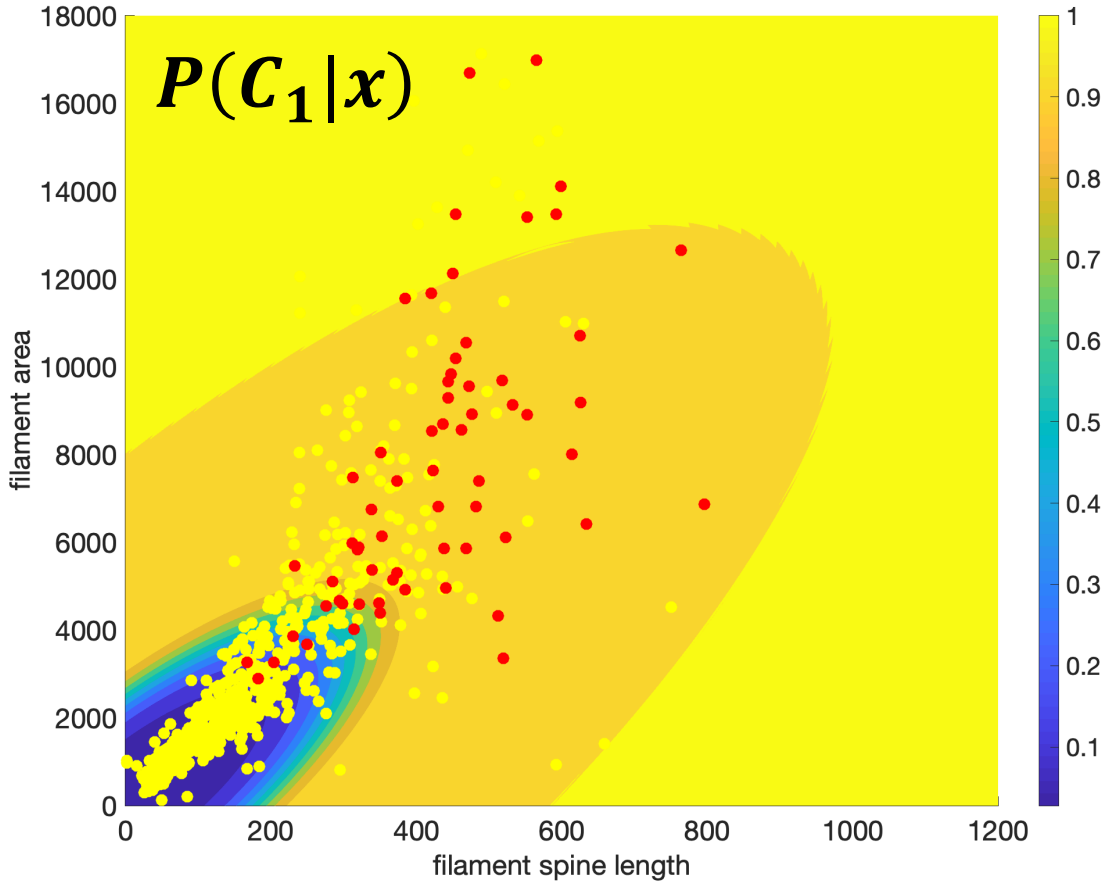
2. Maximum Likelihood

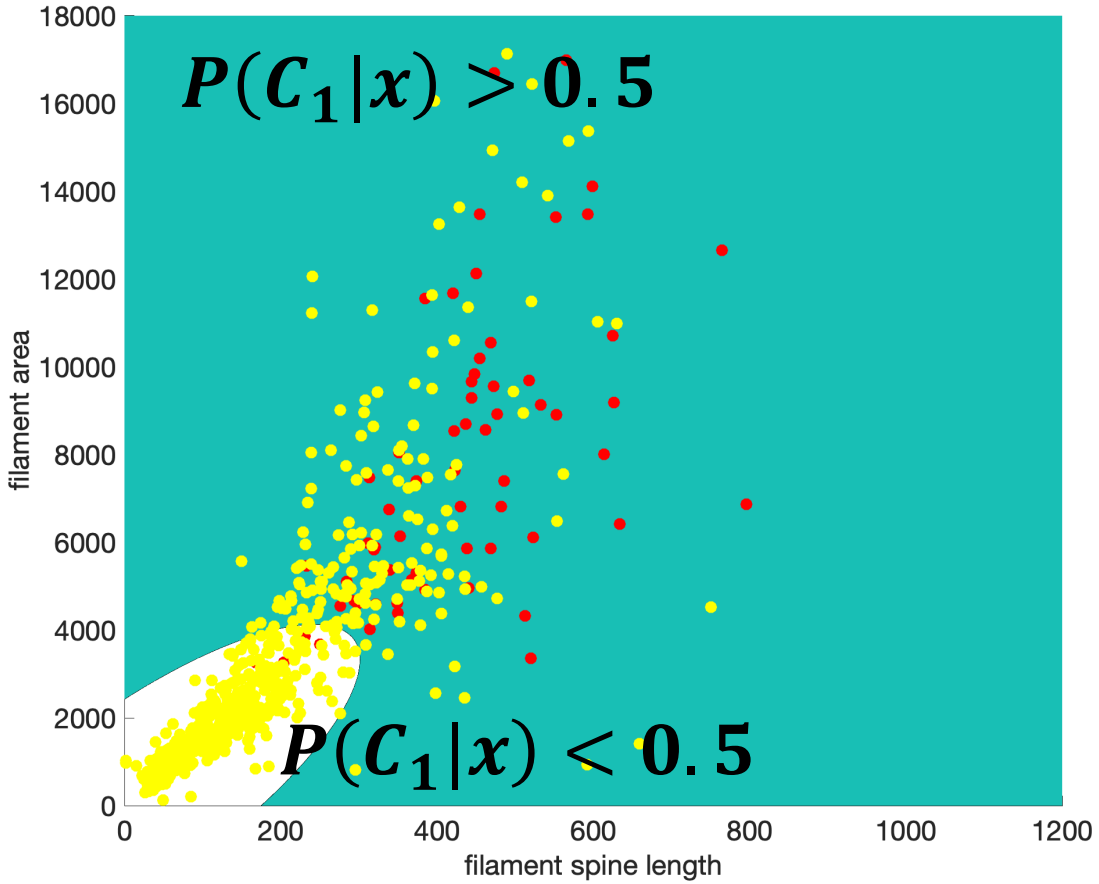




How about the results? **Training: 86.64% accuracy**

2. Maximum Likelihood





How about the results? **Testing: 82.21% accuracy**

2. Maximum Likelihood

Date	Latitude	Longitude	Area	Spine length	Barb Number	Tilt angle	Quiescent filament
2011-07-01	83.42	338.75	5160.42	319.65	11	-28.47	1
2011-11-16	221.32	-611.63	1159.69	81.15	3	23.29	0
2014-12-29	-431.07	-362.29	2762.41	134.87	6	11.12	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

6 features;

μ^1, μ^2: 6-dimension vector;

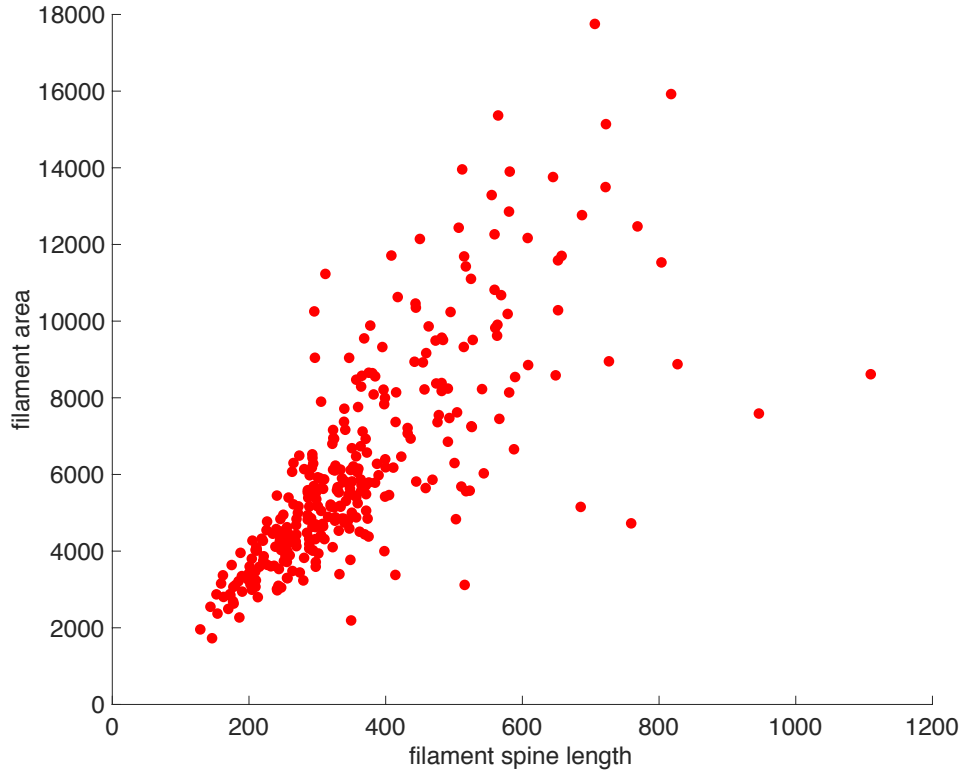
Σ^1, Σ^2: 6×6 matrices

How about the results?

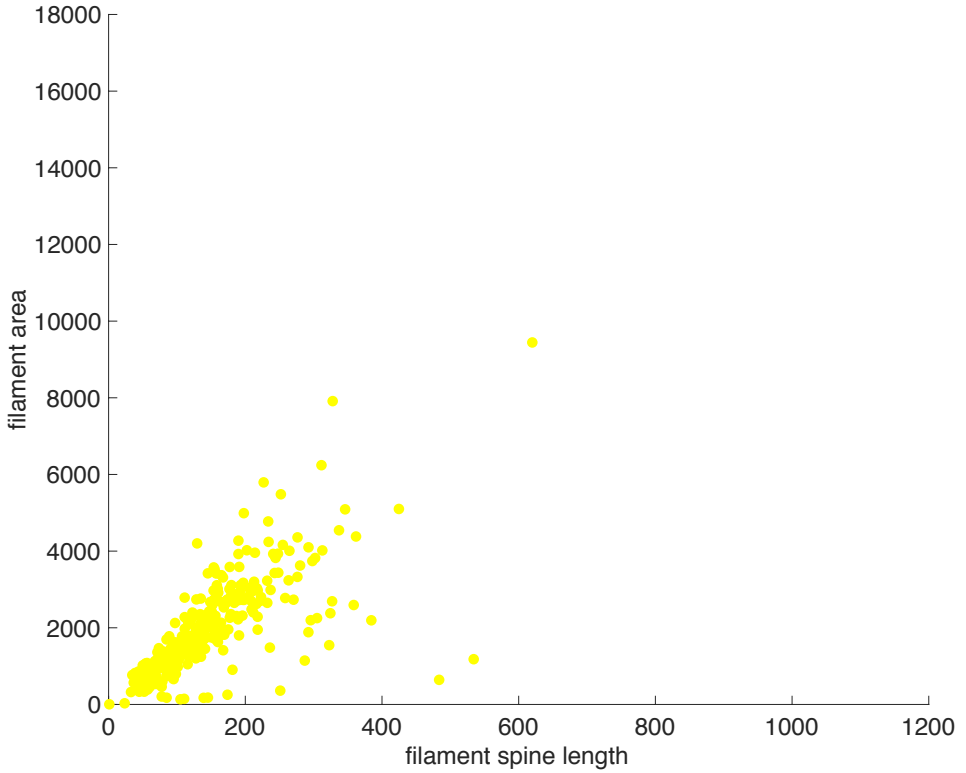
Training: 98.45% accuracy
Testing: 89.53% accuracy

2. Maximum Likelihood

Class 1: Quiescent filament



Class 2: Non-quiescent filament



$$\mu^1 = \begin{bmatrix} 358.9 \\ 6128.6 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 20900 & 304700 \\ 304700 & 7745300 \end{bmatrix}$$

$$\mu^2 = \begin{bmatrix} 129 \\ 1716.9 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 7100 & 81100 \\ 81100 & 1558500 \end{bmatrix}$$

2. Maximum Likelihood

If we use the same Σ, less parameters.

How about the results?

2. Maximum Likelihood

Quiescent filament samples:

$x^1, x^2, x^3, \ldots\ldots, x^{324}$

Non-quiescent filament samples:

$x^{325}, x^{326}, x^{327}, \ldots\ldots, x^{711}$

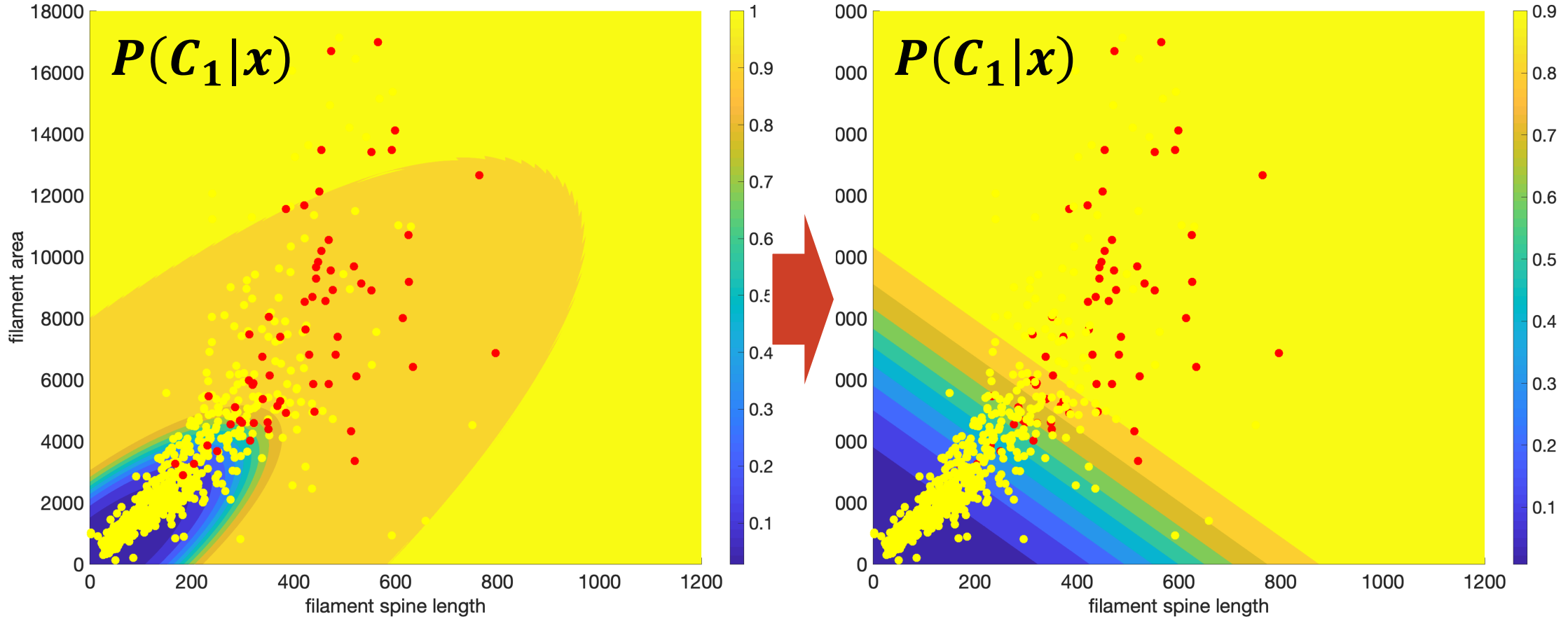
Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1, \mu^2, \Sigma)$

$$L(\mu^1, \mu^2, \Sigma) = f_{\mu^1,\Sigma}(x^1)\, f_{\mu^1,\Sigma}(x^3)\, f_{\mu^1,\Sigma}(x^3) \ldots\ldots\, f_{\mu^1,\Sigma}(x^{324}) \times$$
$$f_{\mu^2,\Sigma}(x^{325})\, f_{\mu^2,\Sigma}(x^{326})\, f_{\mu^2,\Sigma}(x^{327}) \ldots\ldots\, f_{\mu^2,\Sigma}(x^{711})$$

μ^1, μ^2 is the same : $\mu^1 = \begin{bmatrix} 358.9 \\ 6128.6 \end{bmatrix}$ $\mu^2 = \begin{bmatrix} 129 \\ 1716.9 \end{bmatrix}$

$$\Sigma = \frac{324}{711}\Sigma^1 + \frac{387}{711}\Sigma^2$$

2. Maximum Likelihood



$$\mu^1 = \begin{bmatrix} 358.9 \\ 6128.6 \end{bmatrix} \quad \mu^2 = \begin{bmatrix} 129 \\ 1716.9 \end{bmatrix} \quad \Sigma = \frac{324}{711}\Sigma^1 + \frac{387}{711}\Sigma^2 = \begin{bmatrix} 13400 & 183000 \\ 183000 & 4377800 \end{bmatrix}$$

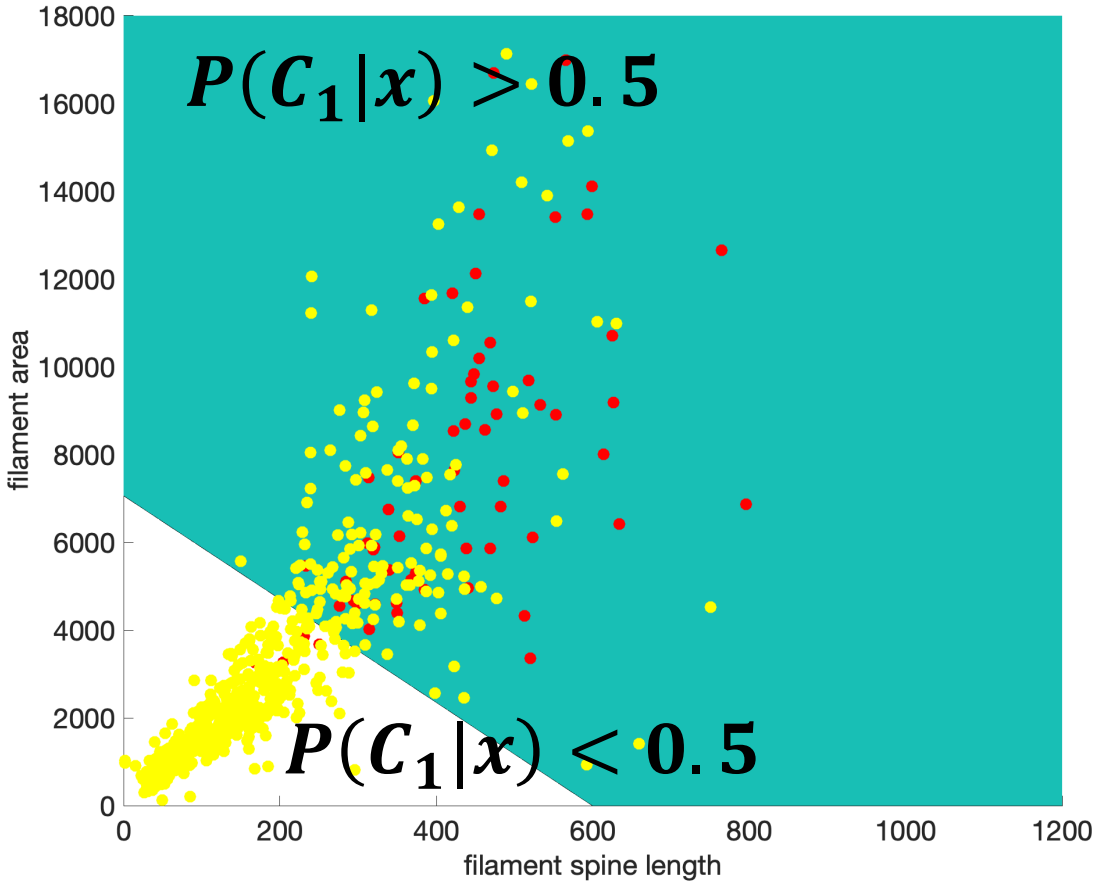
2. Maximum Likelihood

How about the results?

- Training: 88.05% accuracy
- Training: 86.64% accuracy
- Testing: 87.52% accuracy
- Testing: 82.21% accuracy

Consider all 6 features:

- Training: 100% accuracy
- Training: 98.45% accuracy
- Testing: 100% accuracy
- Testing: 89.53% accuracy



2. Maximum Likelihood

Step 1
define a set of function

Step 2
goodness of function

Step 3
pick the best function

$$\boldsymbol{X} \rightarrow f(x) = \begin{cases} \boldsymbol{P(C_1|x) = \dfrac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}} \\ \text{If } \boldsymbol{P(C_1|x) > 0.5} \quad \text{Output} = \text{class 1} \\ else \quad \text{Output} = \text{class 2} \end{cases}$$

Likelihood: $\mu^*, \Sigma^* = \arg\max_{\mu,\Sigma} L(\mu, \Sigma)$

The mean μ^* and covariance Σ^* that maximizing the likelihood
e.g., the probability of generating data

Generative model: quite easy !

Bayes classifier

03

3. Bayes classifier

3.1 Bayes classifier

Can I use different probability distribution?

Yes. You can always use the distribution which you like.

$$\boldsymbol{X} = [x_1, x_2, x_3, \ldots\ldots, x_n]^T$$

$$\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{C_1}) = \boldsymbol{P}(x_1|\boldsymbol{C_1})\boldsymbol{P}(x_2|\boldsymbol{C_1})\boldsymbol{P}(x_3|\boldsymbol{C_1}) \cdots\cdots \boldsymbol{P}(x_n|\boldsymbol{C_1})$$

For multi-features: you may assume they are from **1D Gaussian distributions**.

For binary features: you may assume they are from **Bernoulli distributions**.

If you assume all the dimensions (features) are independent, then you are using **Naïve Bayes Classifier**.

3. Bayes classifier

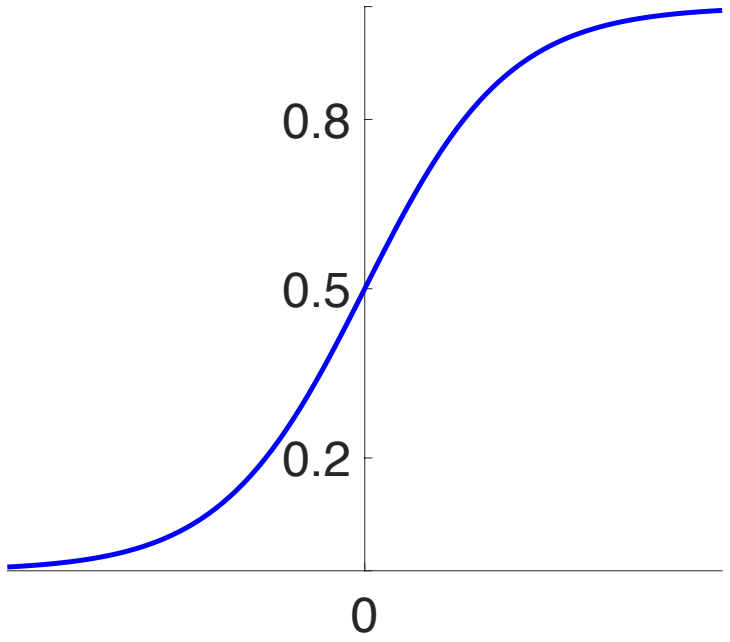
3.2 Posterior probability

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1}{1+\frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1+exp(-z)} = \sigma(z)$$

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

Sigmoid function



3. Bayes classifier

3.2 Posterior probability

$$P(C_1|x) = \sigma(z) \qquad \text{Sigmoid function} \qquad z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{P(C_1)}{P(C_2)} \quad \frac{P(C_1)}{P(C_2)} = \frac{\frac{N_1}{N_1+N_2}}{\frac{N_2}{N_1+N_2}} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}$$

$$P(x|C_2) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma^2|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

3. Bayes classifier

3.2 Posterior probability

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\} \quad P(x|C_2) = \frac{1}{(2\pi^{D/2})} \frac{1}{|\Sigma^2|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}$$

$$\ln \frac{\cancel{\frac{1}{(2\pi^{D/2})}} \frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)\right\}}{\cancel{\frac{1}{(2\pi^{D/2})}} \frac{1}{|\Sigma^2|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right\}}$$

$$= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)]\right\}$$

$$= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)]$$

3. Bayes classifier

3.2 Posterior probability

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{N_1}{N_2}$$

$$\ln \frac{P(x|C_1)}{P(x|C_2)} = \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}\left[(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1) - (x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2)\right]$$

$$(x-\mu^1)^T(\Sigma^1)^{-1}(x-\mu^1)$$

$$= x^T(\Sigma^1)^{-1}x - x^T(\Sigma^1)^{-1}\mu^1 - (\mu^1)^T(\Sigma^1)^{-1}\,x + (\mu^1)^T(\Sigma^1)^{-1}\mu^1$$

$$= x^T(\Sigma^1)^{-1}x - 2(\mu^1)^T(\Sigma^1)^{-1}x + (\mu^1)^T(\Sigma^1)^{-1}\mu^1$$

$$(x-\mu^2)^T(\Sigma^2)^{-1}(x-\mu^2) = x^T(\Sigma^2)^{-1}x - 2(\mu^2)^T(\Sigma^2)^{-1}x + (\mu^2)^T(\Sigma^2)^{-1}\mu^2$$

$$z = \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}x^T(\Sigma^1)^{-1}x + (\mu^1)^T(\Sigma^1)^{-1}x - \frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1 +$$
$$\frac{1}{2}x^T(\Sigma^2)^{-1}x - (\mu^2)^T(\Sigma^2)^{-1}x + \frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2 + \ln \frac{N_1}{N_2}$$

3. Bayes classifier

3.2 Posterior probability

$$P(C_1|x) = \sigma(z)$$

$$z = \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2} x^T (\Sigma^1)^{-1} x + (\mu^1)^T (\Sigma^1)^{-1} x - \frac{1}{2} (\mu^1)^T (\Sigma^1)^{-1} \mu^1 +$$
$$\frac{1}{2} x^T (\Sigma^2)^{-1} x - (\mu^2)^T (\Sigma^2)^{-1} x + \frac{1}{2} (\mu^2)^T (\Sigma^2)^{-1} \mu^2 + \ln \frac{N_1}{N_2}$$

$$\Sigma^1 = \Sigma^2 = \Sigma$$

$$z = (\mu^1 - \mu^2)^T (\Sigma^1)^{-1} x - \frac{1}{2} (\mu^1)^T (\Sigma^1)^{-1} \mu^1 + \frac{1}{2} (\mu^2)^T (\Sigma^2)^{-1} \mu^2 + \ln \frac{N_1}{N_2}$$

W^T $\quad$ b

$$P(C_1|x) = \sigma(wx + b)$$

In generative model, we estimate N_1 , N_2, μ^1, μ^2, Σ, then determine w and b.

Generative Models

Hao, Qi
School of Astronomy and Space Science

THANKS