Gaussian Mixture Models

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

June 5, 2018

Selayang Pandang

Bahan Bacaan

1. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. (Section 9.3)
2. VanderPlas, J. (2016). Python Data Science Handbook. (In Depth: Gaussian Mixture Models)
   `http://nbviewer.jupyter.org/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.12-Gaussian-Mixtures.ipynb`
3. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics. (Section 14.3.12)

Motivasi

Coba kelompokkan berita-berita berikut...

News Clustering

Antihero Sergio Ramos Berpotensi Membuatmu Jadi Moralis

Sergio Ramos memantik orang untuk bicara tentang moral, etika, dan sportivitas. Itulah arti penting antihero.

News Clustering

Debar dan Getar Jiwa Nabi Muhammad Kala Menerima Wahyu Pertama

Peristiwa turunnya wahyu pertama adalah momen paling menggetarkan dalam hidup Nabi Muhammad.

News Clustering

Mohamed Salah di Antara Pemain Muslim, Puasa, dan Liga Champions

Gambar: Agama? Olahraga?

Apakah sepakbola harus dibedakan dengan olahraga?
Bagaimana dengan fikih dan akidah?

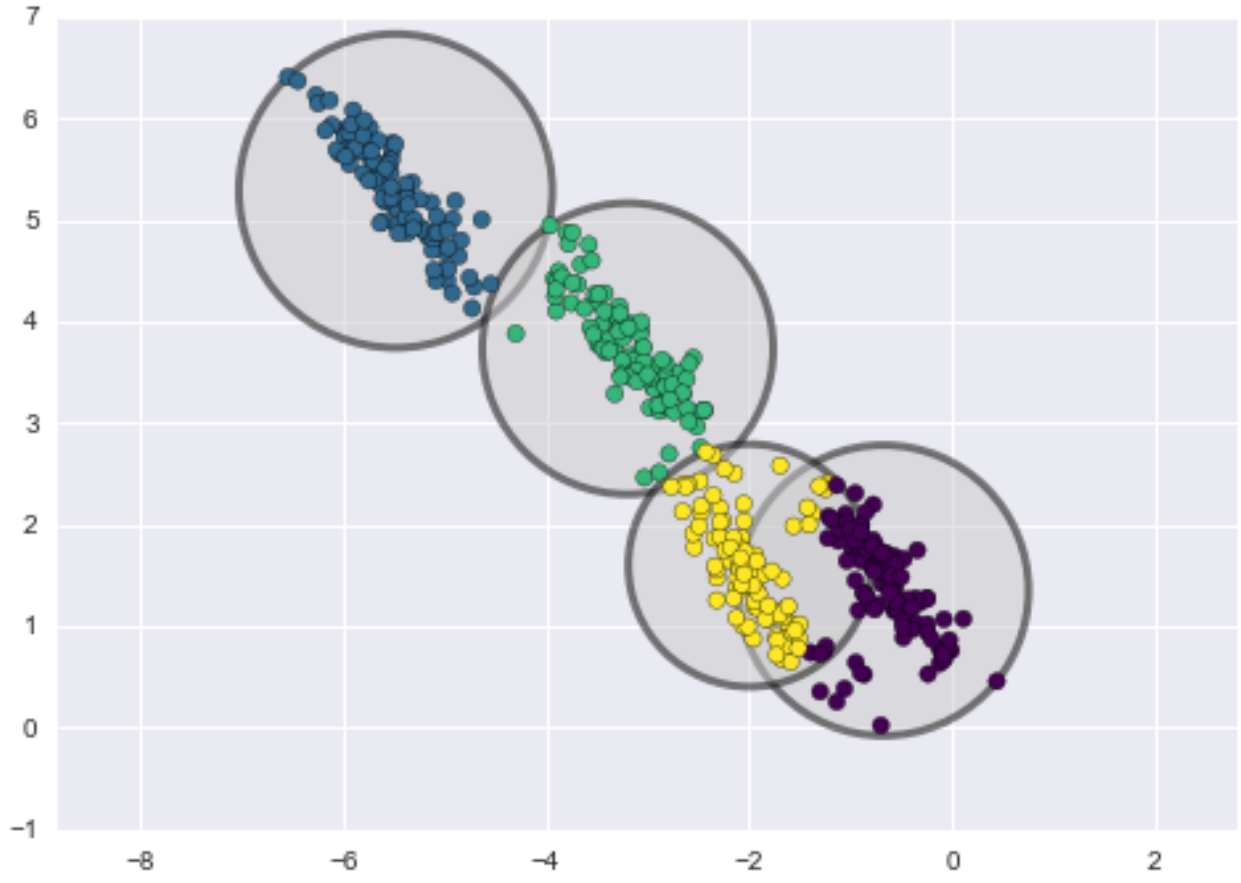
Jenis-jenis Clustering

1. Tujuan:
    1. Monothetic: *common property*
    2. Polythetic: kemiripan data dengan pengukuran jarak
2. Irisan:
    1. Hard clustering
    2. Soft clustering
3. Flat vs hierarchical

k-Means

Polythetic, hard boundaries, flat

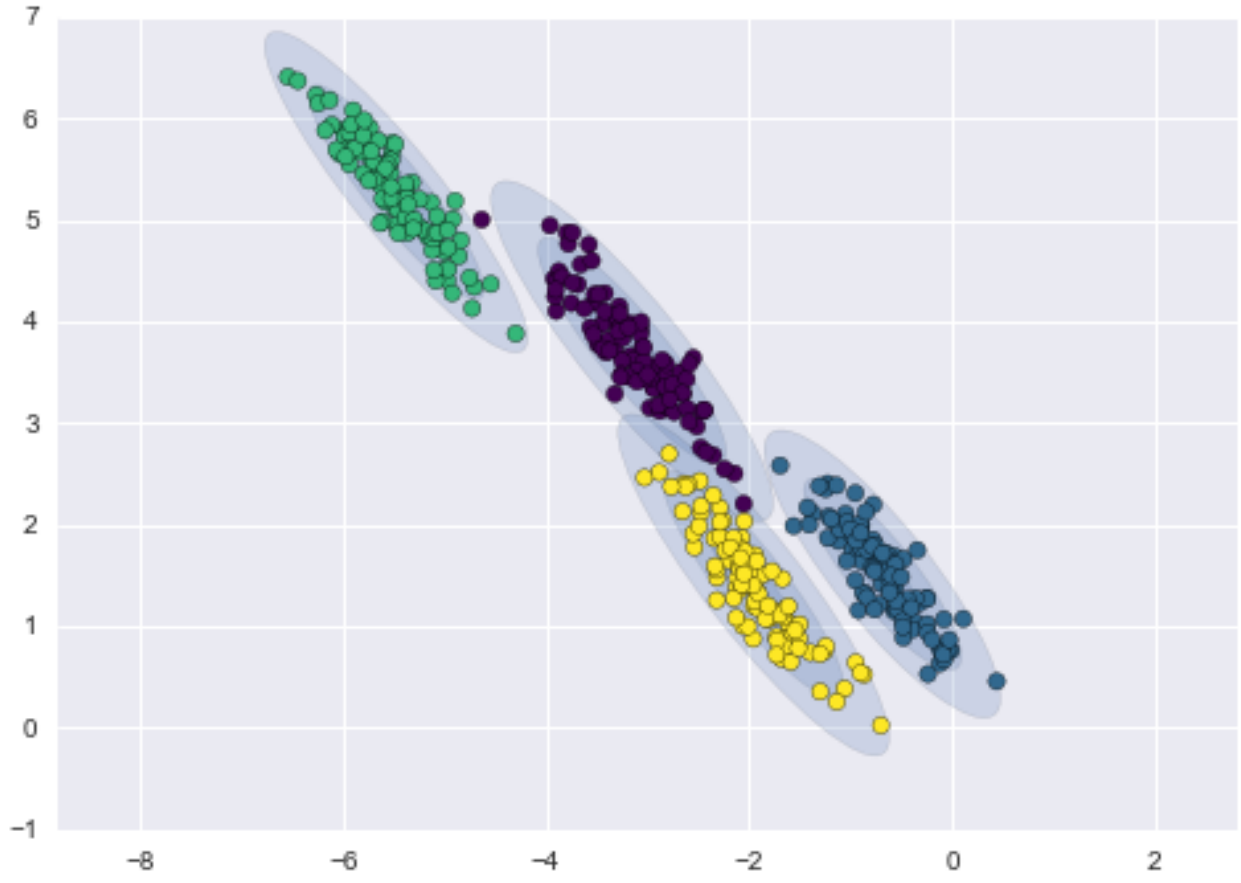
k-Means



Gaussian Mixture Models

Polythetic, soft boundaries, flat

GMM

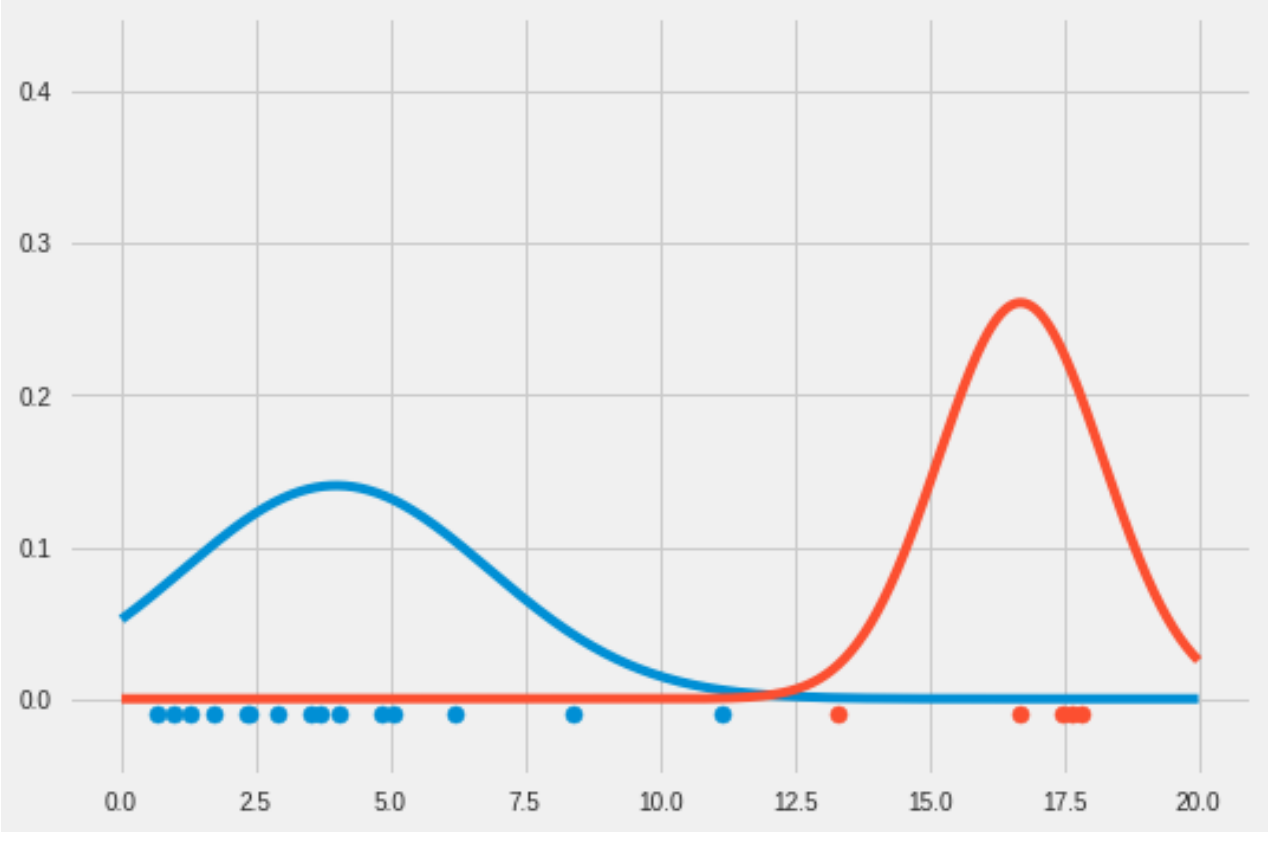


Gaussian Mixture Models

Mixture Models

- Pendekatan probabilistik untuk *clustering*
- Setiap klaster adalah model generatif, e.g. Gaussian atau multinomial
- Menggunakan parameter
- Didasarkan pada algoritma Expectation Maximisation (EM)

Mixture Models 1D



Bagaimana kalau kita tidak tahu kelasnya?

Expectation Maximisation (EM)

1. Inisialisasi dengan dua Gaussians secara acak (μ_a, σ_a^2), (μ_b, σ_b^2)
2. Ulangi hingga konvergen
   a. **E-step**: Apakah x_i terlihat masuk ke a atau b, i.e. $P(a|x_i)$?[1]

$$a_i = P(a|x_i) = \frac{P(x_i|a)P(a)}{P(x_i)}$$

$$b_i = P(b|x_i) = 1 - a_i$$

   b. **M-step**: Perbaiki nilai (μ_a, σ_a^2), (μ_b, σ_b^2)

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \ldots + a_n x_n}{a_1 + a_2 + \ldots + a_n}$$

$$\sigma_a^2 = \frac{a_1 (x_1 - \mu_a)^2 + \ldots + a_n (x_n - \mu_a)^2}{a_1 + a_2 + \ldots + a_n}$$

[1] Bayes' rule!

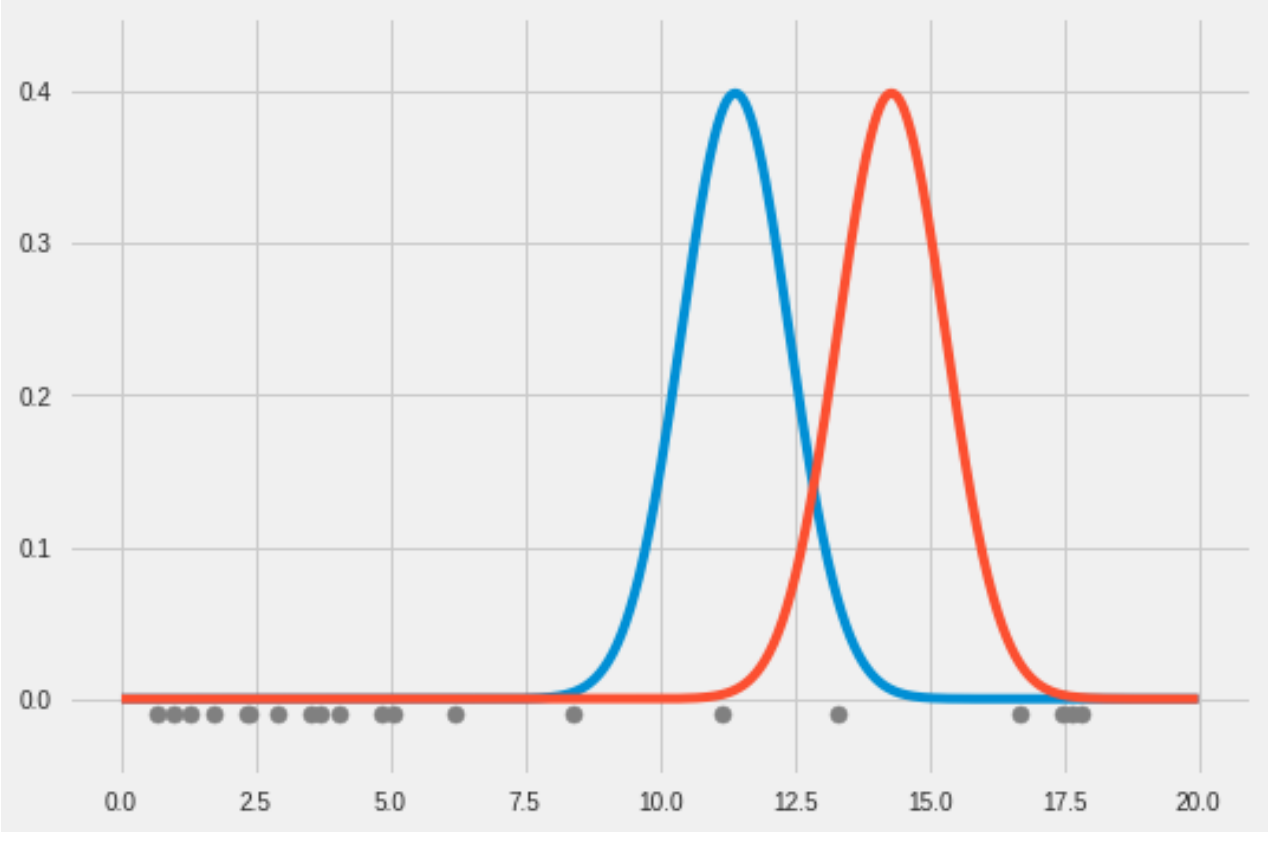
Prior dari Bayes' Rule

- Bisa dibuat tetap, atau
- Dibuat berubah-ubah, i.e.

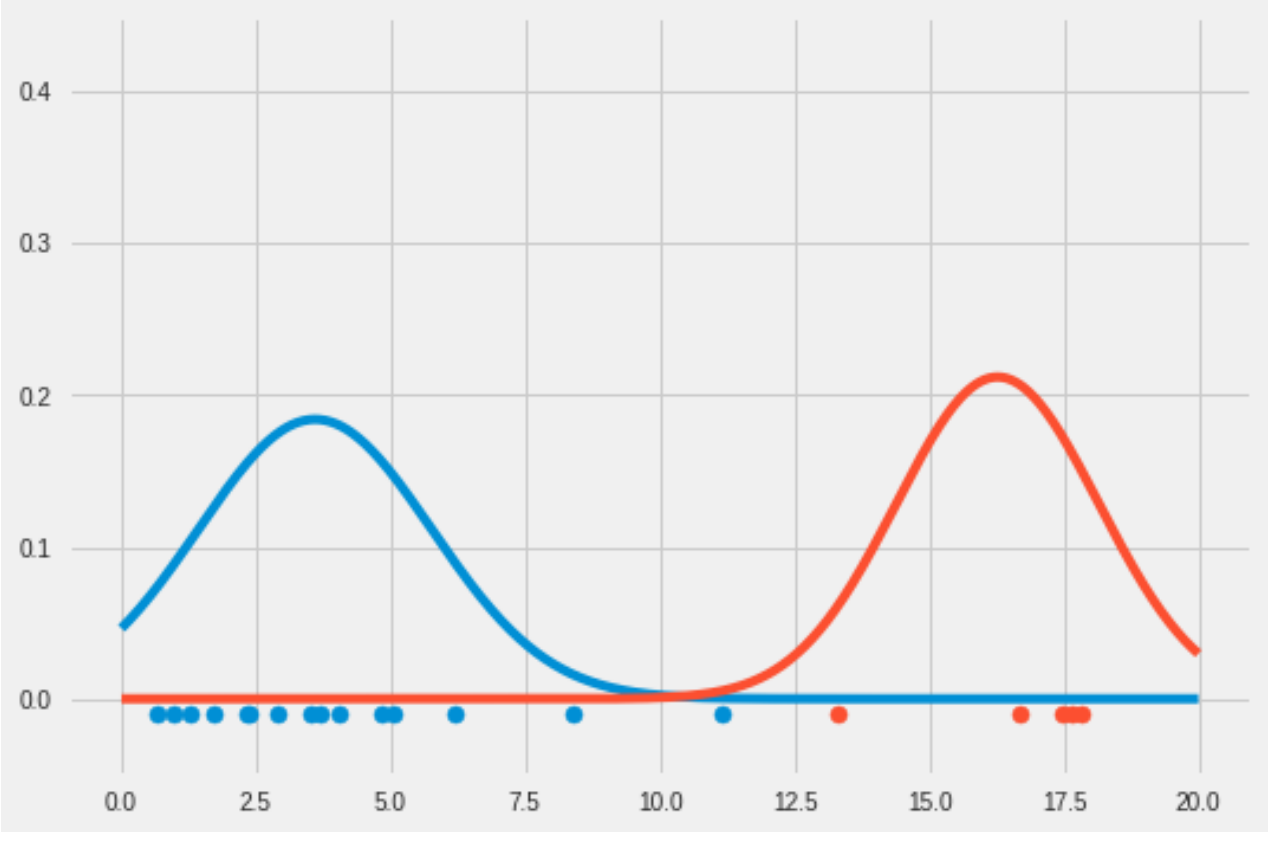
$$P(a) = \frac{a_1 + a_2 + ... + a_n}{n}$$

$$P(b) = 1 - P(a)$$

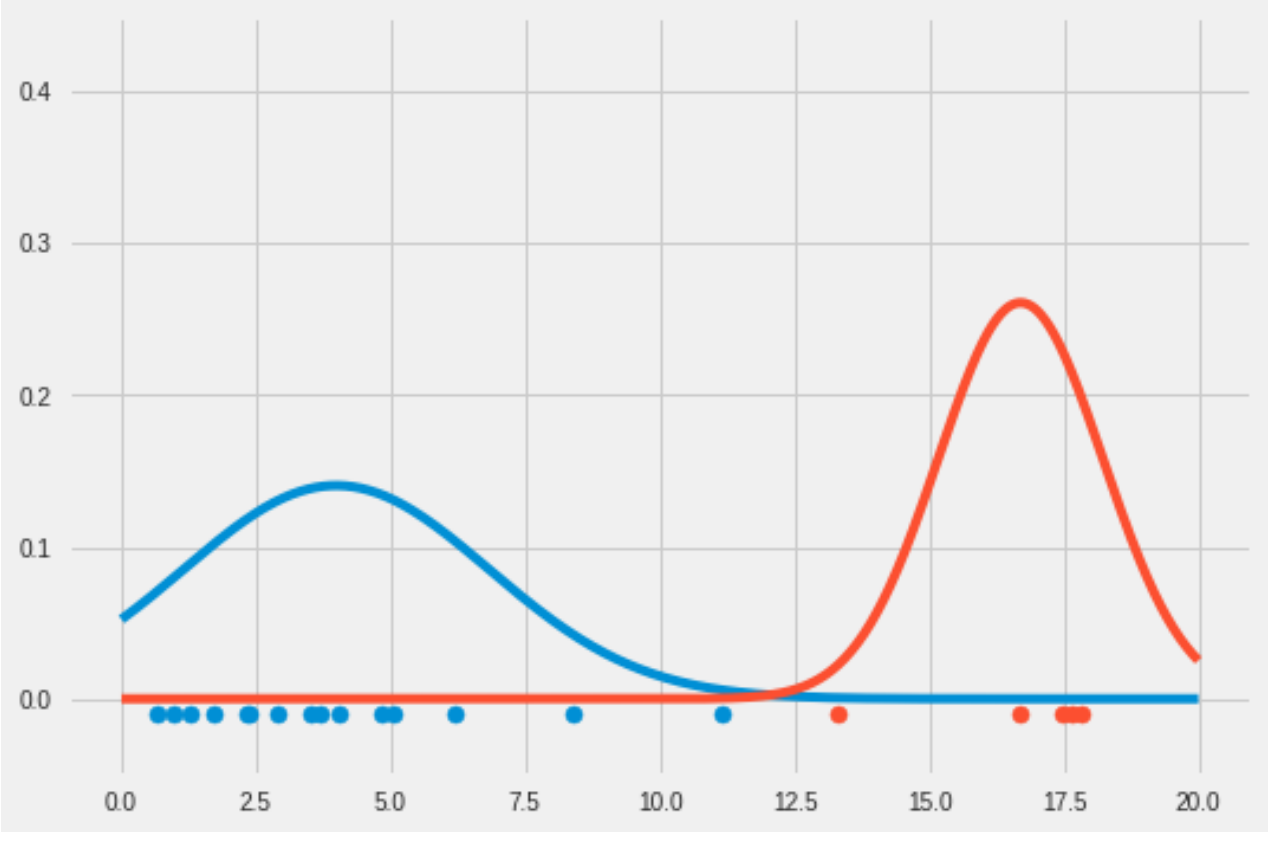
Mixture Models 1D



Mixture Models 1D



Mixture Models 1D



Berapa nilai K?

- Model probabilistik $\rightarrow$ *maximum likelihood*

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(x_i|k)P(k)$$

$$\mathcal{L} = \log P(x_i, ..., x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i|k)P(k)$$

Berapa nilai K?

- Model probabilistik $\rightarrow$ *maximum likelihood*

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(x_i|k)P(k)$$

$$\mathcal{L} = \log P(x_i, ..., x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i|k)P(k)$$

- $\mathcal{L}$ bisa dimaksimalkan dengan membuat $K = n \rightarrow$ *overfitting*!

Berapa nilai K?

- Model probabilistik $\rightarrow$ *maximum likelihood*

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(x_i|k)P(k)$$

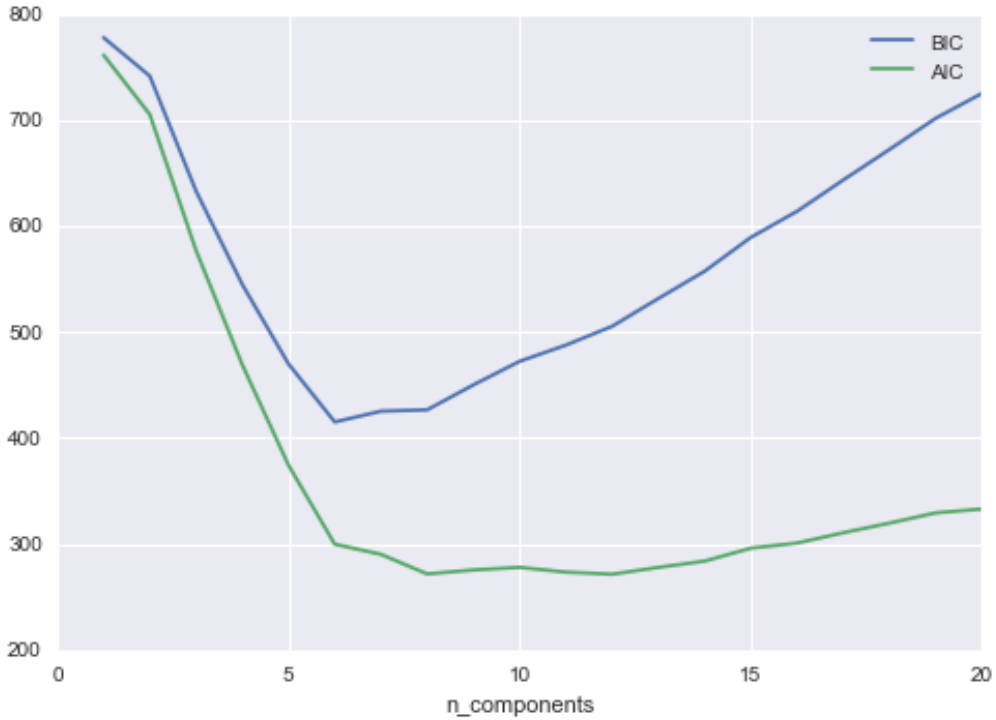
$$\mathcal{L} = \log P(x_i, ..., x_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} P(x_i|k)P(k)$$

- $\mathcal{L}$ bisa dimaksimalkan dengan membuat $K = n \rightarrow$ *overfitting*!
- Occam's razor
  - Bayes. Inf Criterion (BIC): $\max_p(\mathcal{L} - \frac{1}{2} p \log n)$
  - Akaike Inf Criterion (AIC): $\min_p(2p - \mathcal{L})$

  dengan $\mathcal{L}$ adalah *log likelihood* dan p adalah jumlah parameter

Tenang, sudah ada di scikit-learn!

AIC dan BIC



Gambar: Nilai terbaik adalah saat `n_components` antara 8-12 [VanderPlas, 2016]

Ikhtisar

1. Jenis-jenis clustering: tujuan, irisan, flat vs hierarchical
2. GMM adalah pendekatan probabilistik untuk *clustering*
3. Algoritma Expectation-Maximisation (EM)
4. Konsep AIC dan BIC

Pertemuan Berikutnya

1. Siapkan presentasi 5 menit per orang untuk penjelasan topik makalah
2. Konten: latar belakang, studi terkait, dan metode yang akan digunakan
3. Poin penting: Perbaiki studi literatur!

Referensi

Jake VanderPlas (2016)

In Depth: Gaussian Mixture Models

http://nbviewer.jupyter.org/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.12-Gaussian-Mixtures.ipynb

Terima kasih