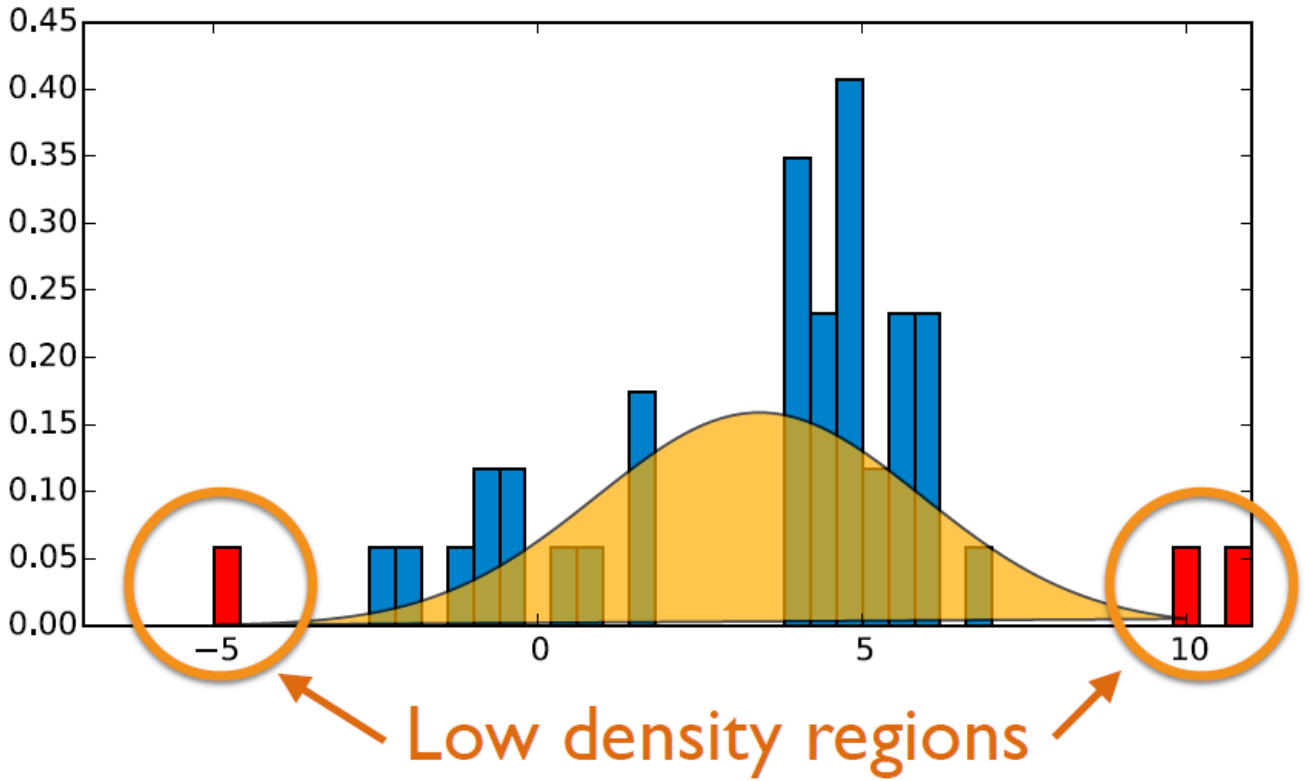
Anomaly Detection:
(Mixture of) Gaussian Density Estimation

Pilsung Kang
School of Industrial Management Engineering
Korea University

Density-based Novelty Detection

Gramfort (2006)

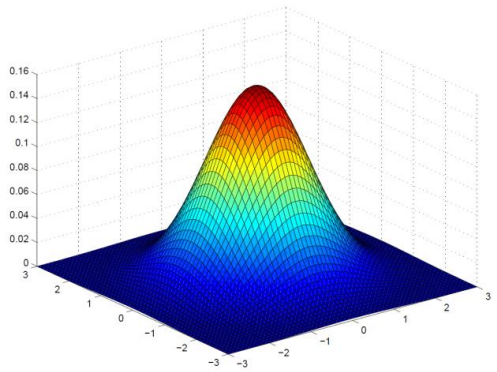
- Purpose
  - ✓ Estimate the data-driven density function
  - ✓ If a new instance has a low probability according the trained density function, it will be identified as novel



Density-based Novelty Detection

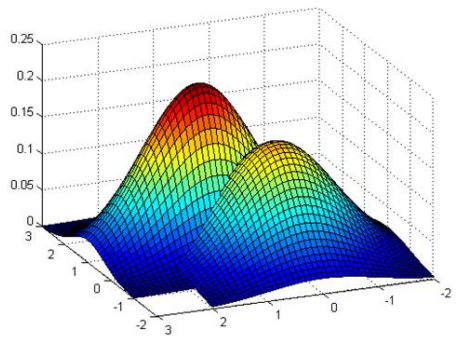
- Purpose
  - ✓ Estimate the data-driven density function
  - ✓ If a new instance has a low probability according the trained density function, it will be identified as novel

Gaussian Density Estimation



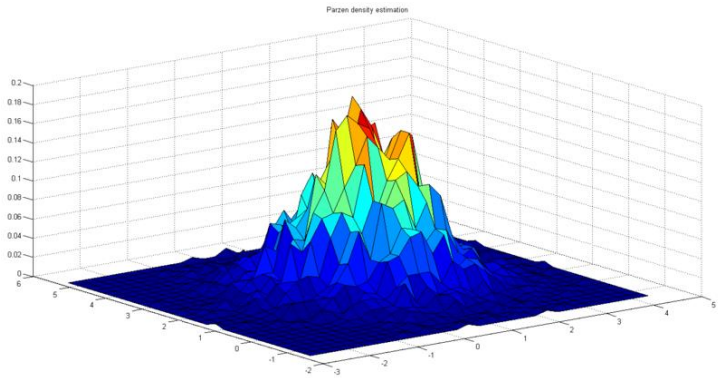
Number of modals = 1

Mixture of Gaussian Density Estimation



1 < Number of modals < Number of instances

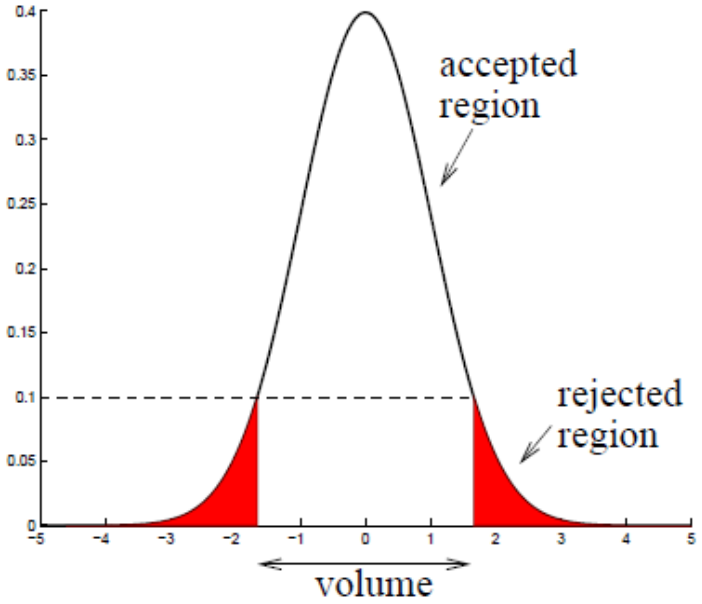
Kernel Density Estimation

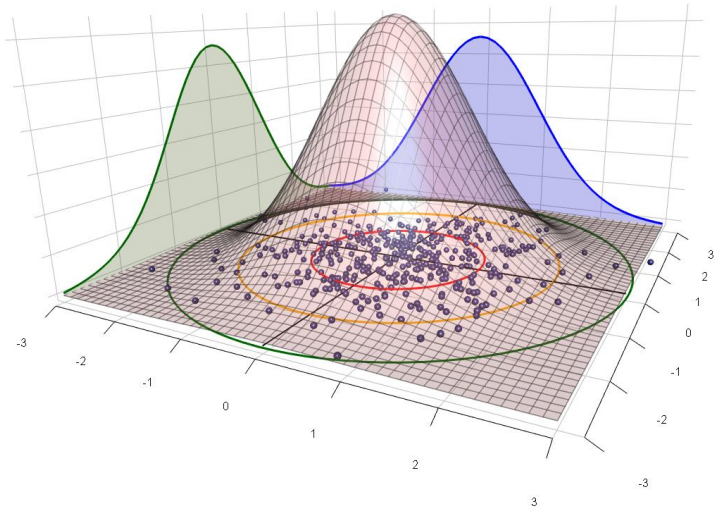


Number of modals = Number of instances

Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ Assume that the observed data are drawn from a Gaussian distribution





$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

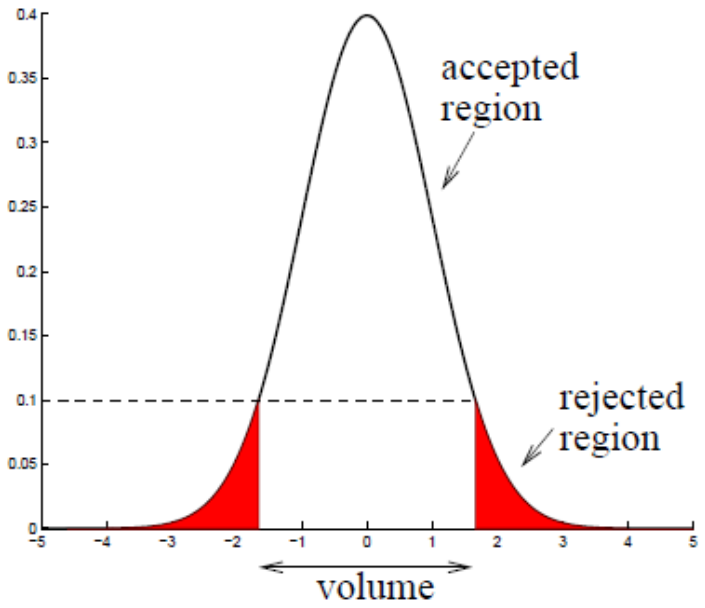
$$\boldsymbol{\mu} = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathbf{X}^+} \mathbf{x}_i \text{ (mean vector)}, \quad \boldsymbol{\Sigma} = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathbf{X}^+} (\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T \text{ (covariance matrix)}$$

Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ Advantages
    - Insensitive to scaling of the data

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

    - Possible to compute analytically the optimal threshold



Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ Parameter estimation: μ and σ^2
  - ✓ For one-dimensional data,

$$L = \prod_{i=1}^{N} P(x_i|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$Log\ L = -\frac{1}{2}\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} - \frac{N}{2}\log(2\pi\sigma^2)$$

Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ Maximum likelihood estimation, let's set $\gamma = 1/\sigma^2$

$$Log\ L = -\frac{1}{2}\sum_{i=1}^{N}\gamma(x_i-\mu)^2 - \frac{N}{2}\log(2\pi) + \frac{N}{2}\log(\gamma)$$

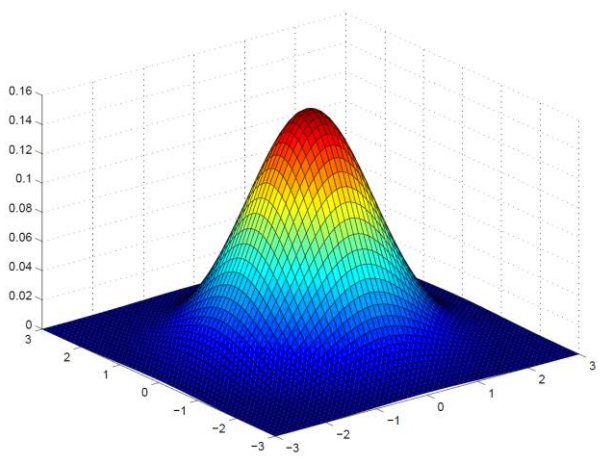
$$\frac{\partial Log\ L}{\partial \mu} = \gamma\sum_{i=1}^{N}(x_i-\mu) = 0 \rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N}x_i$$

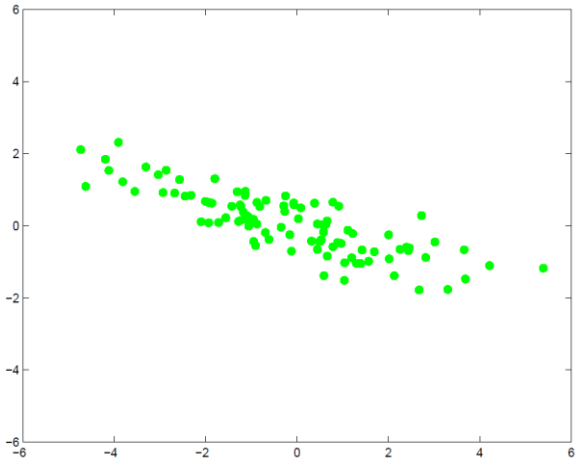
$$\frac{\partial Log\ L}{\partial \gamma} = -\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{N}{2\gamma} = 0 \rightarrow \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

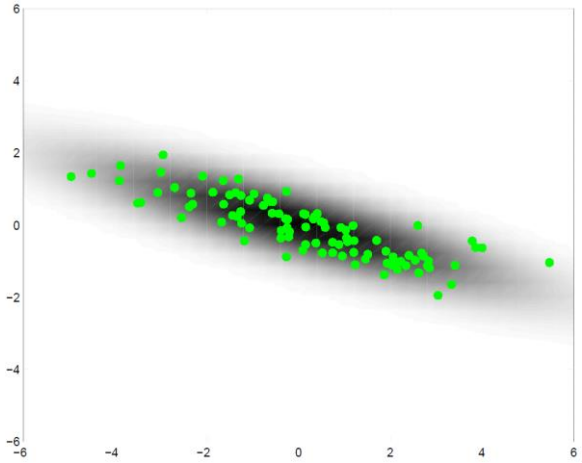
Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ In general,

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \quad \boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T$$



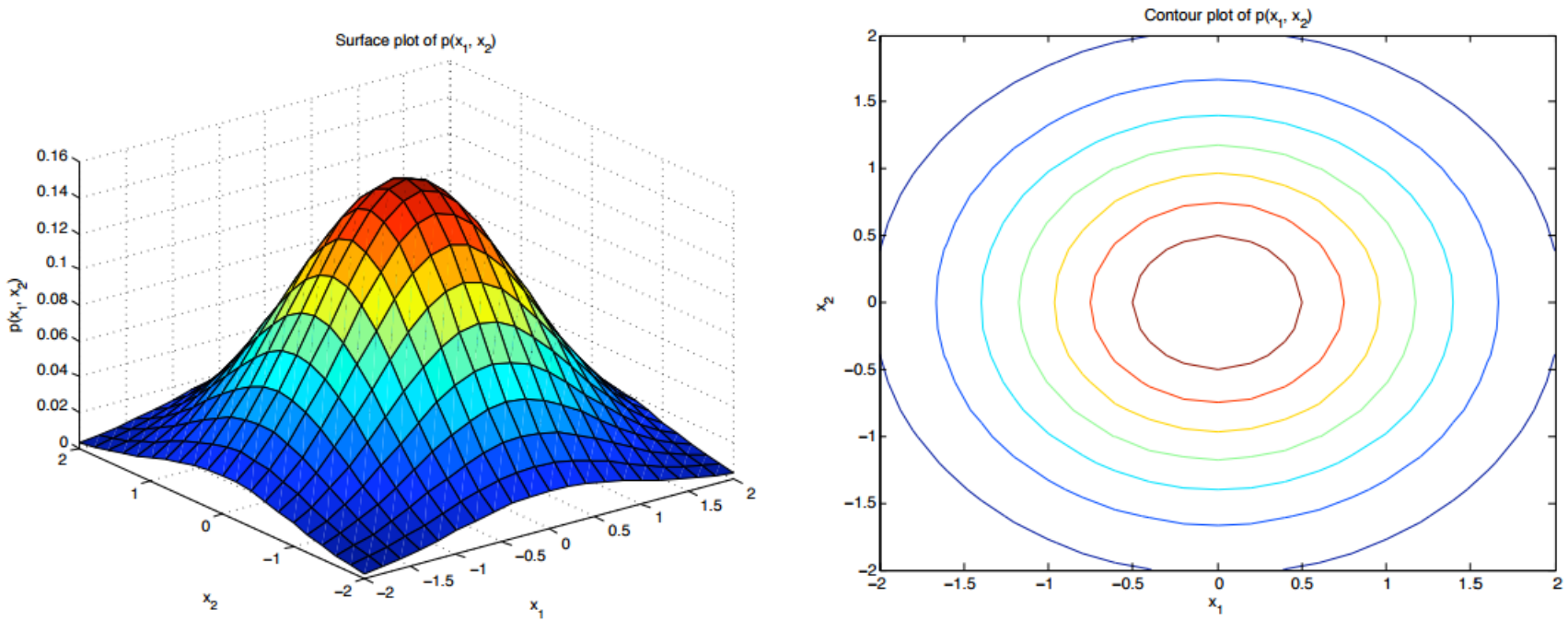




Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ The shape of Gaussian distribution according to the Covariance matrix type

Spherical $\quad \Sigma = \sigma^2 \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$

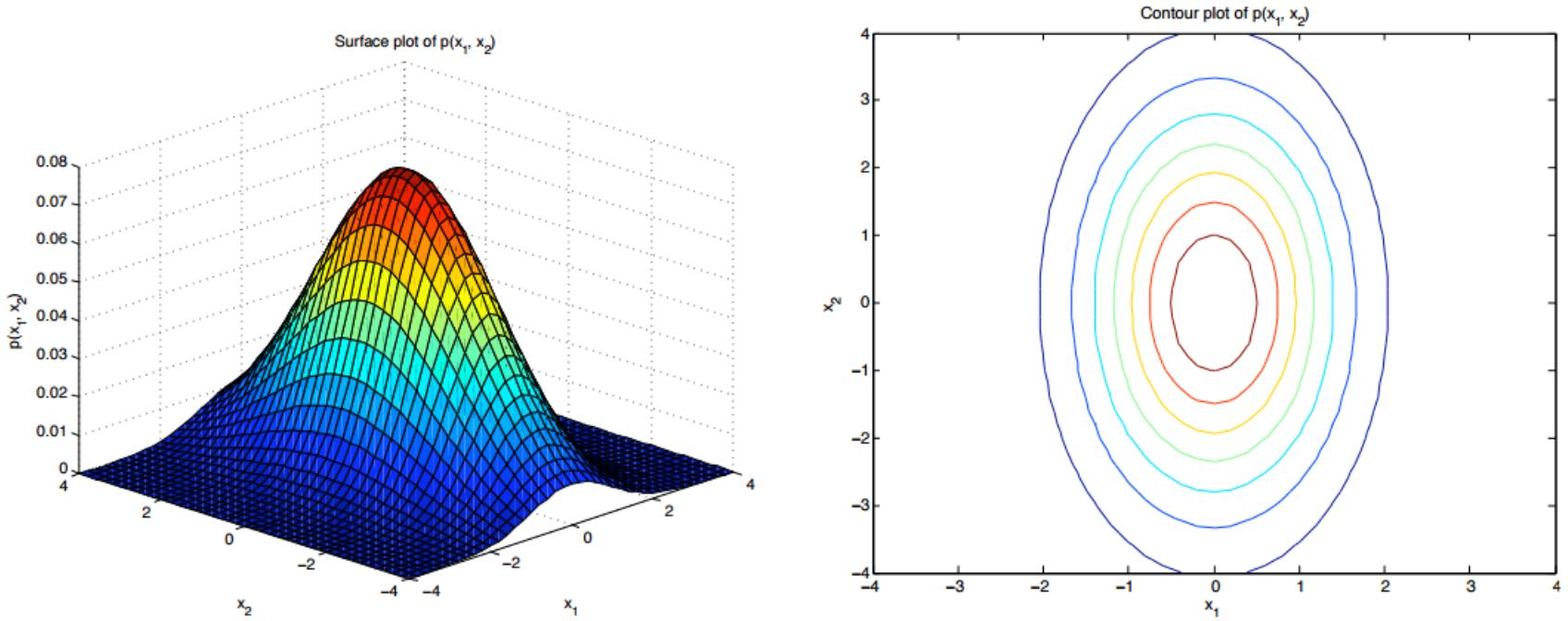


(a) Spherical Gaussian (diagonal covariance, equal variances)

Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ The shape of Gaussian distribution according to the Covariance matrix type

Diagonal $\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d^2 \end{bmatrix}$

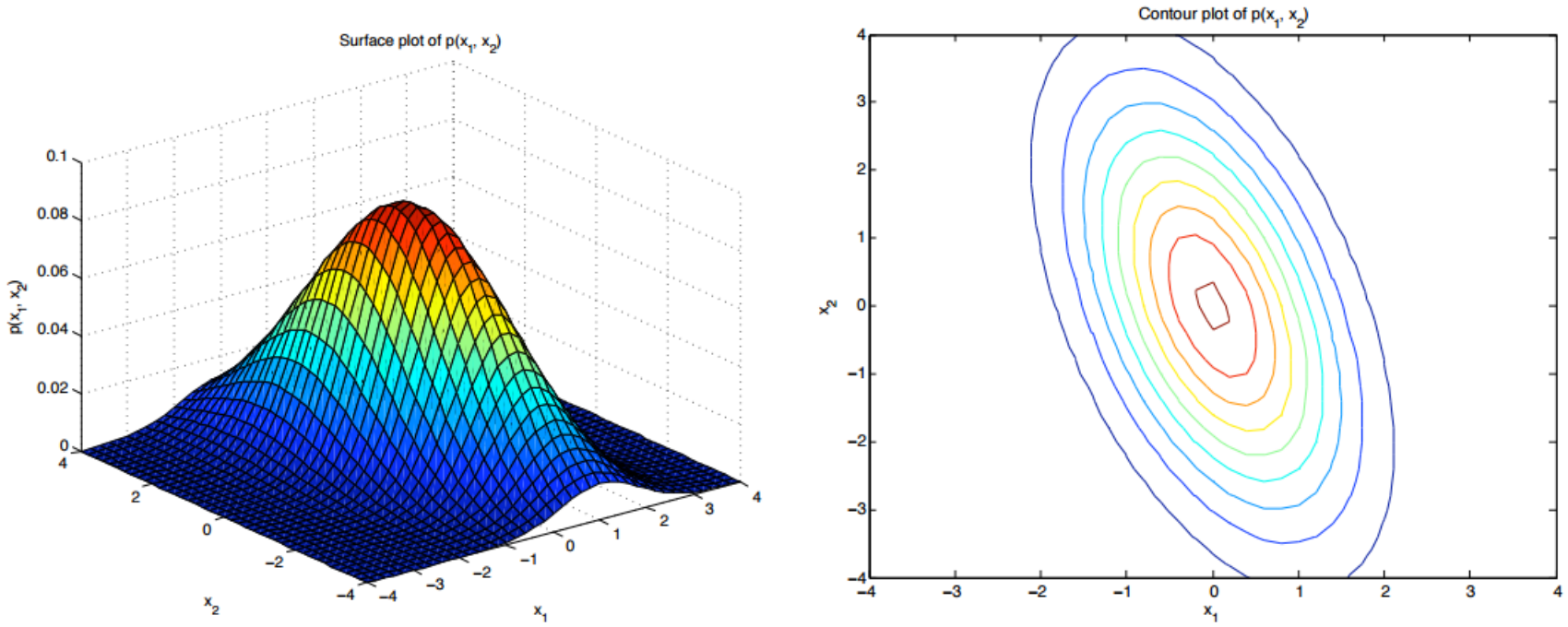


(b) Gaussian with diagonal covariance matrix

Gaussian Density Estimation

- Gaussian Density Estimation
  - ✓ The shape of Gaussian distribution according to the Covariance matrix type

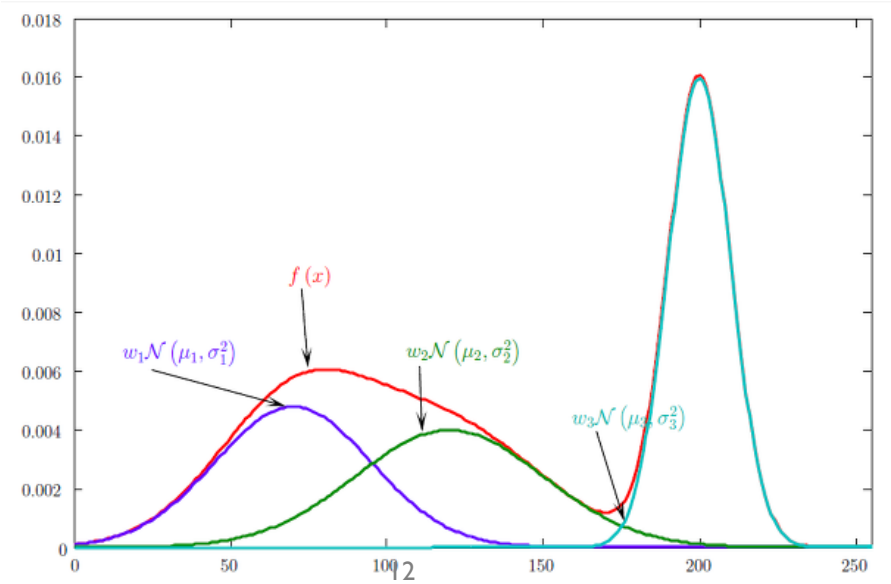
Full $\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$



(c) Gaussian with full covariance matrix

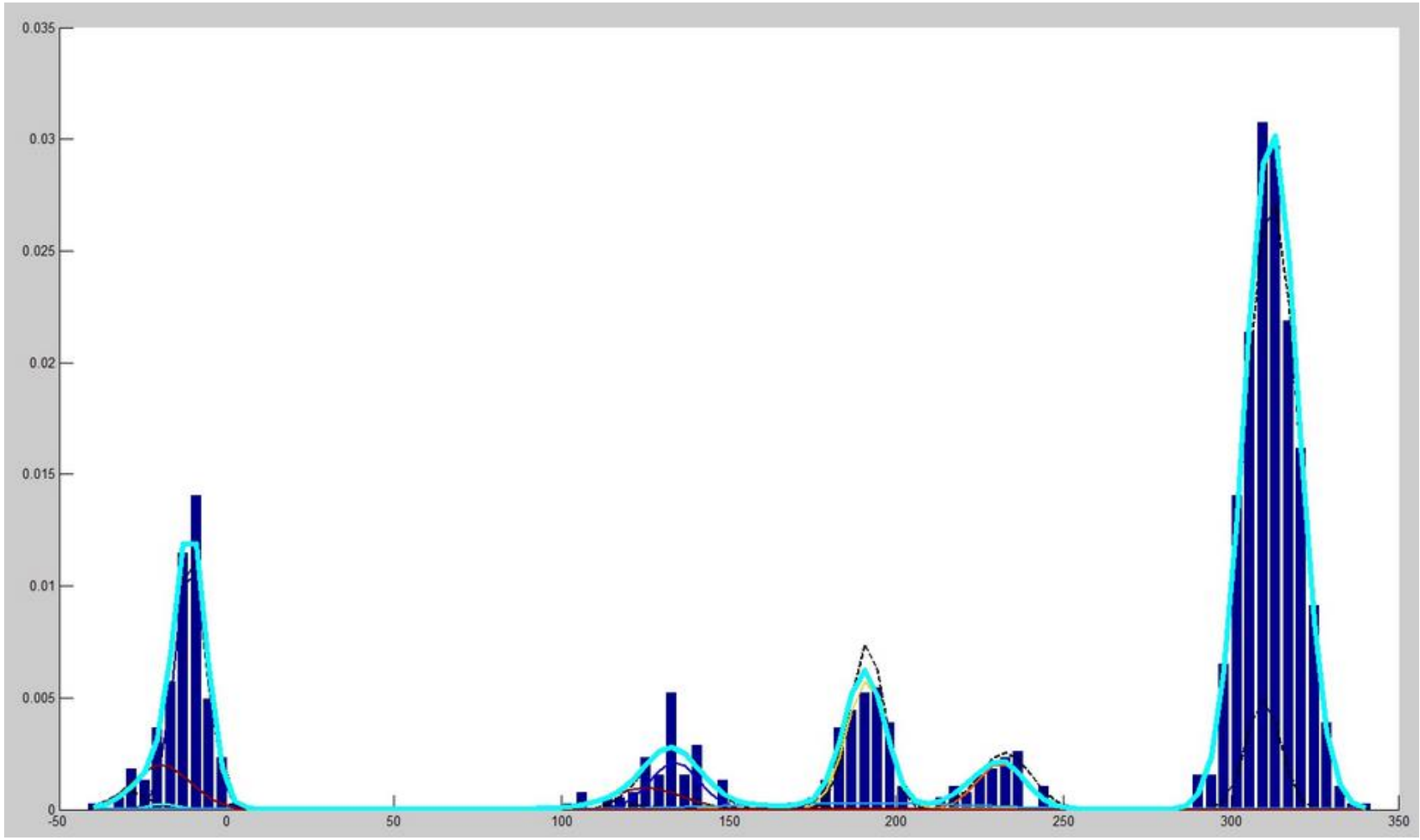
Mixture of Gaussian Density Estimation

- Mixture of Gaussian (MoG) Density Estimation
  - ✓ Gaussian Density Estimation
    - assumes a very strong model of the data: unimodal and convex
  - ✓ MoG
    - an extension of Gaussian that allows multi-modal distribution
    - a linear combination of normal distributions
    - Has a smaller bias than the single Gaussian distribution, but requires far more data for training



Mixture of Gaussian Density Estimation

- MoG example



Mixture of Gaussian Density Estimation

- Components of MoG
  - ✓ Probability of an instance belonging to the normal class

$$p(\mathbf{x}|\lambda) = \sum_{m=1}^{M} w_m g(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

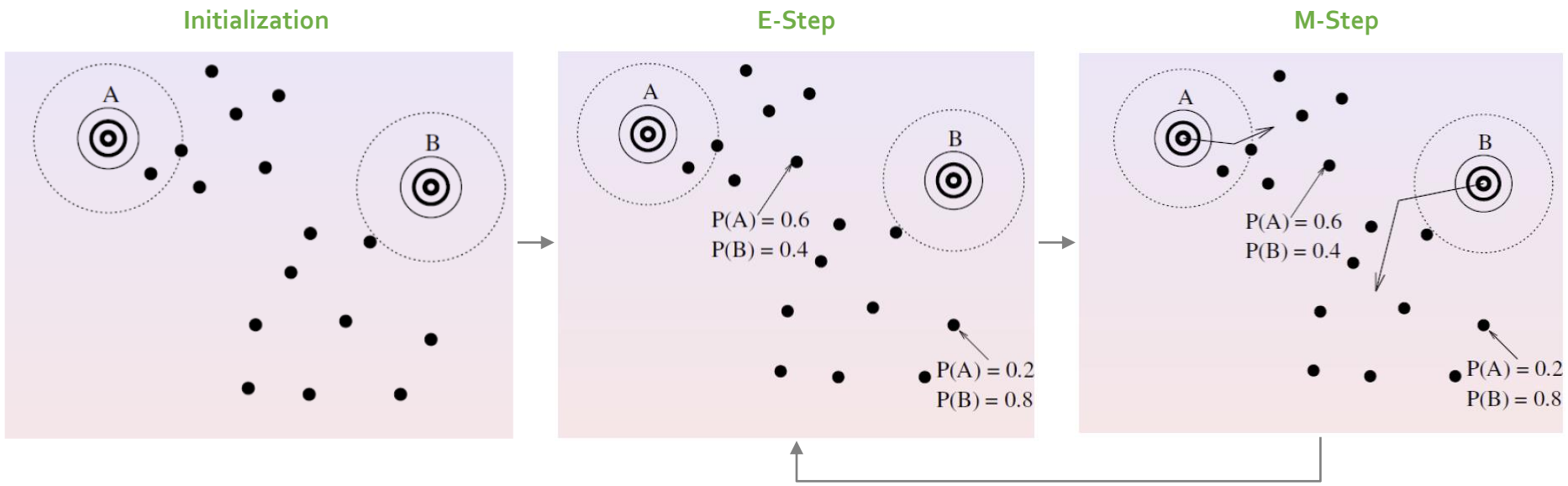
  - ✓ Distribution of each Gaussian model

$$g(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_m|^{1/2}} exp\left[\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right]$$

$$\lambda = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}, \quad m = 1, \cdots, M$$

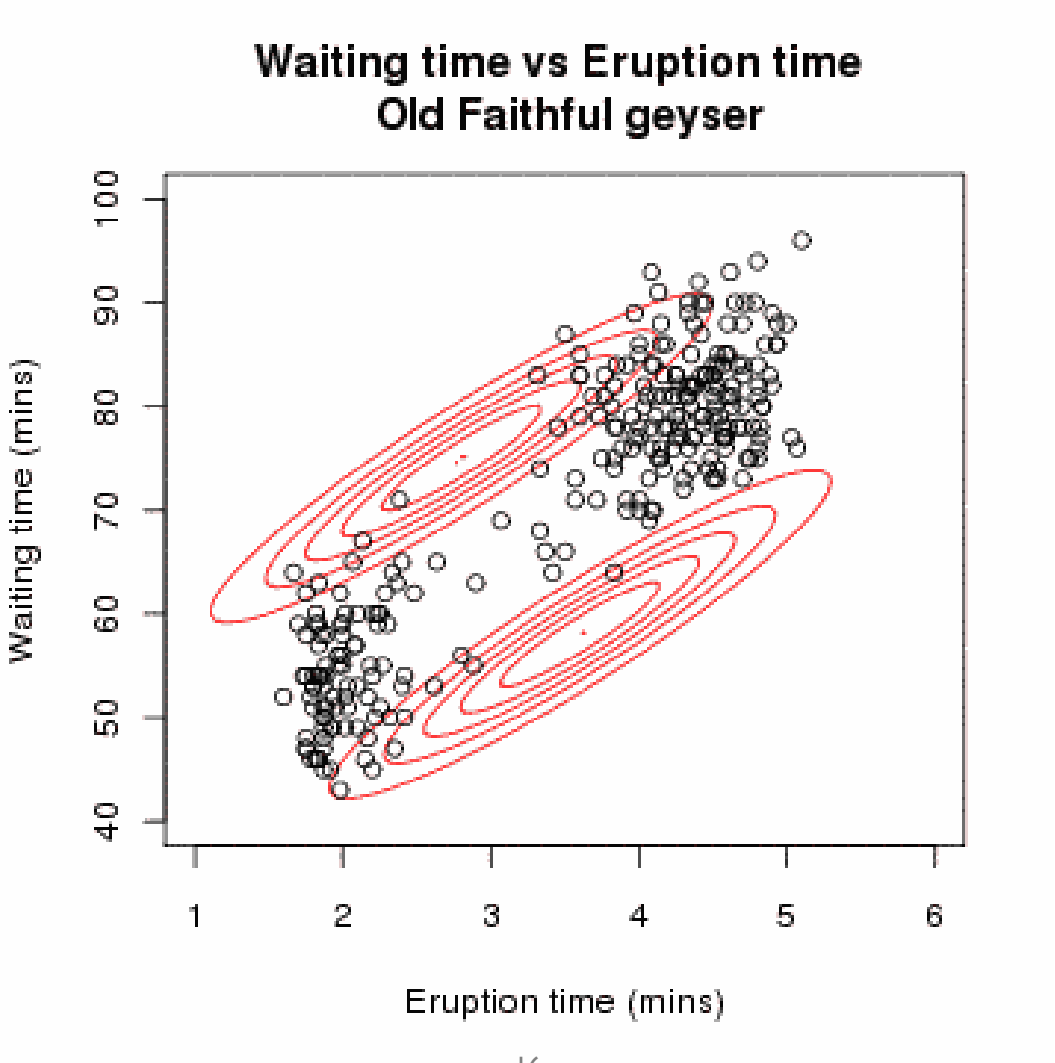
Mixture of Gaussian Density Estimation

- Expectation-Maximization Algorithm
  - ✓ E-Step: Given the current estimate of the parameters, compute the conditional probabilities
  - ✓ M-Step: Update the parameters to maximize the expected likelihood found in the E-Step



Mixture of Gaussian Density Estimation

- Expectation-Maximization Algorithm: Illustrative example



Mixture of Gaussian Density Estimation

- EM algorithm for MoG
  - ✓ Expectation

$$p(m|\mathbf{x}_i, \lambda) = \frac{w_m g(\mathbf{x}_i|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{k=1}^{M} w_k g(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

  - ✓ Maximization
    - Mixture weight

$$w_m^{(new)} = \frac{1}{N} \sum_{i=1}^{N} p(m|\mathbf{x}_i, \lambda)$$

    - Means and variances

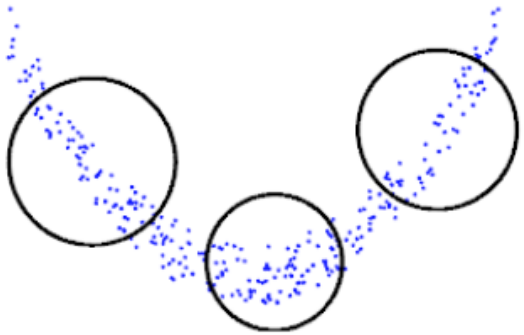
$$\boldsymbol{\mu}_m^{(new)} = \frac{\sum_{i=1}^{N} p(m|\mathbf{x}_i, \lambda)\mathbf{x}_i}{\sum_{i=1}^{N} p(m|\mathbf{x}_i, \lambda)}, \quad \sigma_m^{2(new)} = \frac{\sum_{i=1}^{N} p(m|\mathbf{x}_i, \lambda)\mathbf{x}_i^2}{\sum_{i=1}^{N} p(m|\mathbf{x}_i, \lambda)} - \boldsymbol{\mu}_m^{2(new)}$$

Mixture of Gaussian Density Estimation

- The shape of MoG according to the covariance matrix

Spherical

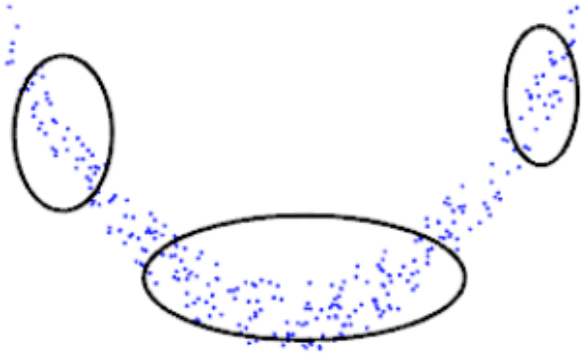
$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$



- Less precise
- Very efficient to compute

Diagonal

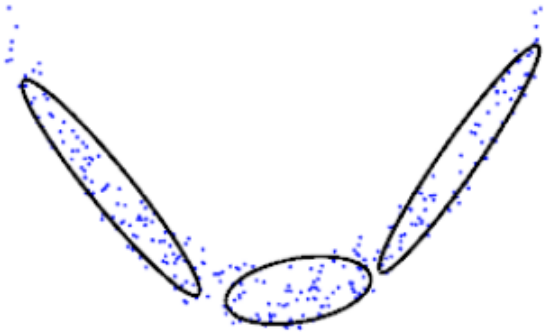
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_d^2 \end{bmatrix}$$



- More precise
- Efficient to compute

Full

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$



- Very precise
- Less efficient to compute


Q & A

References

Research Papers

- Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. (2000). LOF: Identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data: 93-104.
- Chalapathy, R., Menon, A. K., & Chawla, S. (2018). Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys 41(3): 15.
- Harmeling, S. Dornhege, G., Tax, D. Meinecke, F., and Muller, K-R. (2006). From outliers to prototype: Ordering data. Neurocomputing 69(13-15): 1608-1618.
- Hariri, S., Kind, M. C., & Brunner, R. J. (2018). Extended Isolation Forest. arXiv preprint arXiv:1811.02141.
- Kang, P. and Cho, S. (2009). A hybrid novelty score and its use in keystroke dynamics-based user authentication. Pattern Recognition 42(11): 3115-3127.
- Liu, F.T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 413-422). IEEE.
- Liu, F.T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 3.
- Oza, P., & Patel, V. M. (2018). One-class convolutional neural network. IEEE Signal Processing Letters, 26(2), 277-281.
- Perera, P., Nallapati, R., & Xiang, B. (2019). Ocgan: One-class novelty detection using gans with constrained latent representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2898-2906).

References

Research Papers

- Pimentel, M.A.F., Clifton, D.A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection, Signal Processing 99: 215-249.
- Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017, June). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In International Conference on Information Processing in Medical Imaging (pp. 146-157). Springer, Cham.
- Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R. C.(2001). Estimating the support of a high-dimensional distribution. Neural computation 13(7): 1443-1471.
- Tax, D.M. (2001). One-class classification, Ph.D. Thesis, Delft University of Technology, Netherlands.
- Tax, D.M. and Duin, R.P. (2004). Support vector data description. Machine learning 54(1): 45-66.

Other materials

- Pages 28-33 & 36: http://research.cs.tamu.edu/prism/lectures/pr/pr_l7.pdf
- Figures in Auto-encoder section: https://dl.dropboxusercontent.com/u/19557502/6_01_definition.pdf
- Gramfort, A. (2016). Anomaly/Novelty detection with scikit-learn: https://www.slideshare.net/agramfort/anomalynovelty-detection-with-scikitlearn