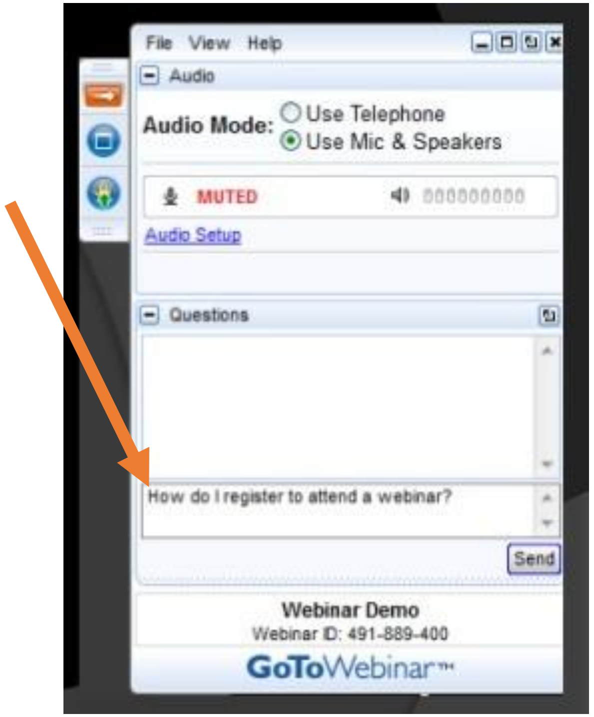
April 28, 2021 | The webinar will begin momentarily

How to Participate

Webinar Logistics



Audio

- If you have trouble connecting to audio with your computer, or if you prefer to use your phone, select "Use Telephone" after joining the webinar and call in using:

  +1 (415) 655-0052
  Access Code: 397-999-675
  Webinar ID: 925-106-267

Polls & Questions

- Please submit your questions using the Questions tab on your screen
- A recording and slide deck will be available after the webinar is complete.
- In the event this webinar shuts off due to connection issues, do NOT close the webinar the window. The webinar room will reopen in just a couple of minutes.

Strengthening University-Industry Partnerships

Strengthening University-Industry Partnerships

Strengthening University-Industry Partnerships

Disclaimer

UIDP materials, which include publications, webinars, videos, and presentations, reflect an amalgamation of the experiences and knowledge of those who participate in UIDP activities. The views and opinions expressed in UIDP materials do not necessarily reflect the official policy or position of any individual organization or the UIDP. At no time should any UIDP materials be used as a replacement for an individual organization's policy, procedures, or legal counsel. UIDP is not a lobbying organization and UIDP materials are not intended to be used to influence government decisions.

UIDP Webinar Presentation

Machine Learning: What It Is and What Its Applications Are

Yiwen Xu, Ph.D.

(yiwen.xu@ndsu.edu)

Assistant Professor

Department of Industrial and Manufacturing Engineering,

North Dakota State University

April 28, 2021

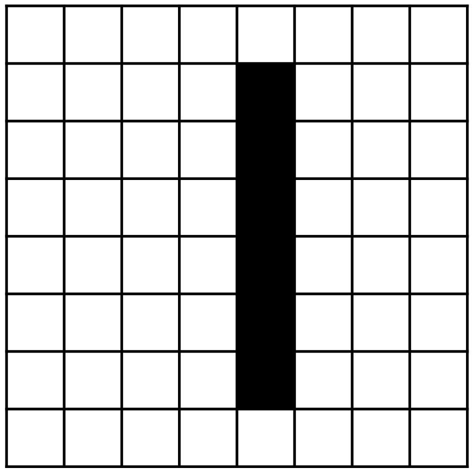
Strengthening University-Industry Partnerships

North Dakota State University

A Big Picture of Machine Learning

- What is machine learning?

  Machine learning (ML) uses interdisciplinary techniques such as optimization, statistics, and linear algebra to create systems that can sift through large volumes of data at high speeds to make predictions or decisions. ML algorithms directly build models based on sample data (i.e., the "training data") to make automated predictions/decisions without being explicitly programed to do so, as opposed to traditional algorithms.



A handwritten "1" in an 8×8 pixeled cell

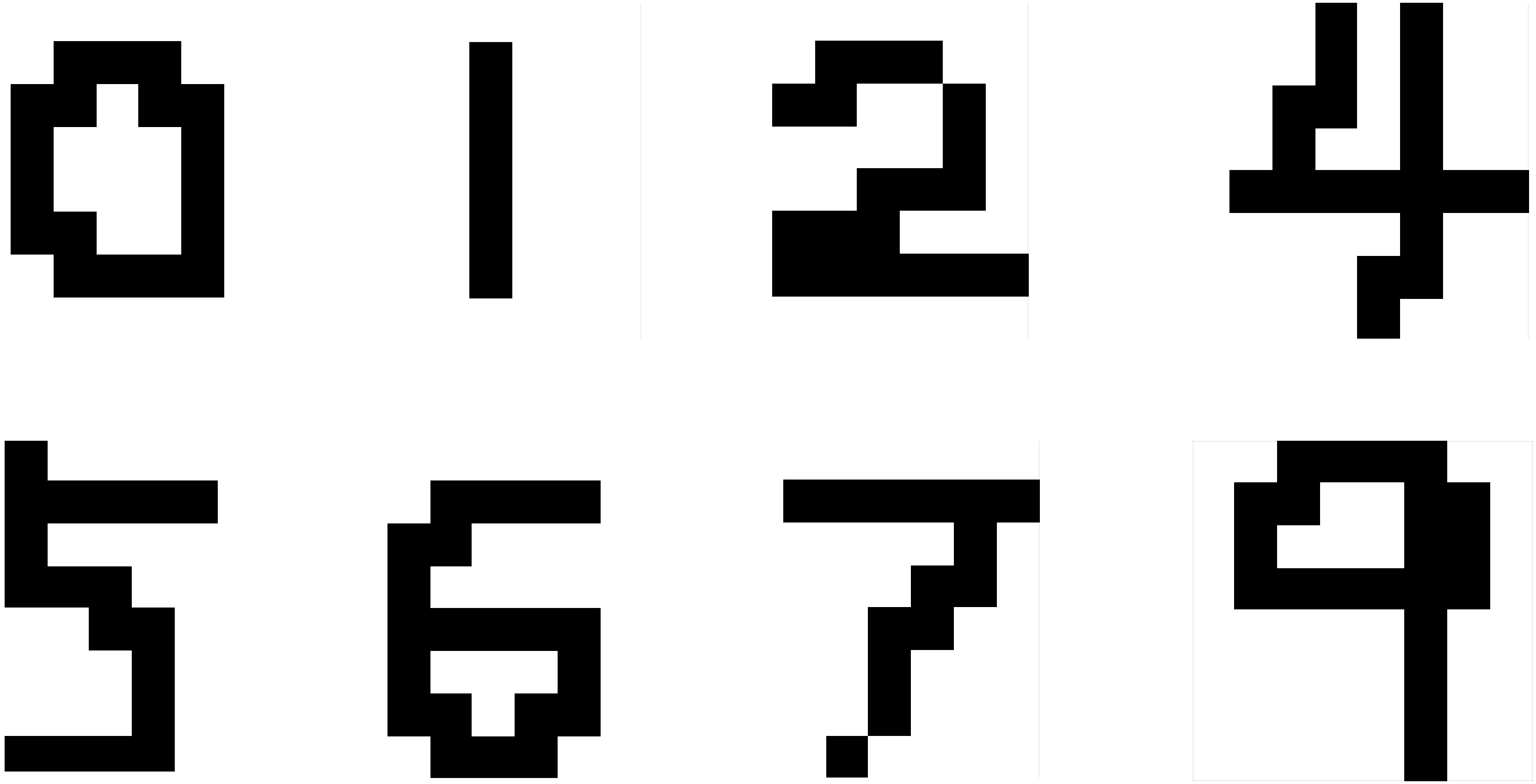
	Traditional Algorithm	**ML Algorithm**
In general	If … then …	Automatically learn the rules/patterns from the training data
For the handwritten digit recognition example	If only exists one column such that at least 5 consecutive vertical pixels are dark, then it is "1".	Training 1000 practical handwritten "1"s and find the "pattern" and recognize a new given "1" automatically

Some Applications of Machine Learning Algorithms

- Application 1 (Handwritten Recognition)

  What are the following numbers? We expect the computer to recognize them automatically.

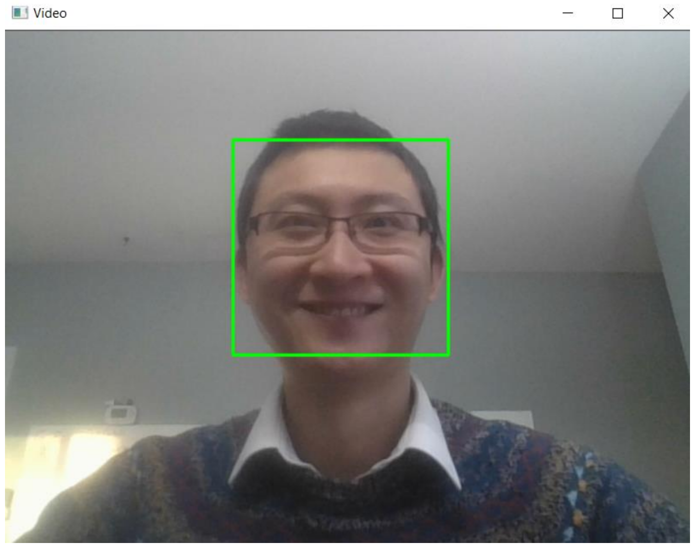
  - We will revisit this example later.

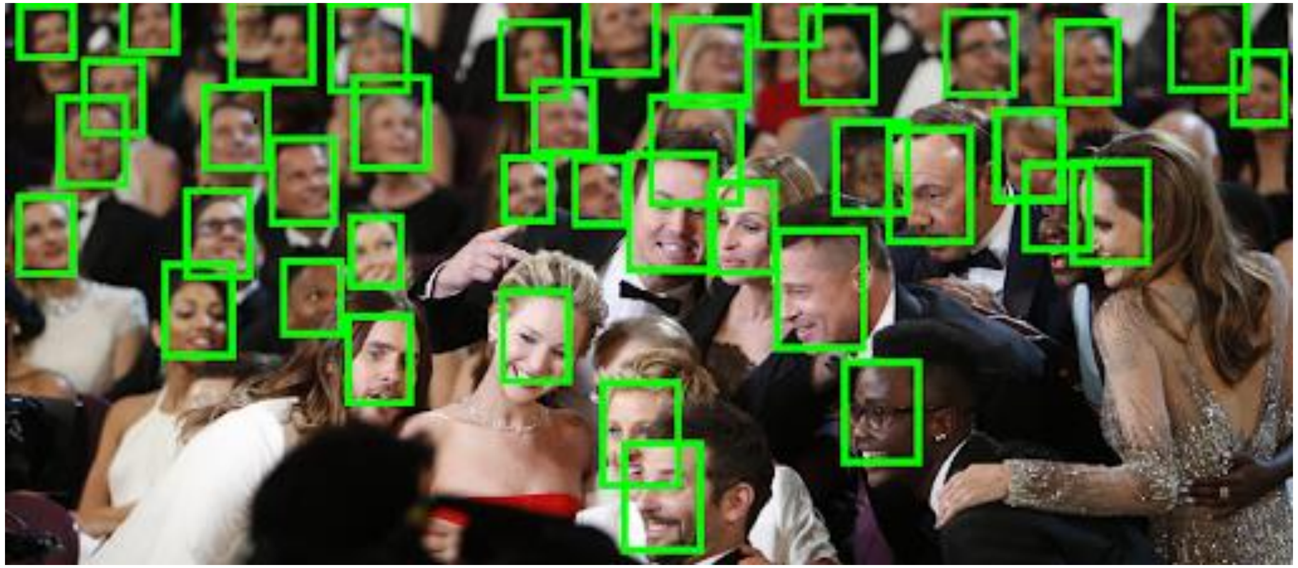


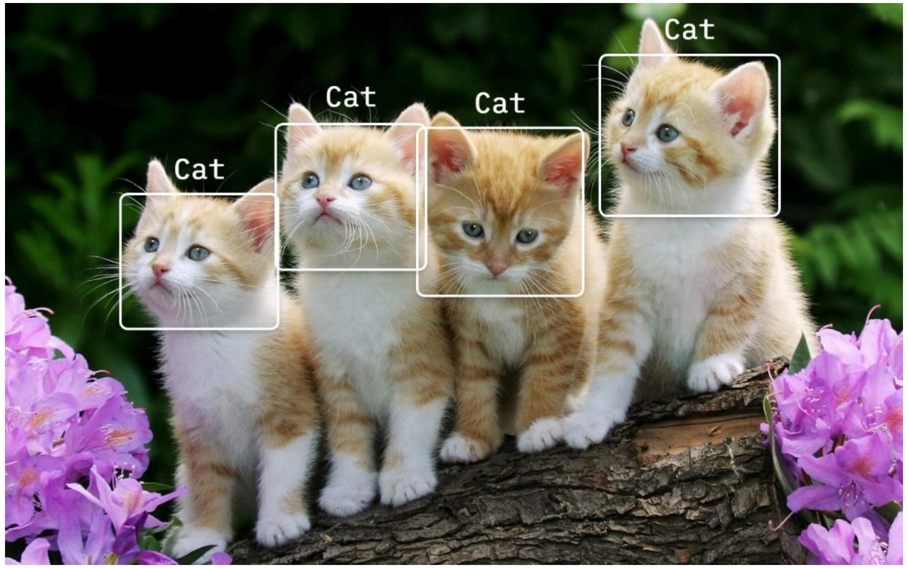
Some Applications of Machine Learning Algorithms

- Application 2 (Face Recognition Recognition)

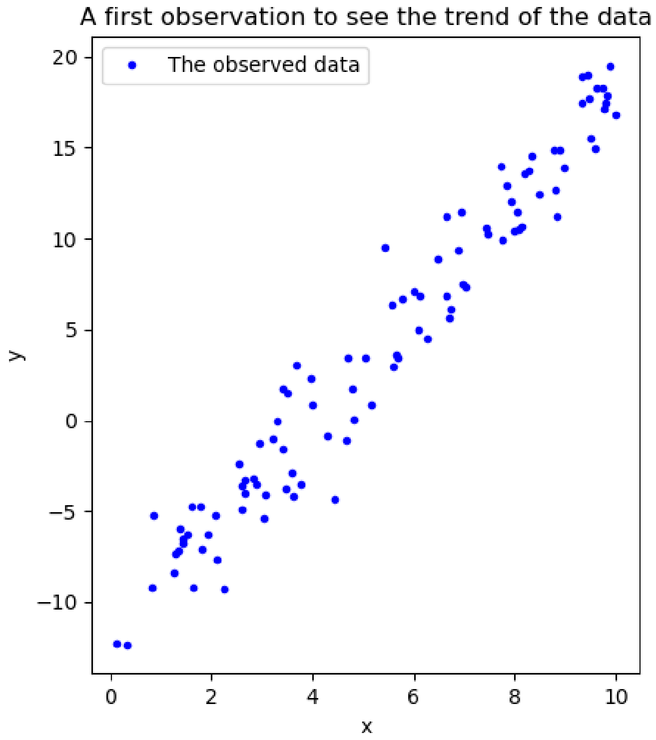
  For a given picture, find the (human) faces inside.
  - We can do a real-time face recognition
  - Not necessarily human faces. E.g., cat face :-)

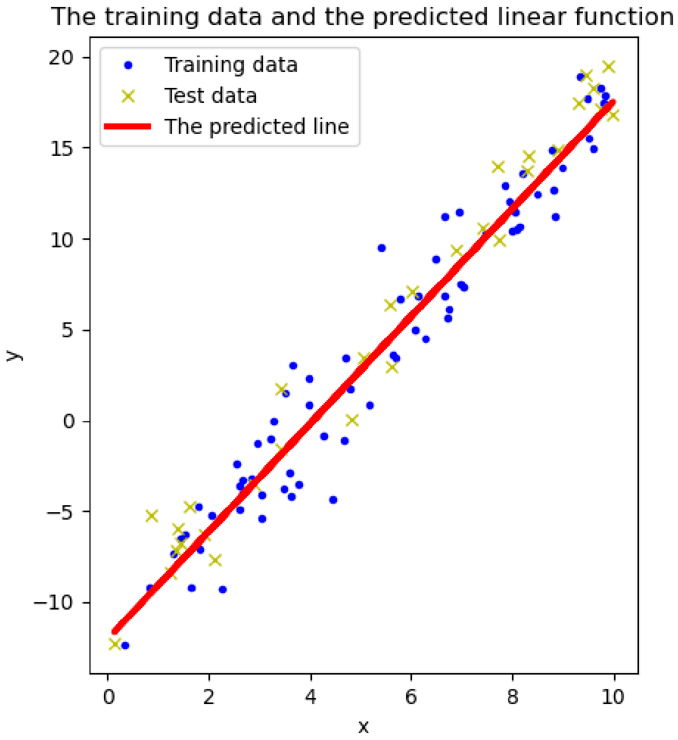






Some Applications of Machine Learning Algorithms





- Application 3 (Data Trend Prediction)

  For given historical data, predict its future trend.

  - We will soon revisit this example

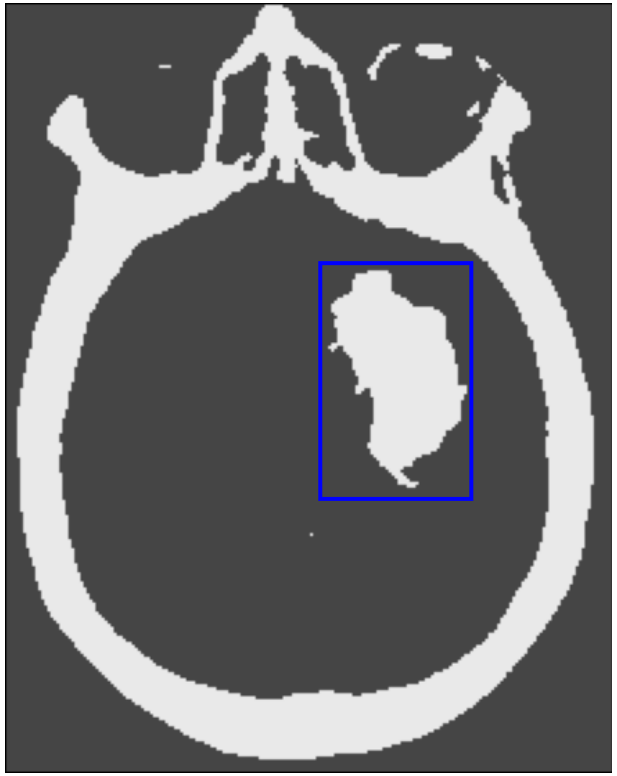
Some Applications of Machine Learning Algorithms

- Application 4 (Tumor Detection)

  For a given CT/MRI film, find the tumor/abnormal area.

  - We will revisit this example later.

The original scan image



The tumor is strengthened

Two Main Categories of Machine Learning Topics

- Supervised learning

  Learn with a "teacher". Mathematically, each input training data x_i includes a response y_i which works as the "teacher".

  - Regression
  - Classification

- Unsupervised learning

  Learn without a "teacher". Mathematically, for each training data x_i, we do NOT have a given response.

  - Clustering
  - Data dimension reduction

Regression

Based on n training data (each with an input vector $\boldsymbol{x}_i$ and a response y_i) and a function $\hat{y} = f(\boldsymbol{x}; \boldsymbol{\beta})$, where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ represent the input data and parameters, the machine will determine the parameters $\boldsymbol{\beta}$ that make function f fit historical data the "best".

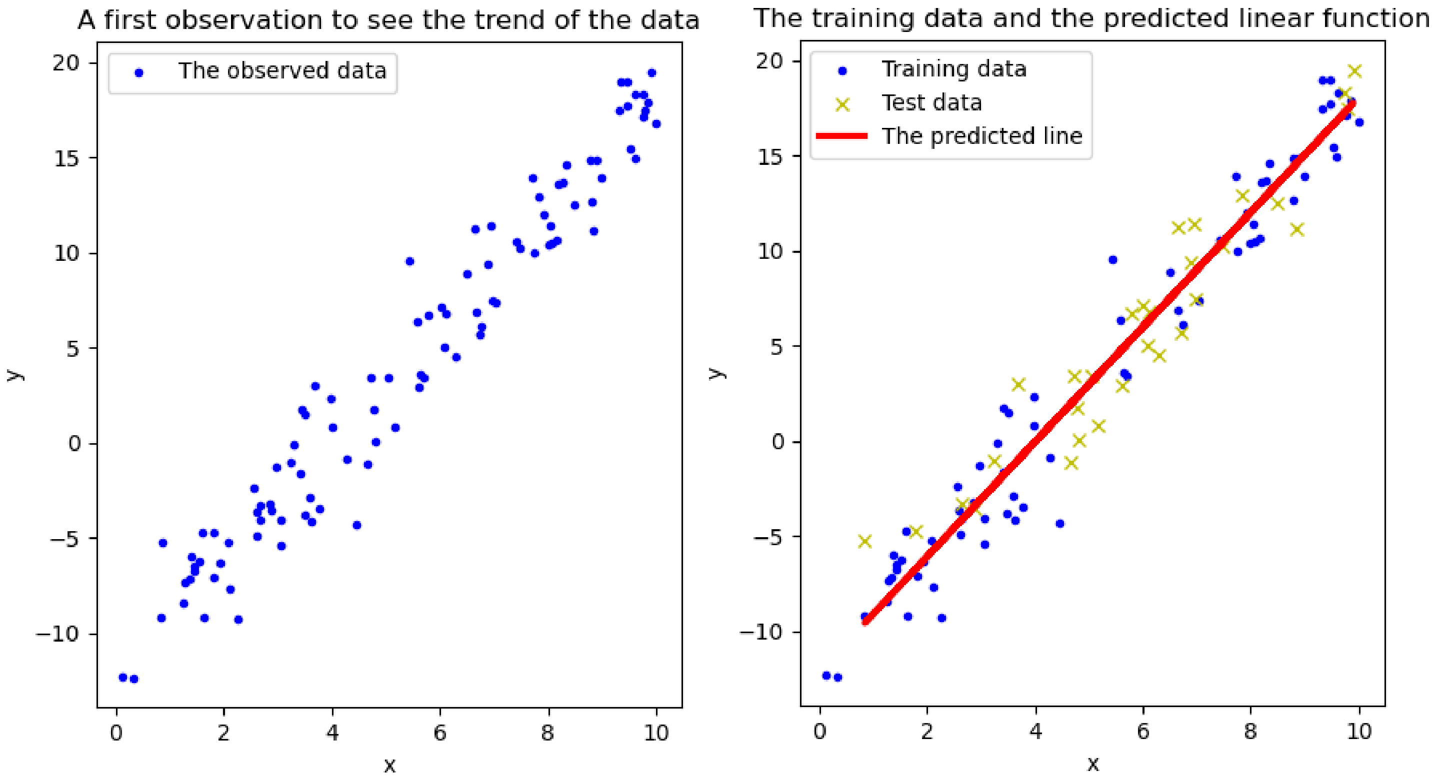
- Most common f (linear regression) :

  $$\hat{y} := f(\boldsymbol{X}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  - In fact, we can fit for any f, not necessarily linear.
- fit the "best": "best" under a given measure.
  - Most common measure (least squares regression):

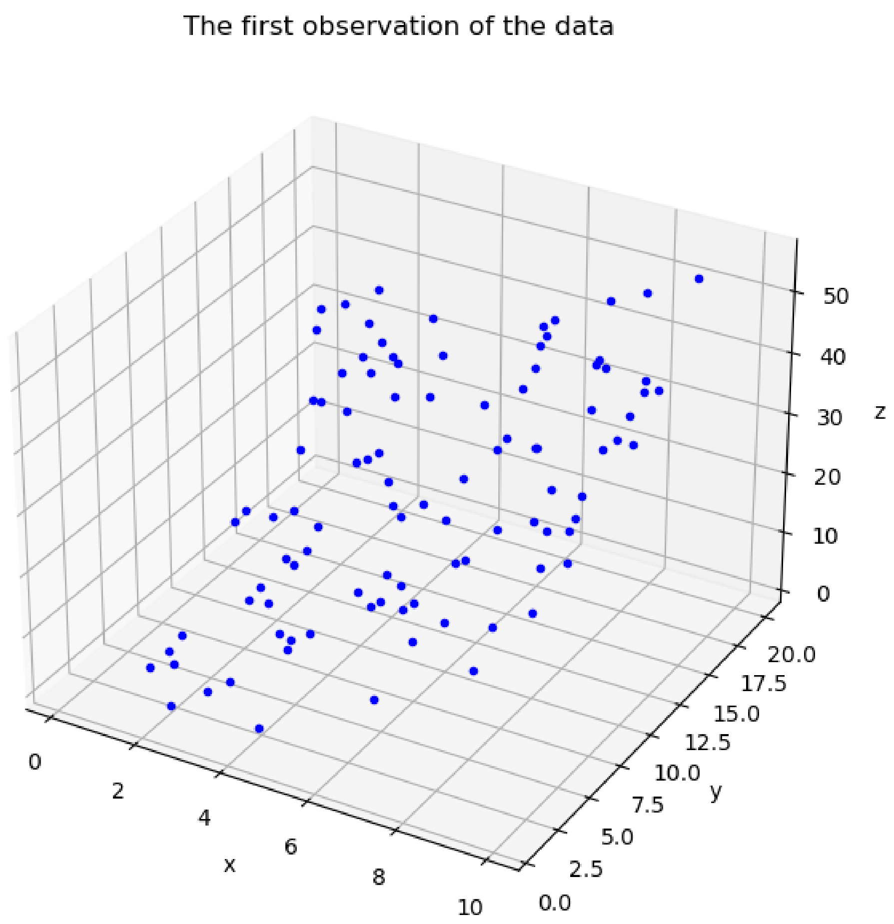
    $$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2$$

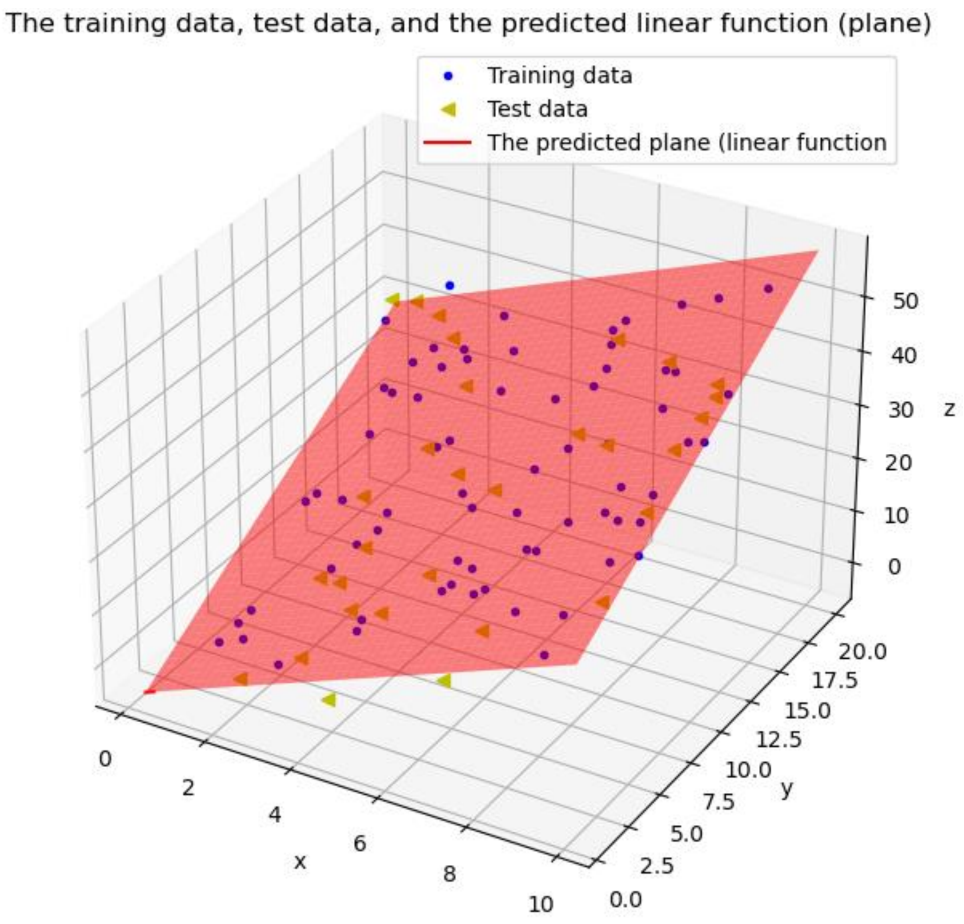
  - Some other measure: Baysian regression, SVM regression, Lasso regression, Ridge regression, etc.



Example 1-1 A 2D linear regression

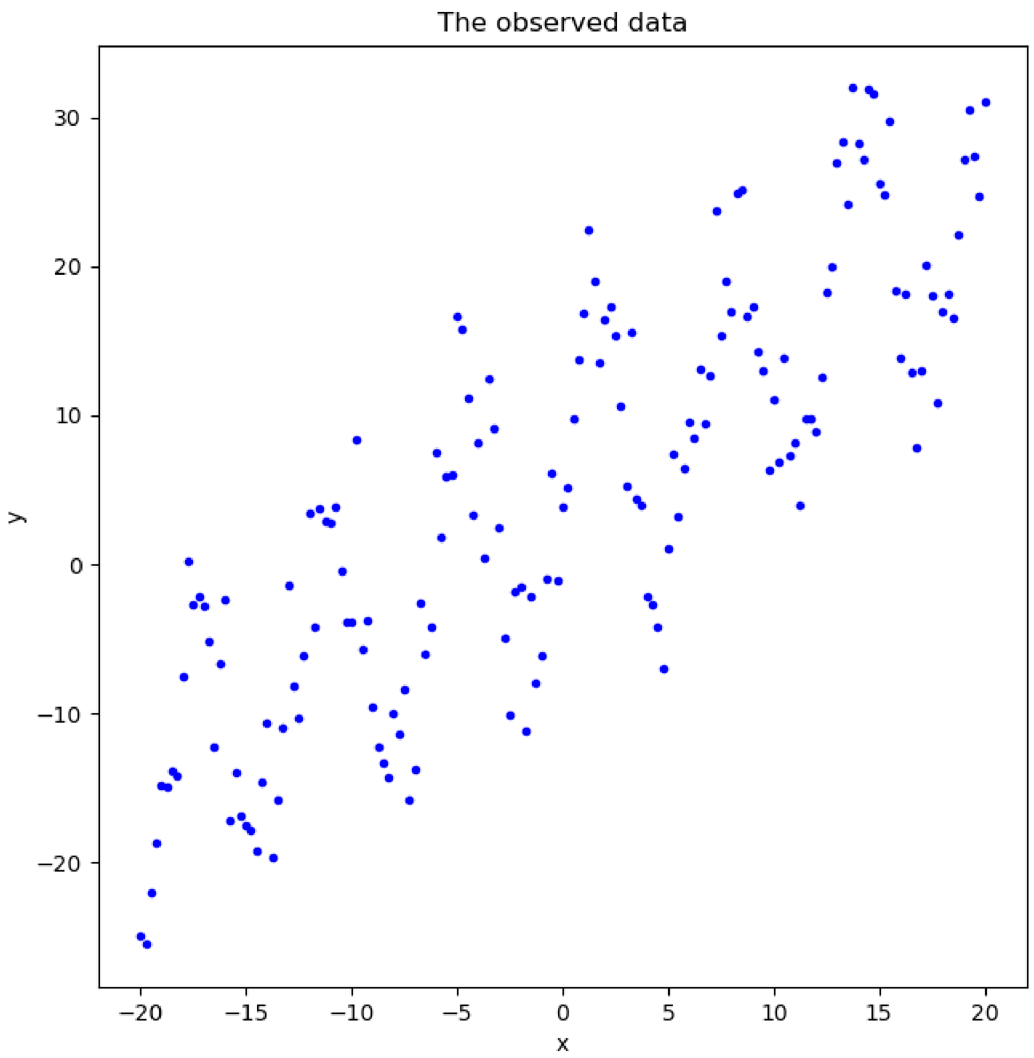
Regression

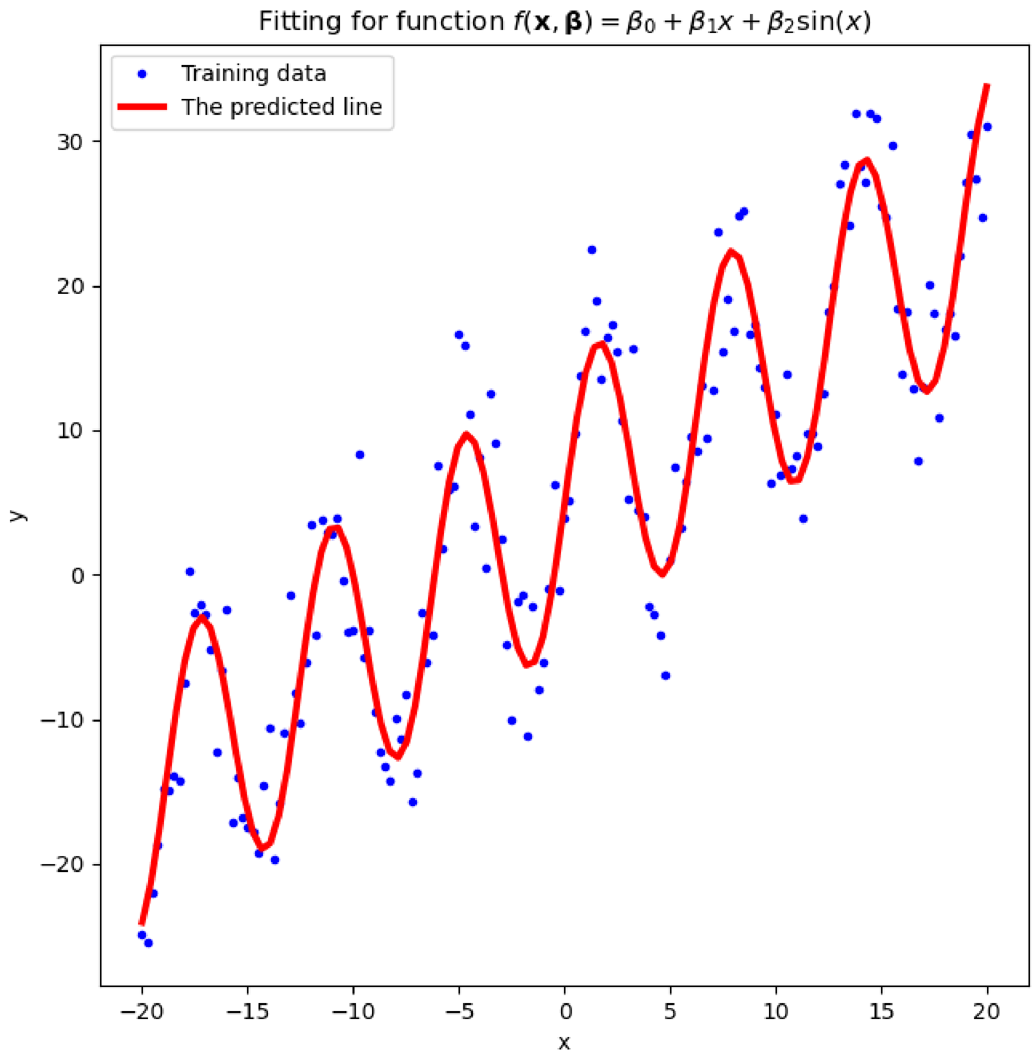




Example 1-2 A 3D linear regression

Regression





Example 1-3 A 2D nonlinear regression in which $f(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

Classification in General

- Definition of Classification in ML

  Based on given n training data (each with an input vector $\boldsymbol{x}_i$ and a class label y_i), the machine need to determine the class for a new data.

Training data, each with a class label: "cats" or "dogs"

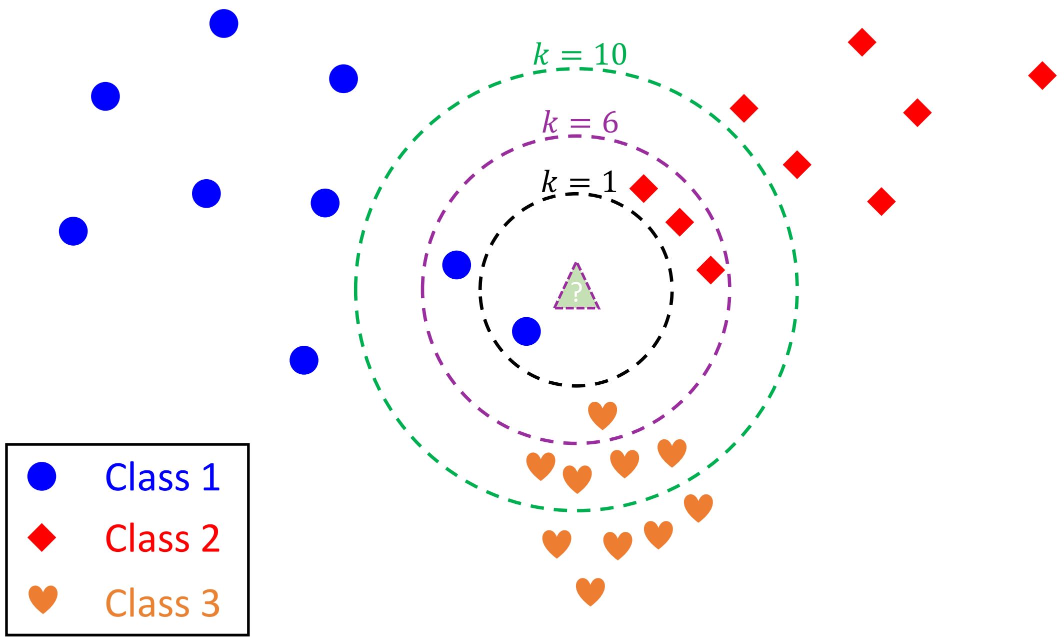
Target data: what is/are included in this picture?

Classification in General

- Classifiers to be introduced in this presentation
  - Distance-based classifier
    - k-nearest neighbor (kNN)
    - Support vector machine (SVM; strictly, a margin-based classifier)
  - Probability distribution-based classifier
    - Naïve Bayes
    - Logistic regression
    - Discriminant analysis

Classification: k-nearest neighbor (kNN)

We classify the target data $\boldsymbol{x}$ by the majority class label of its k nearest neighbor from the training data.



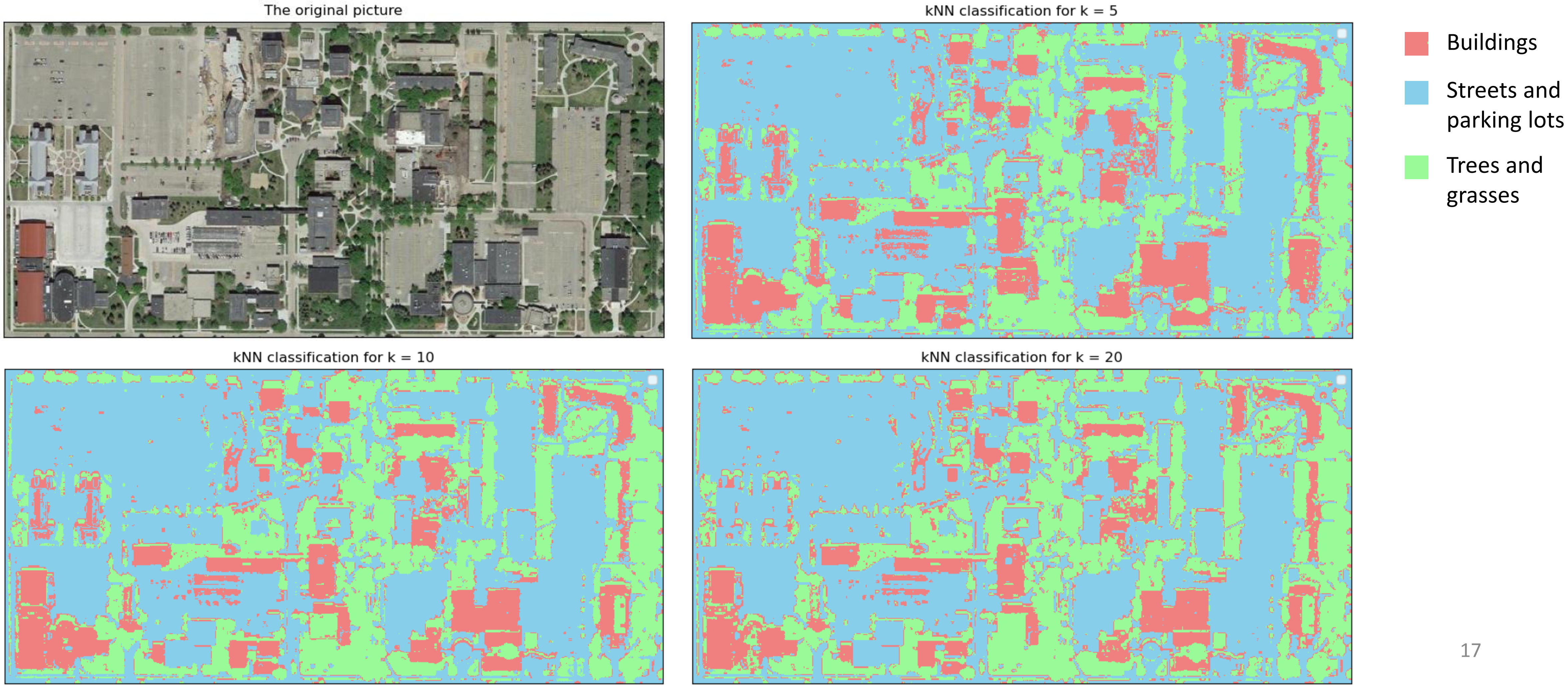
- If $k = 1$, we classify the target data to Class 1
- If $k = 6$, we classify the target data to Class 2
- If $k = 10$, we classify the target data to Class 3

Classification: k-nearest neighbor (kNN)

Example 2-1 [Classification of NDSU campus geography]

A screen shot of part of NDSU campus from google satellite (with label turned off):

https://www.google.com/maps/@46.8960412,-96.8045104,556m/data=!3m1!1e3

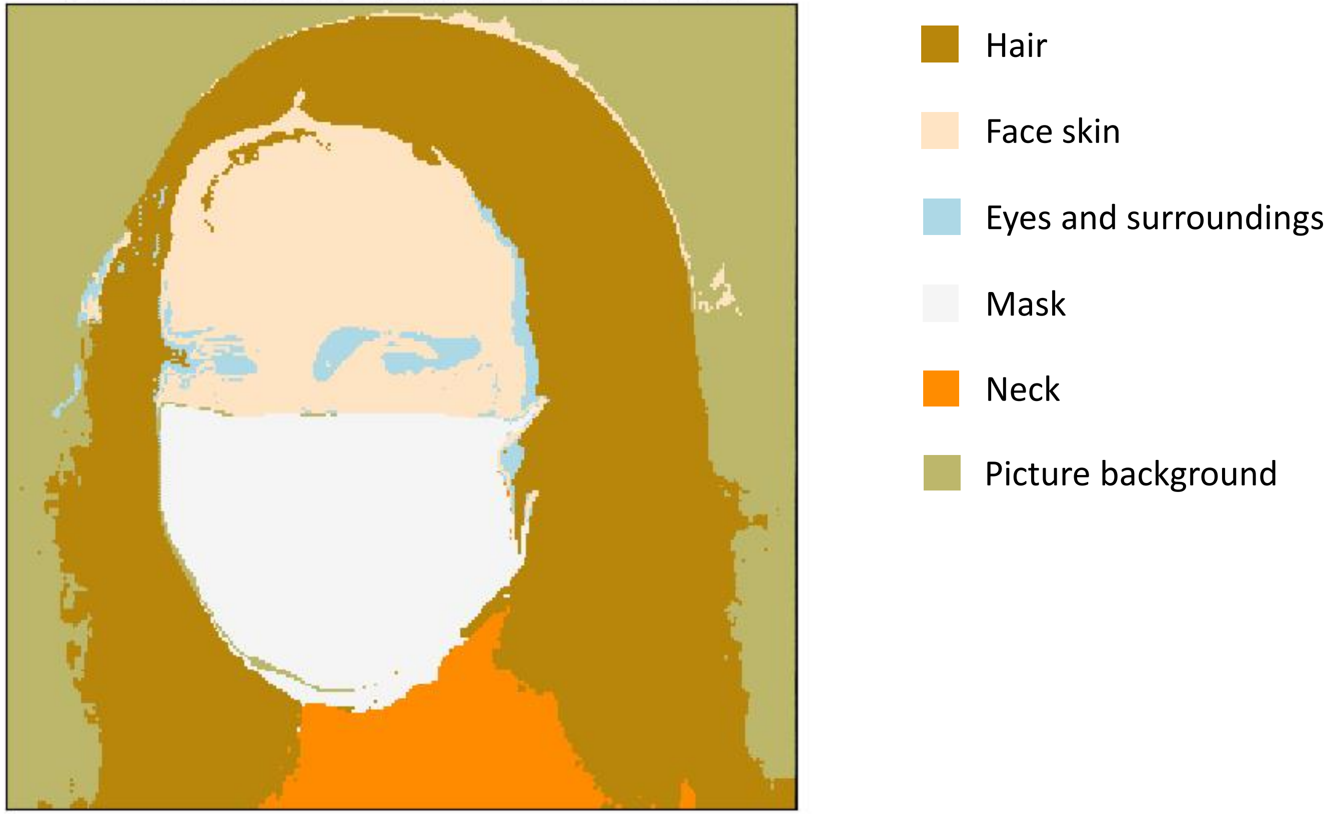


Classification: k-nearest neighbor (kNN)

Example 2-2 [Face part detection of Monalisa with a mask]

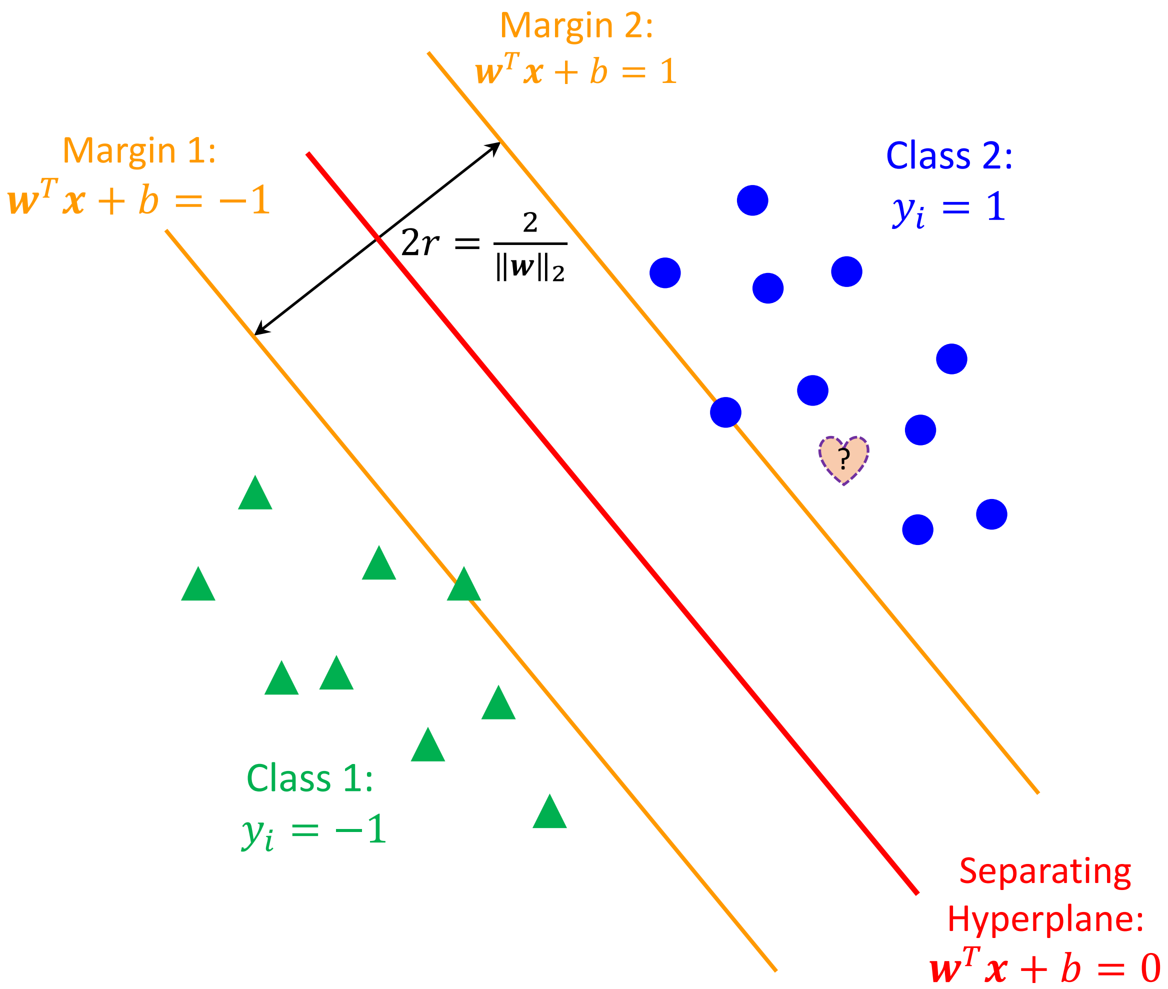
The original picture

RGB-and-Distance-based 2NN face parts prediction



Classification: Support Vector Machine (SVM)

In the training, we try to maximize the distance of the margins (to maximize the difference between classes). We classify the target data $\boldsymbol{x}$ to the class label whose margin area includes $\boldsymbol{x}$.



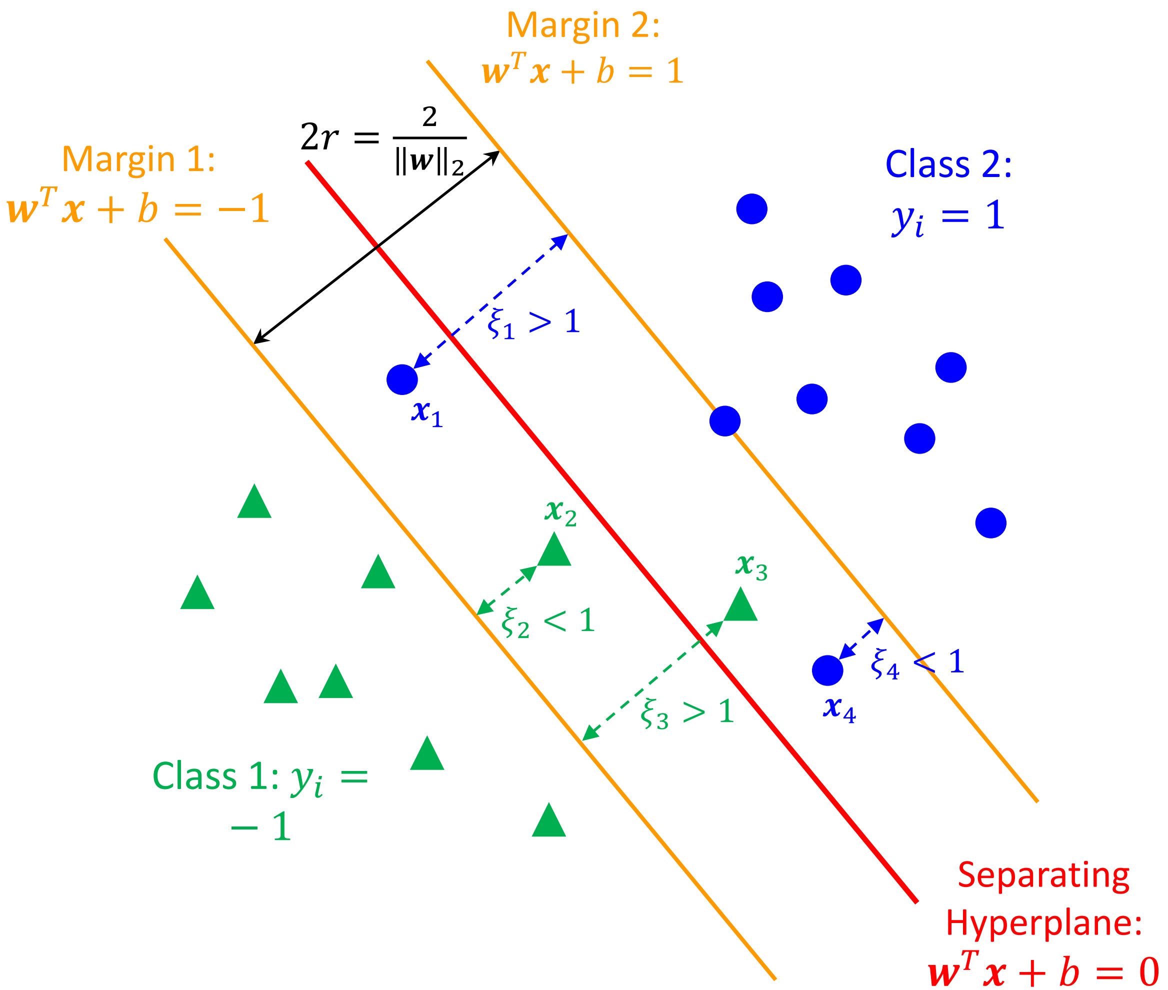
Hard-margin SVM (error intolerable):

$$\min_{\boldsymbol{w},\, b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

s.t. $\quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \ \ i = 1, \cdots, n$

Classification: Support Vector Machine (SVM)

We can add penalty items to make the SVM model error tolerable.



Soft-margin SVM (error tolerable):

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \cdots, n$$

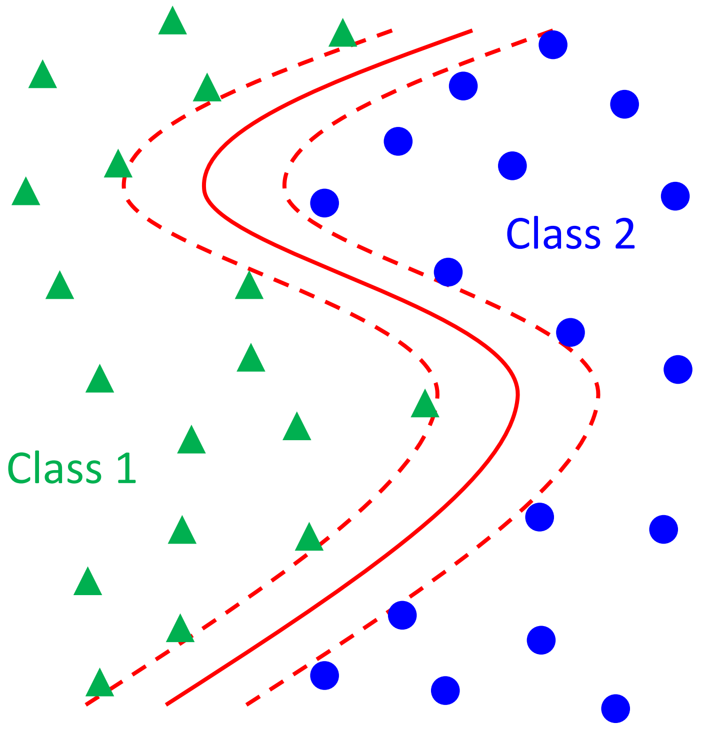
$$\xi_i \geq 0, \quad i = 1, \cdots, n$$

ξ_i: the “error”

C: the regularization parameter, indicating the weight of the penalty

Classification: Support Vector Machine (SVM)

Using the kernel trick, we can even create nonlinear margins!



SVM with kernel trick

SVM with Kernel Trick (error tolerable):

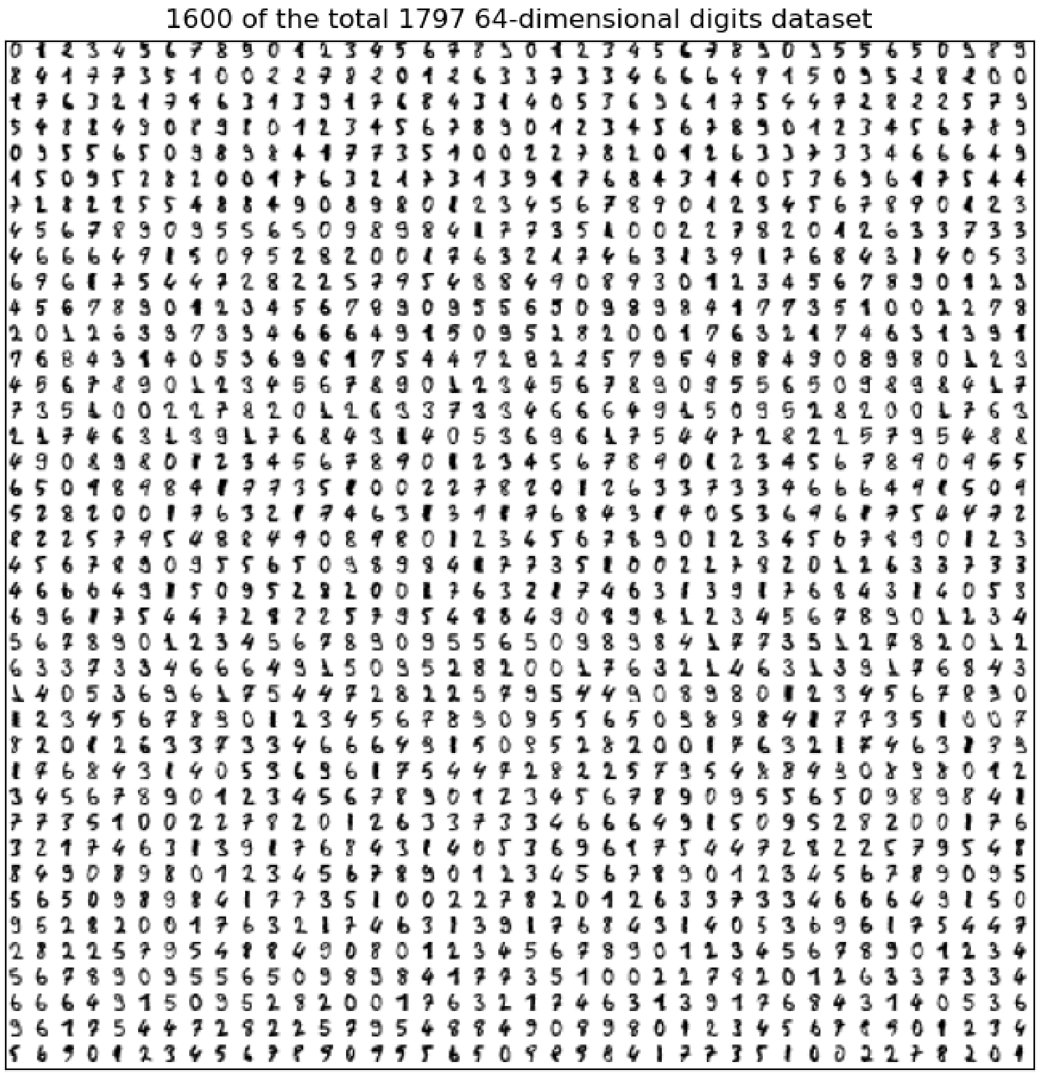
$$\max_{\boldsymbol{\mu}} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_i\mu_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) + \sum_{i=1}^{n}\mu_i$$

s.t.
$$\sum_{i=1}^{n}\mu_i y_i = 0$$
$$0 \le \mu_i \le C, \quad i = 1,\cdots,n$$

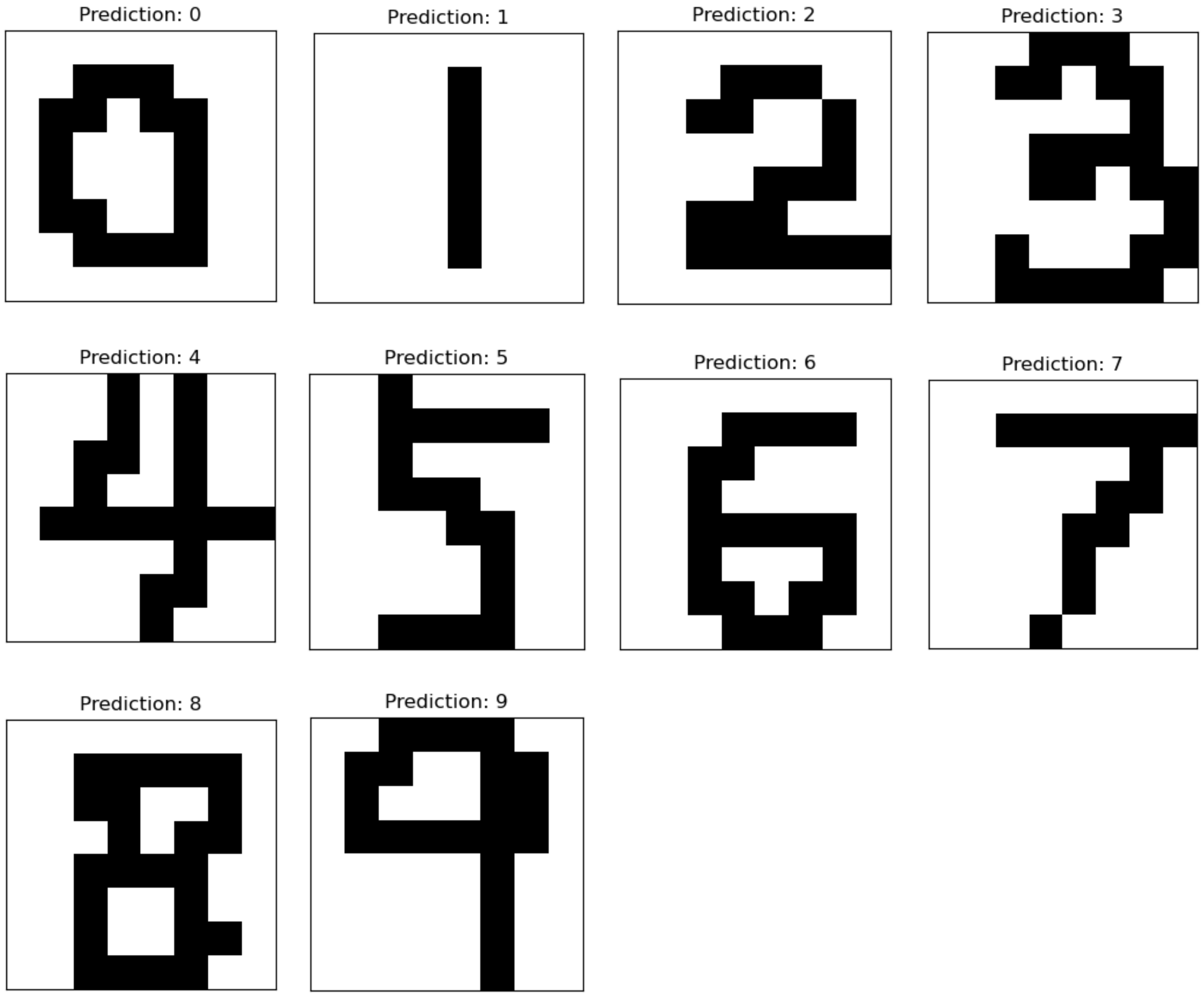
- $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$: the kernel function
  - Gaussian (RBF) kernel: $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\left\{-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right\}$
  - p^{th}-degree polynomial function: $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \left(r + \langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle\right)^p$
  - Sigmoid kernel: $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \tanh\left(\gamma\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle + r\right)$

Classification: Support Vector Machine (SVM)

Example 3 [Handwritten digit recognition]



Part of the training data



Test results

Classification: Naïve Bayes

Naïve Bayes Assumption: given the sample label, all the p features are mutually independent. By Bayes formula and the chain rule of conditional probability:

$$\mathbb{P}\{Y = y|\boldsymbol{X} = \boldsymbol{x}\} = \mathbb{P}\{Y = y\} \prod_{j=1}^{p} \mathbb{P}\{X_j = x_j|Y = y\}$$

- Classification for a target data $\boldsymbol{x}$ (choosing the label with the maximum posterior probability):

$$\hat{y}_{NB}(\boldsymbol{x}) = \arg \max_{y \in \{1,\cdots,m\}} \sum_{j=1}^{p} \log\left(\mathbb{P}\{X_j = x_j|Y = y\}\right) + \log(\mathbb{P}\{Y = y\})$$

Classification: Naïve Bayes

Example 4-1 [Categorical Naïve Bayes for health condition detection]

Columns		Explanation
Output:	Healthy?	0: unhealthy, 1: health
Input:	Gender	0: female, 1: male
	Age Class	2: 20-29, 3: 30-39, 4: 40-49
	Measure	An integer between 0 and 5, indicating the value of the measure

We expect to tell whether or not a person is healthy based on his/her gender, age class, and a medical measure (e.g., blood pressure). The notations of the training data are showed in the above table. Are the following two people healthy?

a) $\boldsymbol{t}_1 = (0,2,0)$, i.e., (gender = female, age class = 20-29, measure = 0)
b) $\boldsymbol{t}_2 = (1,4,0)$

```
The relative posteriors are: [[0.39085596 0.60914404]
 [0.56525163 0.43474837]]
The health conditions are (0: unhealthy; 1: health): [1 0]
```

Classification: Naïve Bayes

Example 4-2 [Multinomial Naïve Bayes for spam email detection]

> *"Congratulations! You have been selected to win $500,000. Click the following link to create an account and get your dollars!"*

We expect the computer to detect spam emails automatically.

- We capture the number of occurrence of some "sensitive" words ("dollar", "select", "reward", "claim", "win") and then based on them to do classification (using multinomial distribution). So the target data: $\boldsymbol{x} = [2,1,0,0,1]$.

```
The relative posterior probability of the test data is: [[0.388 0.612]]
This email is in class (class 0: not spam; class 1: spam): 1
```

- The probability that this email is a spam email is 61.2%, versus 38.8% not a spam.
- Thus, we predict it to be a spam email.

Classification: Logistic Regression

Referring to logic function $logit(x) = \log\left(\frac{x}{1-x}\right)$, model the problem to be

$$\mathbb{P}\{Y = j \mid \boldsymbol{X} = \boldsymbol{x}\} = \frac{e^{\beta_{j0}+\boldsymbol{\beta}_j^T\boldsymbol{x}}}{1+\sum_{k=1}^{m-1} e^{\beta_{k0}+\boldsymbol{\beta}_k^T\boldsymbol{x}}}, \qquad j = 1, \cdots, m-1$$

and

$$\mathbb{P}\{Y = m \mid \boldsymbol{X} = \boldsymbol{x}\} = \frac{1}{1+\sum_{k=1}^{m-1} e^{\beta_{k0}+\boldsymbol{\beta}_k^T\boldsymbol{x}}}$$

Then, for the label, we assume that

$$\boldsymbol{Y} \sim multinoulli\big(p_1(\boldsymbol{x}; \beta_{10}, \boldsymbol{\beta}_1), p_2(\boldsymbol{x}; \beta_{20}, \boldsymbol{\beta}_2), \cdots, p_m(\boldsymbol{x}; \beta_{m0}, \boldsymbol{\beta}_m)\big),$$

where

$$p_j(\boldsymbol{x}; \beta_{j0}, \boldsymbol{\beta}_j) \coloneqq \mathbb{P}\{Y = j \mid \boldsymbol{X} = \boldsymbol{x}\}, \forall j \in \{1, \cdots, m\}$$

- Training process (maximum likelihood estimation):

$$\left(\widehat{\boldsymbol{\beta}}_0, \widehat{\boldsymbol{B}}\right) = \arg \max_{\boldsymbol{\beta}_0 \in \mathbb{R}^m, \boldsymbol{B} \in \mathbb{R}^{m \times m}} \log L(\boldsymbol{\beta}_0, \boldsymbol{B})$$

- Classification for a target data $\boldsymbol{x}$ (choosing the label with the maximum posterior probability):

$$\hat{y}_{LR}(\boldsymbol{x}) = \arg \max_{j \in \{1, \cdots, m\}} p_j(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}_{j0}, \widehat{\boldsymbol{\beta}}_j)$$

Classification: Logistic Regression

Example 5 [Credit card automatic approval/denial system]

Columns		Explanation
Output:	Approval?	The response. 0: denied, 1: approved
Input:	Age	The age of the applicant (between 18 and 70)
	Married	Currently married (1) or not (0)
	Education	Education level. 5: Ph.D., 4: Master, 3: Bachelor, 2: High School, 1: lower than high school
	YearsEmployed	Number of years being full-time employed
	PriorDefault	Number of defaults in last year
	Income	The annual income of last year (unit: $1000)
	CreditScore	The current credit score

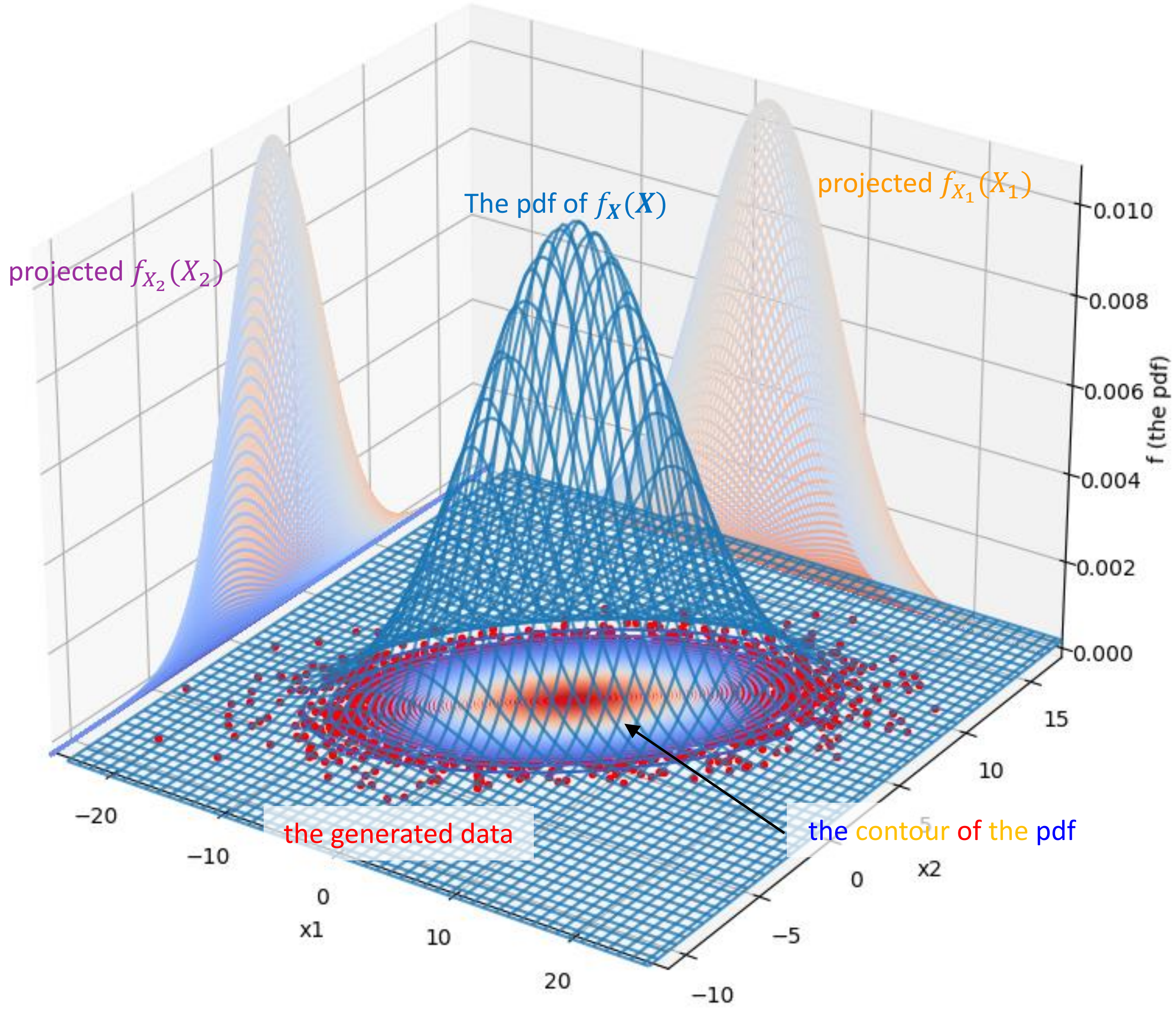
We expect the computer to tell whether or not an application will have a good payoff habit and then make a decision (approval or deny) on his/her credit card application. The notations of the training data are showed in the above table. What is the decision for the following application?

- $\boldsymbol{x} = (40, 1, 3, 16, 0, 50, 600)$

```
The probability of belong to [good payoff habit, not-good payoff habit] is: [[0.24387274 0.75612726]]
The result of the test data is (Approval: 1, Denial: 0): 1
```

Classification: Discriminant Analysis

3D probability density function with projected univariate pdf contours and 2D data



Assuming the input matrix $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (i.e., multivariate Gaussian distribution) in a d-dimensioned feature space. The idea of discriminant analysis is to model the difference between the classes:

$$\log \frac{\mathbb{P}\{Y=k|\boldsymbol{X}=\boldsymbol{x}\}}{\mathbb{P}\{Y=l|\boldsymbol{X}=\boldsymbol{x}\}}, \ k, l \in \{1, \cdots, m\}, \ k \neq l$$

Classification: Discriminant Analysis

- Linear discriminant analysis (LDA)
  - Assumption: $\boldsymbol{X}|Y = y \ \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$
  - Classification for a target data $\boldsymbol{x}$ (choosing the label with the maximum posterior probability):

$$\hat{y}_{LDA}(\boldsymbol{x}) = \arg\max_{k\in\{1,\cdots,m\}} \boldsymbol{x}^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_k - \frac{1}{2} \widehat{\boldsymbol{\mu}}_k{}^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_k + \log \hat{\pi}_k$$

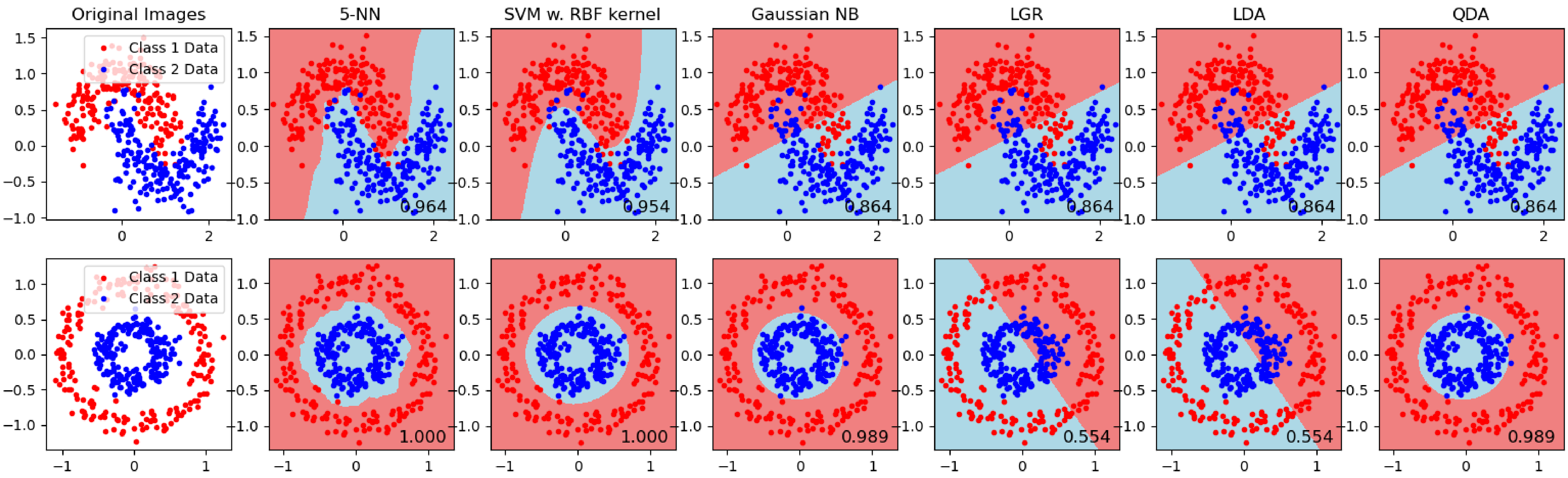
  where

$$\widehat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \delta_{y_i=k}}{n_k}, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-m} \sum_{k=1}^{m} \sum_{i=1}^{n} \delta_{y_i=k} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^T, \quad \hat{\pi}_k = \frac{n_k}{n}$$

- Quadratic discriminant analysis (QDA)
  - Assumption: $\boldsymbol{X}|Y = y \ \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$
  - Classification for a target data $\boldsymbol{x}$ (choosing the label with the maximum posterior probability):

$$\hat{y}_{QDA}(\boldsymbol{x}) = \arg\max_{k\in\{1,\cdots,m\}} -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) - \frac{1}{2} \log(\det(\boldsymbol{\Sigma}_k)) + \log \pi_k$$

Classification: Discriminant Analysis




```
The test accuracy of 5-NN classifier for MOON dataset is:  0.9917
The test accuracy of 5-NN classifier for CIRCLE dataset is:  0.9917

The test accuracy of SVM w. RBF kernel classifier for MOON dataset is:  0.9667
The test accuracy of SVM w. RBF kernel classifier for CIRCLE dataset is:  1.0000

The test accuracy of Gaussian NB classifier for MOON dataset is:  0.8417
The test accuracy of Gaussian NB classifier for CIRCLE dataset is:  0.9917

The test accuracy of LGR classifier for MOON dataset is:  0.8500
The test accuracy of LGR classifier for CIRCLE dataset is:  0.4667

The test accuracy of LDA classifier for MOON dataset is:  0.8417
The test accuracy of LDA classifier for CIRCLE dataset is:  0.4750

The test accuracy of QDA classifier for MOON dataset is:  0.8500
The test accuracy of QDA classifier for CIRCLE dataset is:  0.9917
```

Example 6 [Comparison of the classifiers]

- Two groups of data: “moon” and “circle” shaped data
- Classifiers: 5NN, Gaussian Naïve Bayes, logistic regression, LDA, and QDA,

Two Main Categories of Machine Learning Topics

- Supervised learning

  Learn with a "teacher". Mathematically, each input training data x_i includes a response y_i which works as the "teacher".

  - Regression
  - Classification

- Unsupervised learning

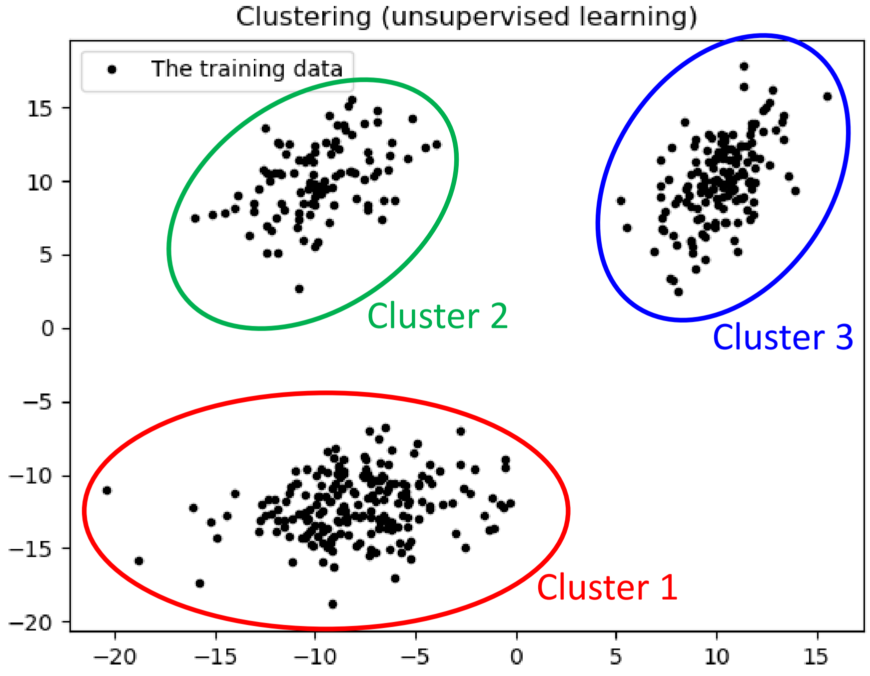
  Learn without a "teacher". Mathematically, for each training data x_i, we do NOT have a given response.

  - Clustering
  - Data dimension reduction

Clustering

Clustering is to group data objects into clusters such that objects in the same cluster are more “similar” (or, more closely related) to each other than to those in other clusters.

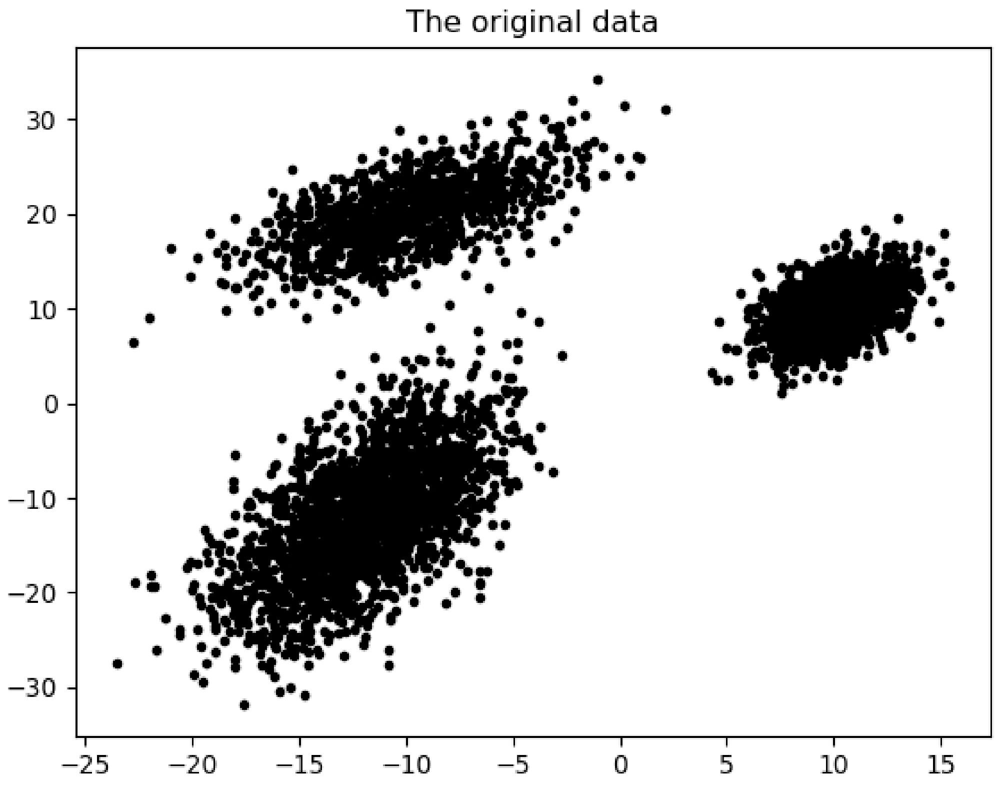
- Clustering is a type of <u>unsupervised learning</u>.
- We will introduce the most popular one: <u>*K-means clustering*</u>.

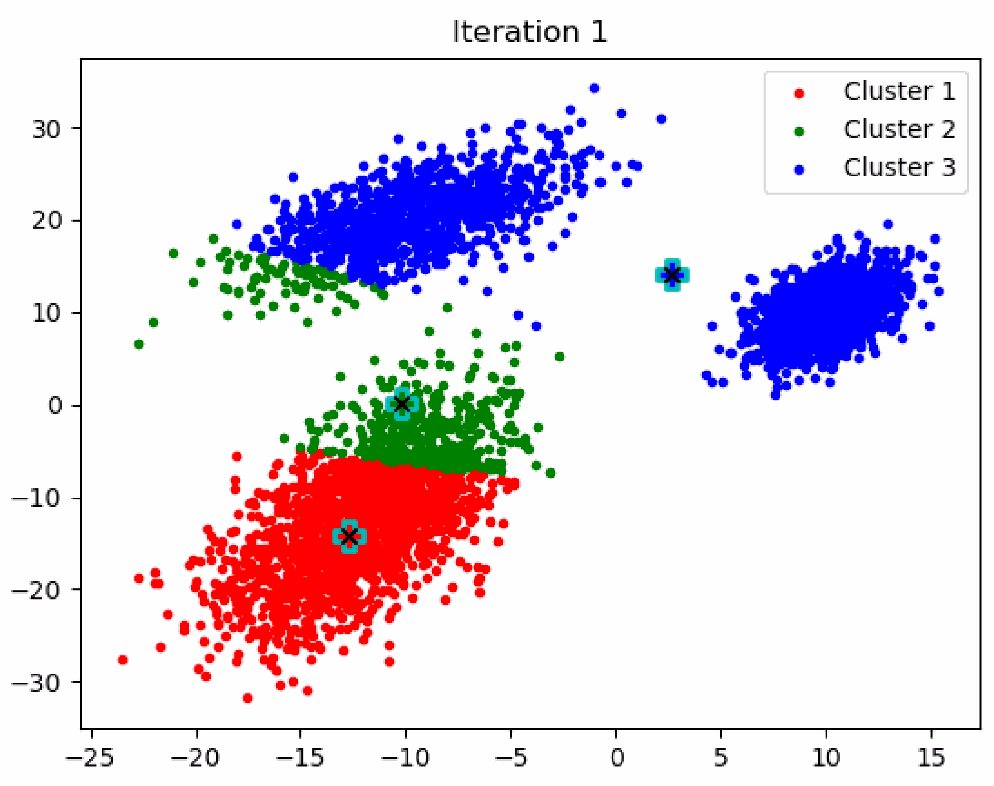


Clustering: K-means Clustering

The motivation of the K-means clustering is to assign the n data into K clusters such that the sum of dissimilarity measure of all the data to the centroid (i.e., the center) of their corresponding clusters is minimized.

Example 7-1 [An animation for the 3-means clustering]





Clustering: K-means Clustering

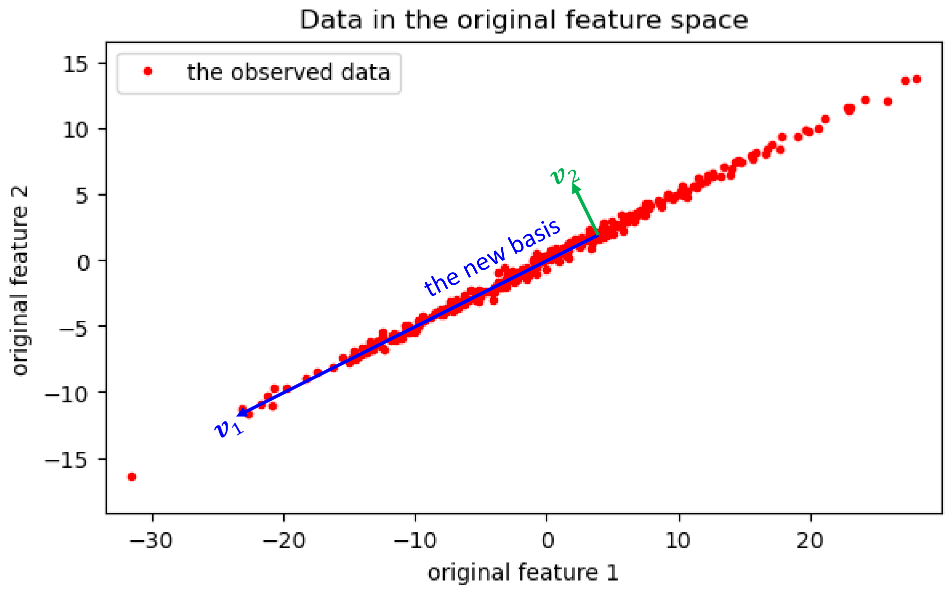
Example 7-2 [Tumor detection]

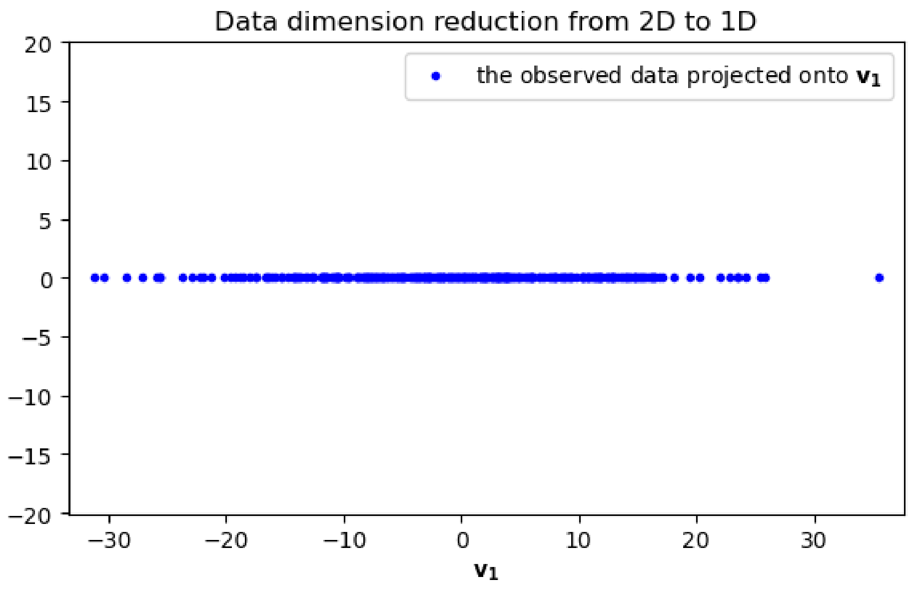
- The tumor part is strengthened.
- We can further process the clustered image. E.g., to measure the size of the tumor and do classification to measure the risk and determine whether a surgery is needed.

Data Dimension Reduction

The data dimension reduction, in general, is a technique to transform a dataset (including n data) from a high-dimensional space into a low-dimensional space (e.g., from p features to t features, where $t < p$ or even $t \ll p$) such that the low-dimensional representation of the data can keep “significant information” of the dataset.

- Applications:
  - storage space saving
  - “important pattern” strengthening

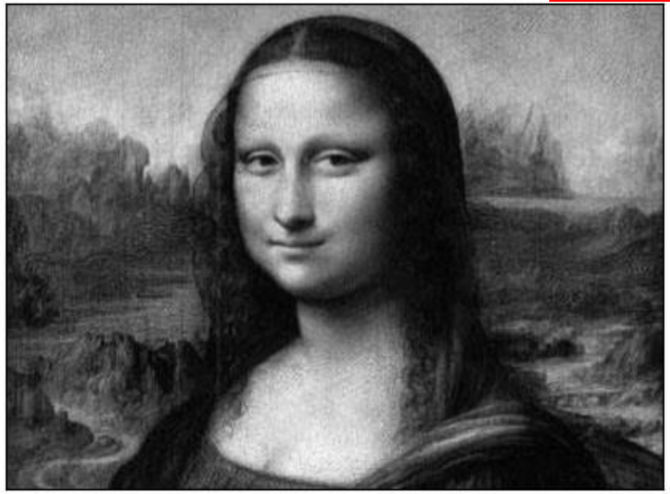




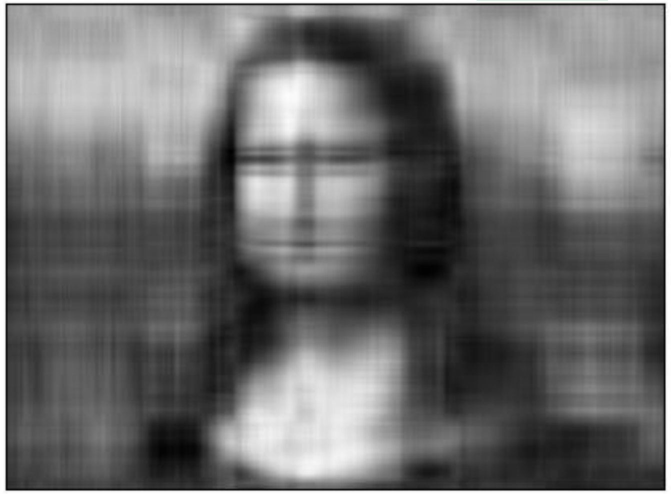
Data Dimension Reduction

Example 8-1 [Image compression by principal component analysis]

The original gray-scaled image: 982·726 = $\boxed{\text{696.22 KB}}$

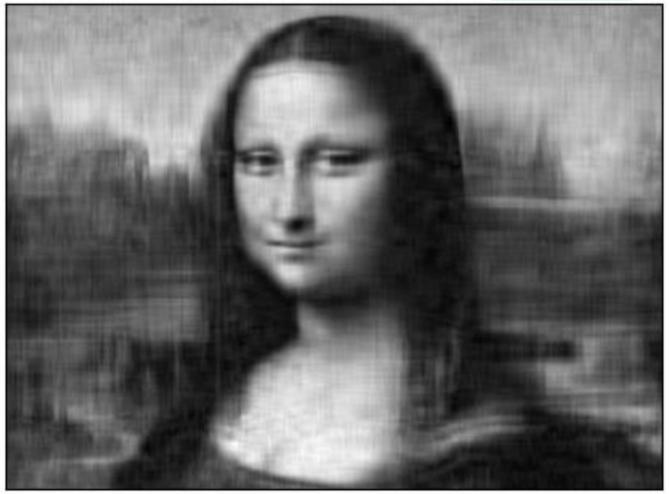


5-PCs reconstruction: 5·726 = $\boxed{\text{3.54 KB}}$



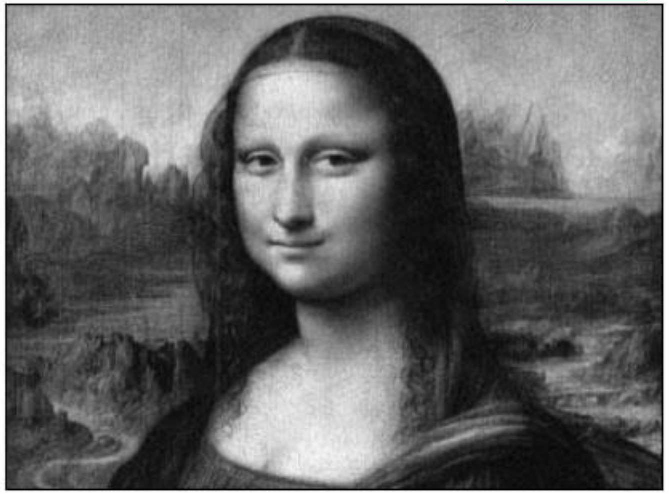
10-PCs reconstruction: 10·726 = $\boxed{\text{7.09 KB}}$

20-PCs reconstruction: 20·726 = $\boxed{\text{14.18 KB}}$



50-PCs reconstruction: 50·726 = $\boxed{\text{35.45 KB}}$

100-PCs reconstruction: 100·726 = $\boxed{\text{70.90 KB}}$



Data Dimension Reduction

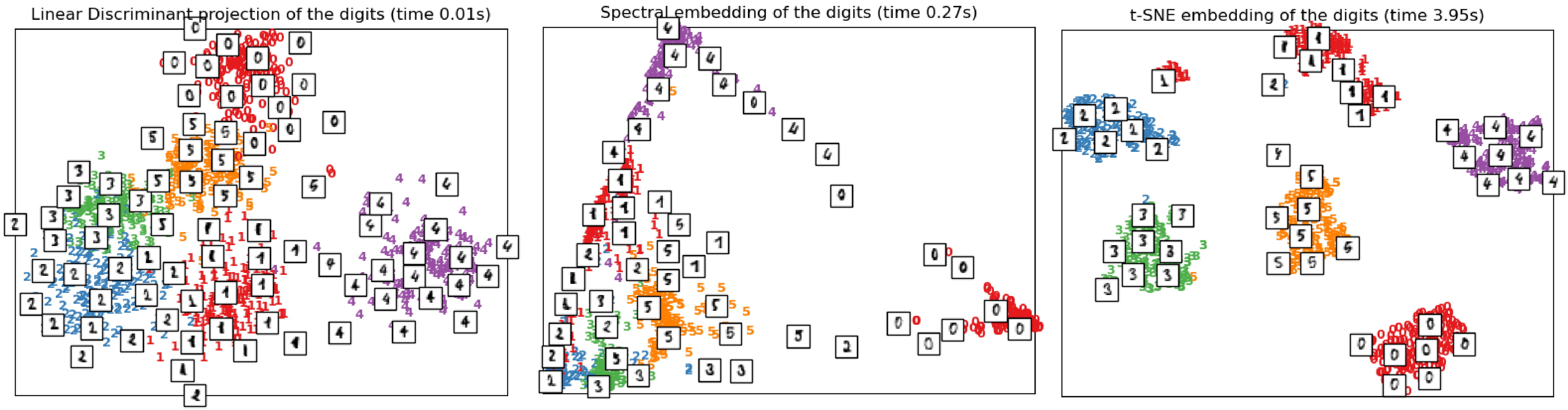
Example 8-2 [Handwritten digit recognition strengthened by dimension reduction techniques]

- Handwritten recognition for digits 0, 1, 2, 3, 4, 5
- By data dimension reduction techniques (i.e., from $8 \times 8 = 64\text{D}$ to 2D), you will see that the distinguishment of data from different classes are (greatly) strengthened.
- Most part of the code in this example is originally from https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html#sphx-glr-auto-examples-manifold-plot-lle-digits-py

A selection from the 64-dimensional digits dataset

The training data

Data Dimension Reduction



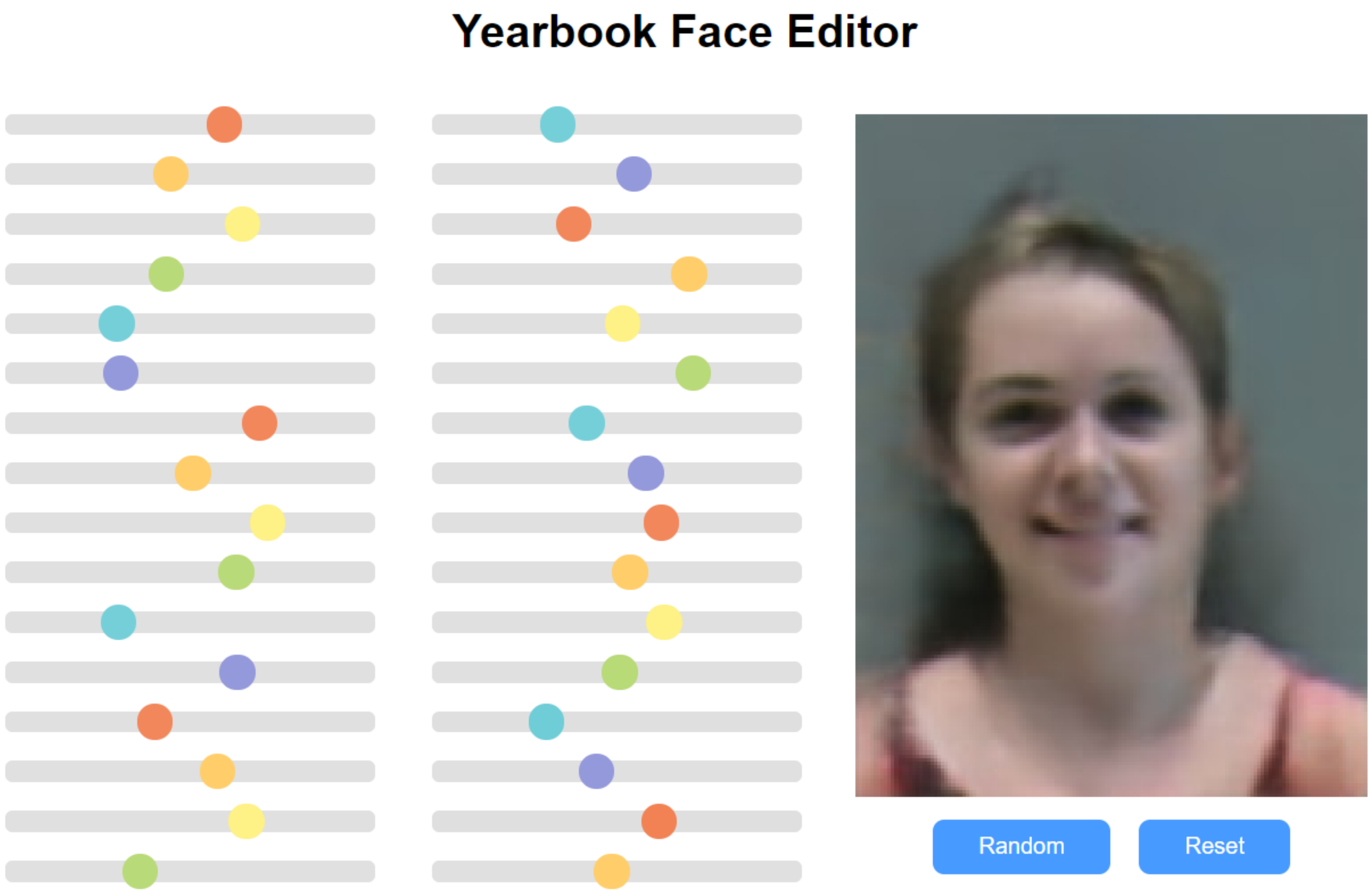
- In this example, classification based on the low-dimensional (2D) spaces of the data is much more efficient than on the original high-dimensional space (64D).

Data Dimension Reduction

Example 8-3 [A face generator]

See http://codeparade.net/faces/

- Each face is composed of 32 features (a dimension reduction!)
- See https://www.youtube.com/watch?v=4VAkrUNLKSo for a further introduction.
- Have fun!



Thank you!

Please feel free to contact me for any questions:

yiwen.xu@ndsu.edu

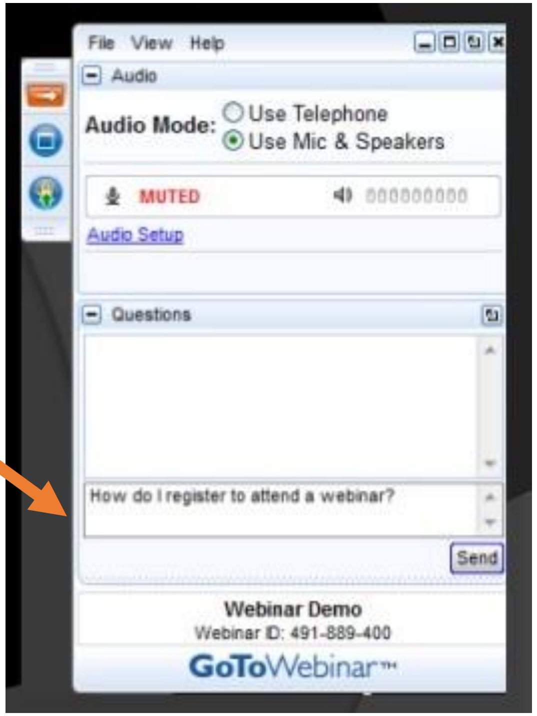
Strengthening
University-Industry
Partnerships

North Dakota State University

Q&A

GoToWebinar

Please submit your questions using the questions tab on your screen.



Open Innovation at Merck KGaA, Darmstadt, Germany / EMD Group

May 12, 2021 | 12 - 1 PM ET

Moderator:
Patrick Limbach
University of Cincinnati

Ulrich Betz
Merck KGaA, Darmstadt, Germany

THANK YOU!

Please check your email and complete the survey so UIDP can better meet your needs.