Data Mining

INFS 4203/7203

Miao Xu

miao.xu@uq.edu.au

The University of Queensland, 2020 Semester 2

More about the assignments

- The assignments not only cover the lecture slides, but also other topics we talked explicitly in class. **Checking the course record!**
- Partially correct can be given partial marks (a, c are correct, selecting a or c will be given partial marks)

Discussions on Piazza

- Discussions are highly welcomed on Piazza
- Everyone is encouraged to join in both asking and answering
- State you point concisely and clearly with support/evidence
- Discuss in a respectable way, regarding different backgrounds
- @157, @165 as an example

> *"(1) rewarding to help, and (2) a good test of my own understanding, or lack thereof, if I can explain a concept to others. It's valuable for everyone if we promote discussion like this."*

Last week: anomaly detection

Outlier vs. noise

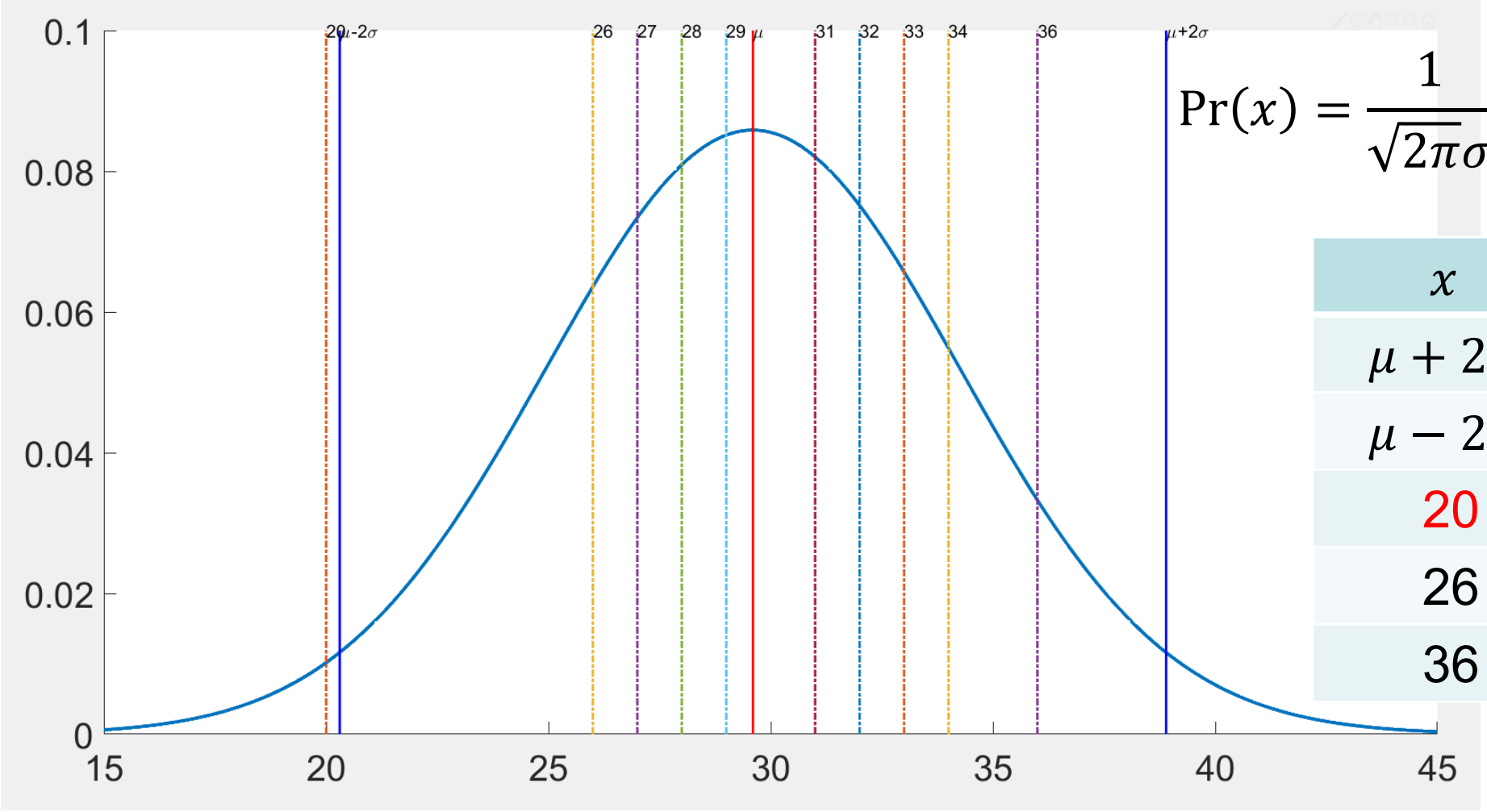
- Noise
  - erroneous, perhaps random perturbation or contaminating
  - not necessarily produce unusual values or objects
  - may not be interesting (not that novel)
- Anomalies:
  - can be caused by noise, may also be caused by adversaries
  - unusual values
  - Interesting (something novel)

They are related but distinct concepts!

- In the clustering part, such kind of point is described as “noise” to be consistent with DBSCAN method
- For example, “min” is sensitive to noise, i.e., the results of the AGNES method with min operation will be highly impacted if there exists some noise points outside of the main cluster structure.
- From “anomaly detection” lectures, all “outliers” will be described as “outlier” or “anomaly” no matter it is noisy or not unless we otherwise specified

Calculate the probability

{20, 26, 27, 28, 29, 31, 32, 33, 34, 36}



$$\Pr(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

x	$\Pr(x)$
$\mu + 2\sigma$	0.0116
$\mu - 2\sigma$	0.0116
20	0.0102
26	0.0636
36	0.0333

More on the $\Pr(x)$

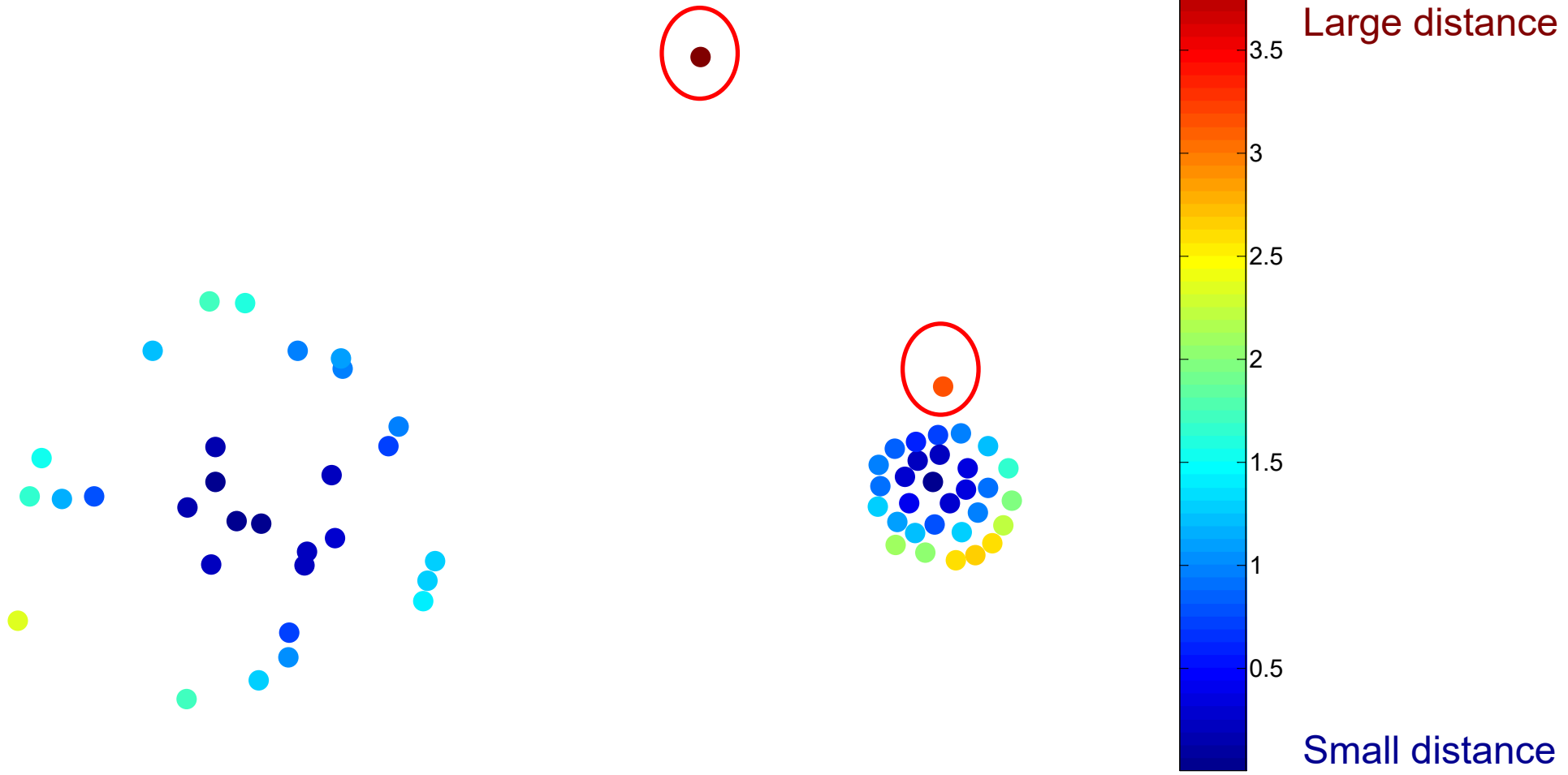
- Probability density function (PDF)
- The value of the PDF at two different samples can be used to infer, in any particular draw of the random variable, how much more likely it is that the random variable would equal one sample compared to the other sample.

$$\Pr(x) = \lim_{\delta \to 0} \text{Probability}(x - \delta, x + \delta\ ) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Thanks to the contribution of Ron for this.

Example with related distance

- For prototype-based clustering



Using relative distance will make the algorithm insensitive to density.

More on sensitive to density in anomaly detection

- More accurate: the algorithm is not easy at detecting outliers when the data set has regions of widely differing densities

Thanks to the contribution of Ron and Farshad for this.

More on the covariance matrix

- {(20, 31), (26, 40), (27, 45), (28, 52), (29, 60), (31, 70), (32, 71), (33, 69), (34, 72), (36, 85)}
- Each point denoted as (x_1, x_2)
- S = $\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$

$$S_{11} = \mathrm{E}[(x_1 - \mathrm{E}[x_1])(x_1 - \mathrm{E}[x_1])]$$

$$S_{12} = \mathrm{E}[(x_1 - \mathrm{E}[x_1])(x_2 - \mathrm{E}[x_2])]$$

$$S_{21} = \mathrm{E}[(x_2 - \mathrm{E}[x_2])(x_1 - \mathrm{E}[x_1])]$$

$$S_{22} = \mathrm{E}[(x_2 - \mathrm{E}[x_2])(x_2 - \mathrm{E}[x_2])]$$

Mahalanobis distance

- $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$: mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$
- $\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{y}$
- Considering the 2-dimensional case

$$\mathrm{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \qquad \mathrm{inv(S)} = \frac{1}{S_{11}S_{22}-S_{12}S_{21}} \begin{bmatrix} S_{22} & -S_{12} \\ -S_{21} & S_{11} \end{bmatrix}$$

- For notation simplicity, assume $\mathrm{A}=\mathrm{S}^{-1}$
- $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{A}(\boldsymbol{x}-\boldsymbol{\mu})}$

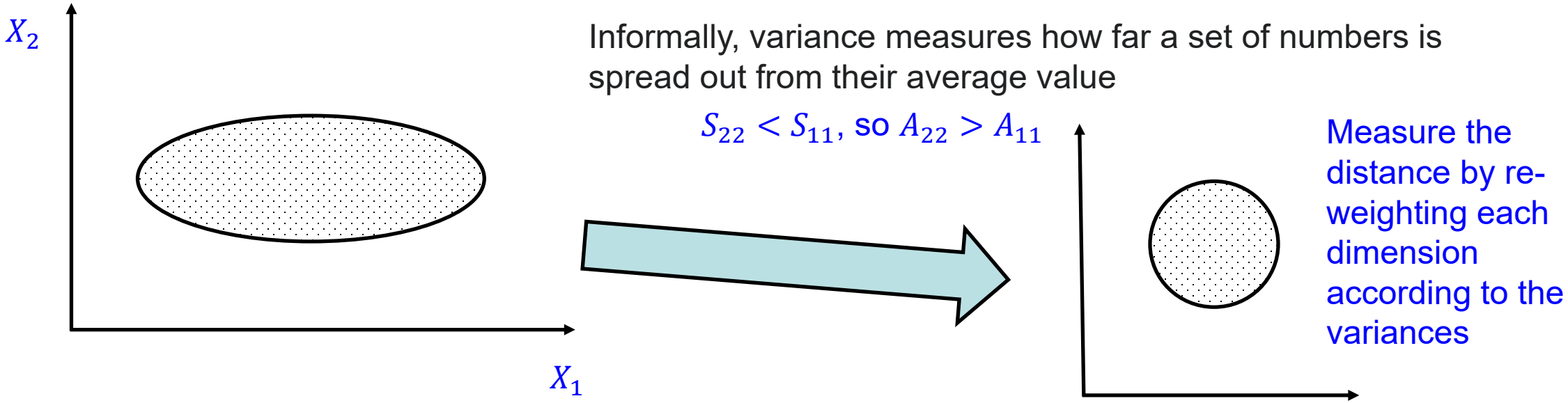
$$= \sqrt{([x_1, x_2] - [\mu_1, \mu_2]) \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}} = \sqrt{A_{11}(x_1-\mu_1)^2 + A_{22}(x_2-\mu_2)^2 + (A_{12}+A_{21})(x_1-\mu_1)(x_2-\mu_2)}$$

Mahalanobis distance-con't

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \qquad \text{Inv}(S) = \frac{1}{S_{11}S_{22} - S_{12}S_{21}} \begin{bmatrix} S_{22} & -S_{12} \\ -S_{21} & S_{11} \end{bmatrix} = A$$

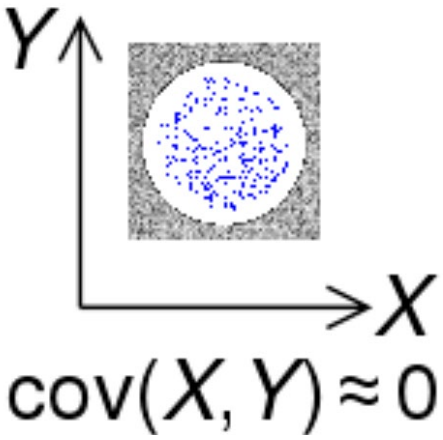
$$\sqrt{A_{11}(x_1 - \mu_1)^2 + A_{22}(x_2 - \mu_2)^2 + (A_{12} + A_{21})(x_1 - \mu_1)(x_2 - \mu_2)}$$

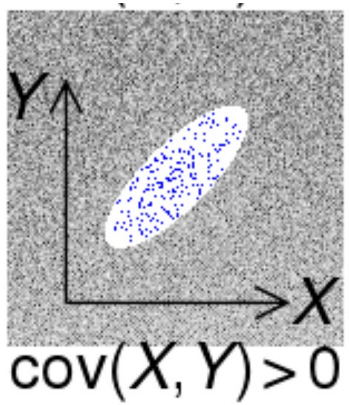
- If S is an identity matrix: $\sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}$
- If S is a diagonal matrix: $\sqrt{A_{11}(x_1 - \mu_1)^2 + A_{22}(x_2 - \mu_2)^2}$

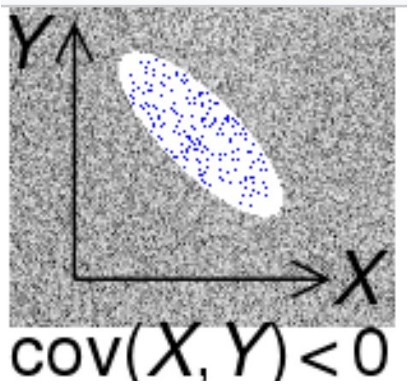


More on the covariance

- For a group of two dimensional data (x_i, y_i)





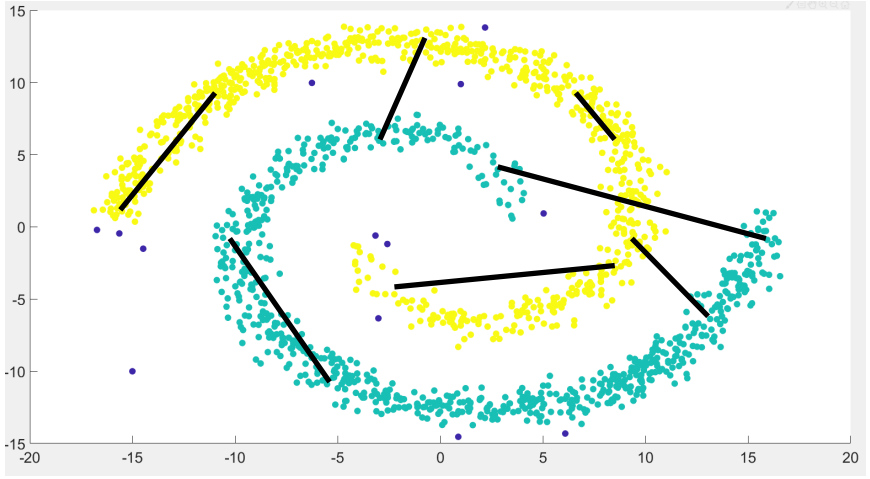


$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \qquad S = \frac{1}{S_{11}S_{22} - S_{12}S_{21}} \begin{bmatrix} S_{22} & -S_{12} \\ -S_{21} & S_{11} \end{bmatrix} = A$$

$$\sqrt{A_{11}(x_1 - \mu_1)^2 + A_{22}(x_2 - \mu_2)^2 + (A_{12} + A_{21})(x_1 - \mu_1)(x_2 - \mu_2)}$$

Mahalanobis distance: application

- Distance metric learning (informally)
  - Sometimes we are afraid the existing distance functions such as Euclidean distance cannot satisfy our needs:



  - If we know some information of which points should be in the same cluster and which points are not, we learn the matrix M in

$$\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\mathrm{M}(\boldsymbol{x}-\boldsymbol{y})}$$

Lecture 7&8: Classification 1

Classification

- Given a collection of records (or called samples)
- Each record (or sample) contains:
  - A object described by a set of attributes (or features/input variables)
  - A label (or class/output variable)
- Find a model, such that the label is a function of the values of all other attributes
- Training set: the collection of records

The goal of classification

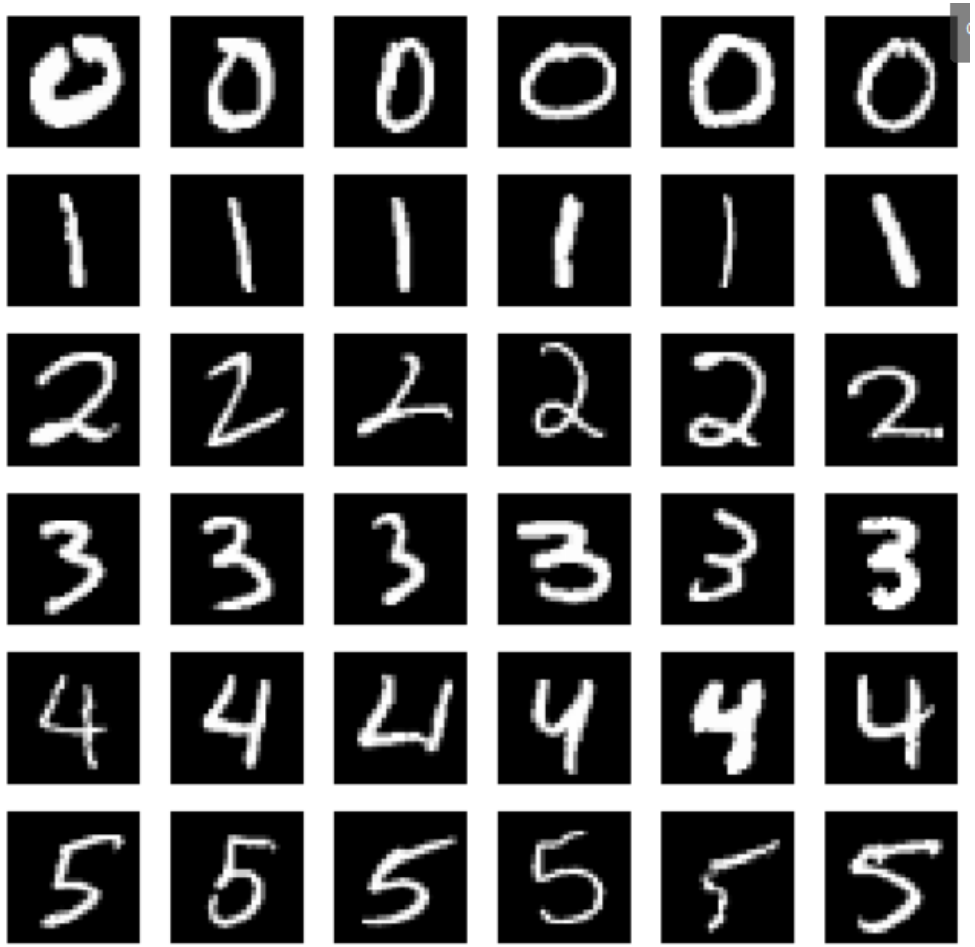
Learn a function from the **training set** as a mapping from the input variables to the labels, thus for previously unseen records (called: **test set**), it should be assigned a label as accurate as possible

Supervised learning

- Label: supervised information
- Thus classification is also called “supervised learning”
- For the same reason, “clustering” is sometimes called “unsupervised learning”

An example of supervised learning

5	0	0	0	0	0	0
4	0	0	0	0	0	0
0	0	0	0	0	0	0
4	0	0	0	0	0	0
0	0	0	0	0	0	0
1	0	0	0	0	0	0
3	0	0	0	0	0	0
1	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
2	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
4	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
5	0	0	0	0	0	0
1	0	0	0	0	0	0



Topics we plan to cover

- Why classification: applications
- What:
  - Pre-processing: normalization, curse of dimensionality
  - Generalization and overfitting
  - Model evaluation and cross-validation
  - Pain of classification: ambiguous supervised information
- How:
  - K-NN
  - Naïve Bayesian
  - Decision Tree
  - Random Forest (a kind of ensemble method)
  - *Logistic regression/neural networks/SVM