14620323

# DEEP LEARNING

Mathematics for Machine Learning

# PENGAMPU

Dr. Fajar Astuti Hermawati, S.Kom.,M.Kom.

Elsen Ronando, S.Si.,M.Si

**Bagus Hardiansyah, S.Kom.,M.Si**

Andrey Kartika Widhy H., S.Kom., M.Kom.

# Capaian Pembelajaran

- Mampu mengidentifikasi konsep matematika dan mesin pembelajaran dasar untuk algoritma deep learning. [C2, A3]

# Bahan Kajian

- Linear Algebra
- **Probability and InformationTheory**
- Numerical Computation

# Probability and InformationTheory

# Why Probability?

- Pembelajaran mesin harus selalu berurusan dengan kuantitas yang tidak pasti dan terkadang kuantitas stokastik (nondeterministik).
  - Ketidakpastian dan stokastik dapat muncul dari banyak sumber.
  - Para peneliti telah membuat argumen yang meyakinkan untuk mengukur ketidakpastian menggunakan probabilitas setidaknya sejak tahun 1980-an.

# Why Probability?

- Dalam kasus dokter yang mendiagnosa pasien, kita menggunakan probabilitas untuk mewakili tingkat kepercayaan (***degree of belief***), dengan 1 menunjukkan kepastian mutlak bahwa pasien terkena flu dan 0 menunjukkan kepastian mutlak bahwa pasien tidak terkena flu.
  - Probabilitas yang pertama, terkait langsung dengan tingkat terjadinya peristiwa, dikenal sebagai ***frequentist probability***,
  - sedangkan yang kedua, terkait dengan tingkat kepastian kualitatif, dikenal sebagai **probabilitas Bayesian**.

# Why Probability?

- Probabilitas dapat dilihat sebagai perpanjangan dari logika untuk menghadapi ketidakpastian.
- Logika menyediakan seperangkat aturan formal untuk menentukan proposisi apa yang tersirat benar atau salah dengan asumsi bahwa beberapa perangkat proposisi lain benar atau salah.
- Teori probabilitas menyediakan seperangkat aturan formal untuk menentukan kemungkinan (likelihood) suatu proposisi menjadi benar mengingat kemungkinan proposisi lain.

# Random Variables

- Variabel acak adalah variabel yang dapat mengambil nilai yang berbeda secara acak.
- Notasi biasanya dengan huruf kecil dalam jenis huruf biasa, dan nilai yang dapat diambilnya dengan huruf skrip huruf kecil.
  - Misalnya, $x_1$ dan $x_2$ keduanya merupakan nilai yang mungkin yang dapat diambil oleh variabel acak $x$.
  - Untuk variabel bernilai vektor, kita menulis variabel acak sebagai **x** dan salah satu nilainya sebagai $x$.

# Random Variables

- Variabel acak mungkin diskrit atau kontinu.
- Variabel acak diskrit adalah salah satu yang memiliki jumlah state yang terbatas atau tak terbatas.
  - Perhatikan bahwa status ini belum tentu bilangan bulat; mereka juga bisa diberi nama sub state yang tidak dianggap memiliki nilai numerik.
- Variabel acak kontinu dikaitkan dengan nilai nyata.

# Probability Mass Function

- Distribusi probabilitas atas variabel diskrit dapat dijelaskan menggunakan Probability Mass Function (PMF).
- Notasi fungsi massa probabilitas dengan kapital **P**.

# Probability Mass Function

- Domain dari ***P*** harus merupakan himpunan semua keadaan yang mungkin dari x.

$$\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$$

  - Suatu kejadian yang tidak mungkin memiliki probabilitas 0 dan tidak ada keadaan yang kemungkinannya lebih kecil dari itu. Demikian pula, suatu peristiwa yang dijamin akan terjadi memiliki probabilitas 1, dan tidak ada keadaan yang memiliki peluang lebih besar untuk terjadi.

$$\sum_{x \in \mathrm{x}} P(x) = 1.$$

  - Disebut sebagai sebagai dinormalisasi (normalized). Tanpa sifat ini, kita dapat memperoleh probabilitas lebih besar dari satu dengan menghitung probabilitas salah satu dari banyak kejadian yang terjadi

# Probability Mass Function

- Sebagai contoh, pertimbangkan satu variabel acak diskrit x dengan k keadaan berbeda. Kita dapat menempatkan distribusi seragam (uniform distribution) pada *x* —yaitu, membuat setiap statusnya memiliki kemungkinan yang sama— dengan menset PMF menjadi:

$$\text{uniform distribution:} \quad P(\mathrm{x} = x_i) = \frac{1}{k}$$

- untuk semua *i*. Kita dapat melihat bahwa ini sesuai dengan persyaratan untuk fungsi massa probabilitas. Nilai 1/*k* positif karena *k* adalah bilangan bulat positif.

# Probability Density Function

- Saat bekerja dengan variabel acak kontinu, kita menggambarkan distribusi probabilitas menggunakan Probability Density Function (PDF) daripada fungsi massa probabilitas.
- Untuk menjadi Probability Density Function , fungsi $p$ harus memenuhi sifat-sifat berikut:
  - Domain dari $p$ harus merupakan himpunan semua keadaan yang mungkin dari $x$
  - $\forall x \in x$, $p(x) \geq 0$. Note : kita tidak memerlukan $p(x) \leq 1$.
  - $\int p(x)dx = 1$

# Probability Density Function

- Untuk contoh PDF yang sesuai dengan kerapatan probabilitas spesifik atas variabel acak kontinu, pertimbangkan distribusi seragam (uniform distribution) pada interval bilangan real.

$$u(x; a, b) = \frac{1}{b-a}.$$

  - di mana *a* dan *b* adalah titik ujung interval, dengan $b > a$.
  - ";" notasi berarti "diparametrikan oleh"; kita menganggap *x* sebagai argumen dari fungsi, sedangkan *a* dan *b* adalah parameter yang menentukan fungsi.
- Untuk memastikan bahwa tidak ada massa probabilitas di luar interval, kita katakan $u(x;a, b) = 0$ untuk semua $x \notin [a, b]$.
- Di dalam $[a, b]$, $u(x;a, b) = \frac{1}{b-a}$.
- Kita sering menyatakan bahwa *x* mengikuti distribusi seragam pada $[a, b]$ dengan menulis $x \sim U(a, b)$

# Marginal Probability

- Kadang-kadang kita mengetahui distribusi probabilitas pada sekumpulan variabel dan kita ingin mengetahui distribusi probabilitas hanya pada subset dari variabel tersebut.
- Distribusi probabilitas atas subset dikenal sebagai **distribusi probabilitas marjinal.**
- Misalnya, kita memiliki variabel acak diskrit *x* dan *y*, dan kita mengetahui *P*(*x*, *y*). Kita dapat menemukan *P*(*x*) dengan aturan penjumlahan:

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y).$$

# Conditional Probability

- Dalam banyak kasus, kita tertarik pada probabilitas suatu kejadian, mengingat bahwa kejadian lain telah terjadi.
- Ini disebut probabilitas bersyarat.
- Kita menyatakan probabilitas bersyarat bahwa Y = *y* diberikan X = *x* sebagai *P*(Y = y | X = x).
- Probabilitas bersyarat ini dapat dihitung dengan rumus

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}.$$

# Conditional Probability

- Probabilitas bersyarat hanya ditentukan ketika $P(x = x) > 0$.
- Kita tidak dapat menghitung probabilitas bersyarat yang dikondisikan pada peristiwa yang tidak pernah terjadi.

# Chain Rule of Conditional Probabilities

- Setiap distribusi probabilitas gabungan pada banyak variabel acak dapat didekomposisi menjadi distribusi bersyarat hanya pada satu variabel:

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$$

- Pengamatan ini dikenal sebagai **aturan rantai**, atau aturan produk dari probabilitas.
- Misalnya, menerapkan definisi dua kali, kita dapatkan

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}). \end{aligned}$$

# Independence and Conditional Independence

- Dua variabel acak *x* dan *y* adalah **independen** jika distribusi probabilitasnya dapat dinyatakan sebagai produk (hasil kali) dari dua faktor, satu hanya melibatkan *x* dan satu hanya melibatkan *y*:

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y).$$

- Dua variabel acak *x* dan *y* **bebas bersyarat** diberikan variabel acak *z* jika **distribusi probabilitas bersyarat** atas *x* dan *y* difaktorkan dengan cara ini untuk setiap nilai *z*:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z).$$

# Expectation, Variance and Covariance

- **Ekspektasi**, atau **nilai ekspektasi**, dari beberapa fungsi $f(x)$ sehubungan dengan **distribusi probabilitas** $P(x)$ adalah **rata-rata**, atau **nilai rata-rata**, yang diambil $f$ ketika $x$ diambil dari $P$. Untuk variabel diskrit, ini dapat dihitung dengan penjumlahan:

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x) f(x),$$

- sedangkan untuk variabel kontinu, dihitung dengan integral:

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x) f(x) dx.$$

# Expectation, Variance and Covariance

- Ketika identitas distribusi jelas dari konteksnya, kita cukup menulis nama variabel acak yang harapannya berakhir, seperti pada $\mathbb{E}_x[f(x)]$.
- Jika jelas variabel acak mana yang harapannya berakhir, kita dapat menghilangkan subskrip seluruhnya, seperti pada $\mathbb{E}[f(x)]$
- Secara default, kita dapat mengasumsikan bahwa $\mathbb{E}[.]$ rata-rata di atas nilai semua variabel acak di dalam tanda kurung. Demikian pula, bila tidak ada ambiguitas, kita dapat menghilangkan tanda kurung siku.

# Expectation, Variance and Covariance

- Expectation bersifat linier, misalnya:

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathrm{x}}[f(x)] + \beta \mathbb{E}_{\mathrm{x}}[g(x)],$$

- dimana $\alpha$ dan $\beta$ tidak dependen terhadap $x$

# Expectation, Variance and Covariance

- Varians memberikan ukuran seberapa besar nilai fungsi variabel acak *x* bervariasi saat kita mengambil sampel nilai *x* yang berbeda dari distribusi probabilitasnya:

$$\operatorname{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right].$$

- Ketika varians rendah, nilai *f* (*x*) mengelompok mendekati nilai yang diharapkan (expectation value). Akar kuadrat dari varian dikenal sebagai **standar deviasi**.

# Expectation, Variance and Covariance

- Kovarian memberikan pengertian tentang seberapa banyak dua nilai terkait secara linear satu sama lain, serta skala variabel-variabel ini:

$$\operatorname{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right].$$

- Nilai **kovarians absolut yang tinggi** berarti bahwa nilainya sangat banyak berubah dan keduanya jauh dari rata-rata masing-masing pada saat yang bersamaan.
- Jika **tanda kovariansnya positif**, maka kedua variabel tersebut cenderung mengambil nilai yang relatif tinggi secara bersamaan.
- Jika **tanda kovarians negatif**, maka satu variabel cenderung mengambil nilai yang relatif tinggi pada saat yang lain mengambil nilai yang relatif rendah dan sebaliknya.

# Expectation, Variance and Covariance

- Ukuran lain seperti **korelasi** menormalkan kontribusi masing-masing variabel untuk mengukur hanya seberapa banyak variabel terkait, daripada juga dipengaruhi oleh skala variabel yang terpisah.
- **covariance matrix** dari sebuah random vector $\boldsymbol{x} \in \mathbb{R}^n$ merupakan matriks $n \times n$,

$$\operatorname{Cov}(\mathbf{x})_{i,j} = \operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_j).$$

- Elemen diagonal dari covariance memberikan nilai variance:

$$\operatorname{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \operatorname{Var}(\mathrm{x}_i).$$

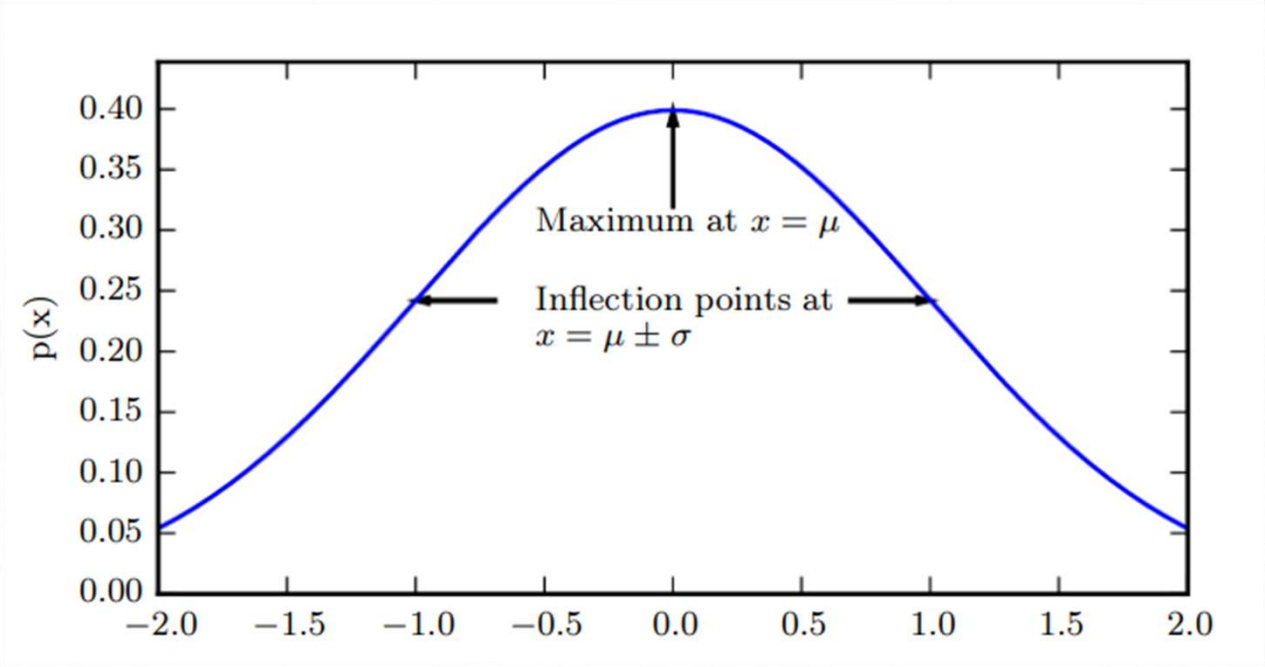
# Gaussian Distribution

- Distribusi yang paling umum digunakan pada bilangan real adalah distribusi normal, juga dikenal sebagai **distribusi Gaussian**

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

- Dua parameter $\mu \in \mathbb{R}$ dan $\sigma \in (0, \infty)$ mengontrol distribusi normal.
- Parameter $\mu$ memberikan koordinat puncak pusat. Ini juga merupakan rata-rata distribusi: $\mathbb{E}[x] = \mu$.
- Deviasi standar dari distribusi diberikan oleh $\sigma$, dan varians oleh $\sigma^2$

# Gaussian Distribution



# Gaussian Distribution

- Distribusi normal menggeneralisasi ke $\mathbb{R}^n$, dalam hal ini dikenal sebagai **multivariate normal distribution**. Ini dapat diparametrikan dengan matriks simetris pasti positif $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

- Parameter $\mu$ masih memberikan rata-rata distribusi, meskipun sekarang bernilai vektor. Parameter $\boldsymbol{\Sigma}$ memberikan matriks kovarians dari distribusi.

# Gaussian Distribution

- Seperti dalam kasus univariat, ketika kita ingin mengevaluasi PDF beberapa kali untuk banyak nilai parameter yang berbeda, kovarians bukanlah cara komputasi yang efisien untuk memparametrikan distribusi, karena kita perlu membalikkan $\boldsymbol{\Sigma}$ untuk mengevaluasi PDF. Sebagai gantinya, kita dapat menggunakan matriks presisi $\boldsymbol{\beta}$:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

- Kita sering memperbaiki matriks kovarians menjadi matriks diagonal. Versi yang lebih sederhana adalah **distribusi Gaussian isotropik**, yang matriks kovariansnya adalah skalar kali matriks identitas.

# Exponential and Laplace Distributions

- Dalam konteks *deep learning*, kita sering ingin memiliki distribusi probabilitas dengan titik tajam di $\boldsymbol{x} = 0$. Untuk melakukannya, kita dapat menggunakan distribusi eksponensial:

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x).$$

- Distribusi eksponensial menggunakan fungsi indikator $\mathbf{1}_{x \geq 0}$ untuk menetapkan probabilitas nol ke semua nilai negatif dari x

# Exponential and Laplace Distributions

- Distribusi probabilitas yang terkait erat yang memungkinkan kita untuk menempatkan puncak massa probabilitas yang tajam pada titik *arbitrer* $\mu$ adalah **distribusi Laplace**

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right).$$

# Dirac Distribution and Empirical Distribution

- Dalam beberapa kasus, kita ingin menentukan bahwa semua massa dalam cluster distribusi probabilitas di sekitar satu titik.
- Ini dapat dicapai dengan mendefinisikan PDF menggunakan **fungsi delta Dirac**, $\delta(x)$:

$$p(x) = \delta(x - \mu).$$

- **Fungsi delta Dirac** didefinisikan sedemikian rupa sehingga bernilai nol di mana-mana kecuali 0, namun berintegrasi dengan 1.
- **Fungsi delta Dirac** bukanlah fungsi biasa yang menghubungkan setiap nilai *x* dengan output bernilai riil;

# Dirac Distribution and Empirical Distribution

- sebaliknya itu adalah jenis objek matematika yang berbeda yang disebut fungsi umum yang didefinisikan dalam sifat-sifatnya ketika diintegrasikan.
- Kita dapat menganggap **fungsi delta Dirac** sebagai titik batas dari serangkaian fungsi yang semakin mengurangi kerapatan pada semua titik selain nol.
- Dengan mendefinisikan $\boldsymbol{p}(x)$ menjadi $\delta$ digeser oleh $-\mu$ kita memperoleh puncak kepadatan probabilitas yang sangat sempit dan sangat tinggi di mana $x = \mu$

# Dirac Distribution and Empirical Distribution

- Penggunaan umum dari **distribusi delta Dirac** adalah sebagai komponen dari **distribusi empiris**,

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

- yang menempatkan massa probabilitas $\frac{1}{m}$ pada masing-masing titik $m$ : $x^{(1)}, \ldots, x^{(m)}$, membentuk kumpulan data atau kumpulan sampel tertentu..

# Dirac Distribution and Empirical Distribution

- Distribusi delta Dirac hanya diperlukan untuk menentukan distribusi empiris atas variabel kontinu.
- Untuk variabel diskrit, situasinya lebih sederhana: distribusi empiris dapat dikonseptualisasikan sebagai **distribusi multinoulli**, dengan probabilitas yang terkait dengan setiap nilai input yang mungkin sama dengan **frekuensi empiris** dari nilai tersebut dalam set pelatihan.

# Mixtures of Distributions

- Juga umum untuk mendefinisikan distribusi probabilitas dengan menggabungkan distribusi probabilitas lain yang lebih sederhana.
- Salah satu cara umum untuk menggabungkan distribusi adalah dengan membuat **distribusi campuran (**Mixtures of Distributions).
- Mixtures Distributions terdiri dari beberapa distribusi komponen.
- Pada setiap percobaan, pilihan distribusi komponen mana yang harus menghasilkan sampel ditentukan dengan mengambil sampel identitas komponen dari distribusi multinoulli:

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i),$$

- di mana $P(c)$ adalah **distribusi multinoulli** atas identitas komponen.

# *Gaussian mixture model*

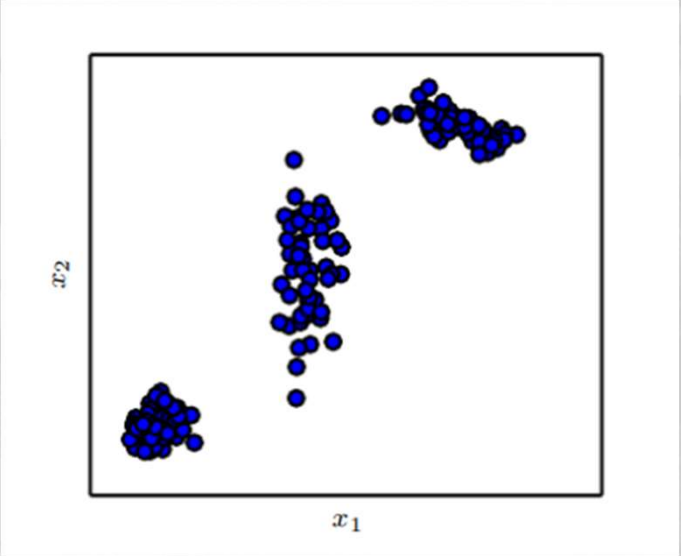
- Jenis model campuran yang sangat kuat dan umum adalah model campuran Gaussian (***Gaussian mixture model***), di mana komponen $p(x \mid c = i)$ adalah *Gaussian*.
- Setiap komponen memiliki mean yang diparametrikan secara terpisah $\mu\,(i)$ dan kovarians $\Sigma(i)$.
- Beberapa campuran dapat memiliki lebih banyak kendala.
- Misalnya, kovarians dapat dibagi ke seluruh komponen melalui batasan $\Sigma^{(i)} = \Sigma,\ \forall_i$. Seperti dengan distribusi Gaussian tunggal, campuran Gaussian mungkin membatasi matriks kovarians untuk setiap komponen menjadi diagonal atau isotropik.

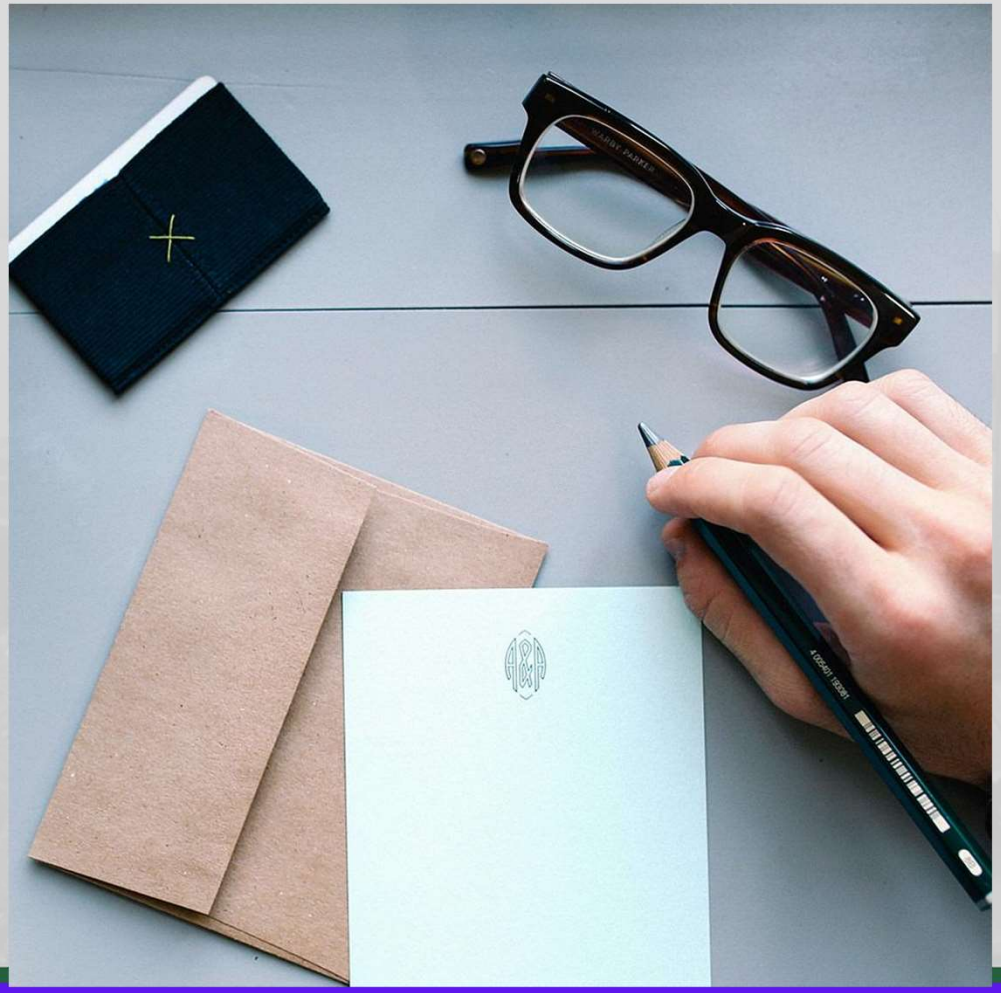
# *Gaussian mixture model*

- Selain rata-rata dan kovarians, parameter **campuran Gaussian** menentukan probabilitas sebelumnya $\alpha_i = P(c = i)$ yang diberikan untuk setiap komponen *i*.
- Kata "***prior***" menunjukkan bahwa itu mengungkapkan keyakinan model tentang *c* sebelum mengamati *x*.
- Sebagai perbandingan, $P(c \mid x)$ adalah probabilitas ***posterior***, karena dihitung setelah pengamatan *x*.
- **Model campuran Gaussian** adalah pendekatan universal dari kerapatan, dalam arti bahwa setiap kerapatan halus dapat didekati dengan jumlah kesalahan bukan nol tertentu oleh **model campuran Gaussian** dengan komponen yang cukup.

# *Gaussian mixture model*

- Sampel dari model campuran Gaussian.
- Dalam contoh ini, ada tiga komponen. Dari kiri ke kanan,
  - komponen pertama memiliki **matriks kovarians isotropik**, artinya memiliki jumlah varians yang sama di setiap arah.
  - Yang kedua memiliki **matriks kovarians diagonal**, artinya dapat mengontrol varians secara terpisah di sepanjang setiap arah yang sejajar sumbu. Contoh ini memiliki lebih banyak variasi sepanjang sumbu $x_2$ daripada sepanjang sumbu $x_1$.
  - Komponen ketiga memiliki **matriks kovarians peringkat penuh**, memungkinkannya untuk mengontrol varians secara terpisah di sepanjang basis arah yang berubah-ubah.





# TERIMA KASIH