5. Pendahuluan Pengenalan Pola

5.1 Pendahuluan

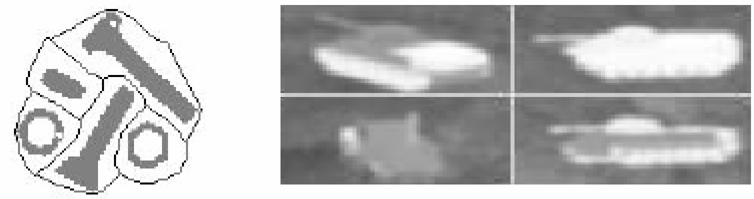
Apakah pengenalan pola itu? Beberapa penulis mendefinisikannya dalam berbagai cara, diantaranya adalah:

- "The assignment of a physical object or event to one of several pre-specified categories" – *Duda and Hart*
- "A problem of estimating density functions in a high-dimensional space and dividing the space into the regions of categories or classes" –*Fukunaga*
- "Given some examples of complex signals and the correct decisions for them, make decisions automatically for a stream of future examples" –*Ripley*
- "The science that concerns the description or classification (recognition) of measurements" –*Schalkoff*
- "The process of giving names ω to observations x", –*Schürmann*
- Pattern Recognition is concerned with answering the question "*What is this?*" –*Morse*

Dari definisi-definisi diatas, secara garis besar dapat dirangkum bahwa pengenalan pola memetakan suatu fitur, yang merupakan ciri utama suatu objek (yang dinyatakan dalam sekumpulan bilangan-bilangan) ke suatu kelas yang sesuai. Proses pemetaan ini menyangkut inferensi, baik secara eksplisit secara statistik (misalnya dalam aturan Bayesian) maupun tak eksplisit dengan suatu jaringan keputusan (misalnya jaringan syaraf tiruan atau logika samar). Pengenalan pola memainkan peran penting dalam berbagai bidang rekayasa, seperti pada contoh-contoh berikut.

Penglihatan Mesini (Machine Vision)

Disini pengenalan pola dipakai untuk memeriksa hasil produksi secara visual, pendeteksi sasaran atau penentuan *friend* atau *foe* dalam suatu peperangan.



Gambar 5.1 *Machine Vision* dipakai untuk mengenali:
(a) perkakas dan (b) tank

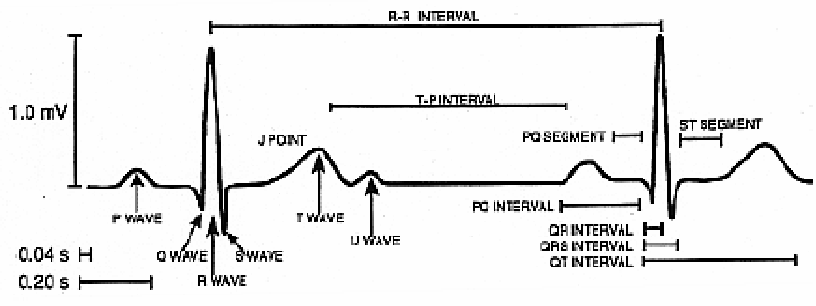
Pengenalan Tulisan Tangan

Pengenalan pola dapat dipakai dalam otomasi proses penyortiran surat di suatu kantor pos dengan mengenali alamat yang dituju, atau pengelolaan cek di bank. Disini, suatu penyapu citra mengangkap citra dijital tulisan tangan dan kemudian sistem pengenal pola menerjemahkan menjadi karakter.

Gambar 5.2 Pengenalan tulisan tangan pada surat pos

Diagnosis Kesehatan Berbantuan Komputer

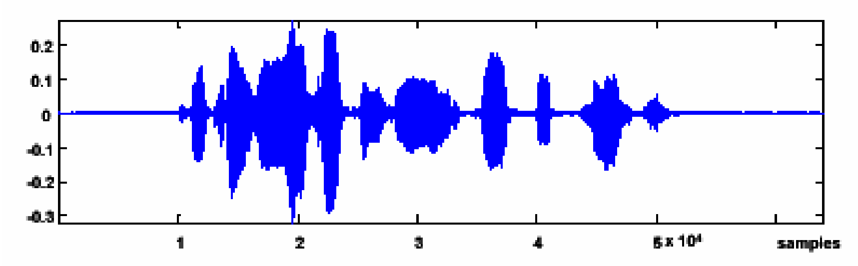
Akhir-akhir ini penegenalan pola mulai diaplikasikan secara luas dalam dunia kedokteran untuk membantu melakukan diagnosis, misalnya deteksi penyakin dari sinyal-sinyal EEG, EKG citra fundus mata atau mammografi sinar-X untuk mendiagnosa knaker. Meskipun diagnosis akhir dilakukan oleh dokter, *screening* awal dengan alat ini sangat membantu jika jumlah pasien sangat banyak dan jumlah spesialis sangat terbatas.



Gambar 5.3 Pengenalan pola untuk analisa medis berbantuan komputer

Pengenalan Ucapan Otomatis

Dalam HCI (Human Computer Interaction), diinginkan terjadinya komunikasi antara manusia dengan computer dengan memakai bahasa alami. Untuk keperluan yang demikian, dalam tahap awal komputer harus mampu menerjemahkan ucapan kedalam tulisan. Program komersil untuk menerjemahkan ucapan menjadi tulisan juga sudah mulai banyak dipakai untuk mempercepat penulisan teks.



Gambar 5.4 Pengenalan ucapan melalui analisa sinyal ucapan

Secara ringkas, beberapa aplikasi pengenalan pola adalah sebagai berikut:

- Pengolahan/segmentasi Citra
- *Computer Vision*
- Pengenalan percakapan
- Pengenalan sasaran otomatis (ATR)
- Pengenalan tulisan tangan (OCR)
- Analisis seismik
- Identifikasi sidik jari
- Inspeksi industri
- Peramalan keuangan
- Diagnosis kesehatan
- Analisis sinyal ECG

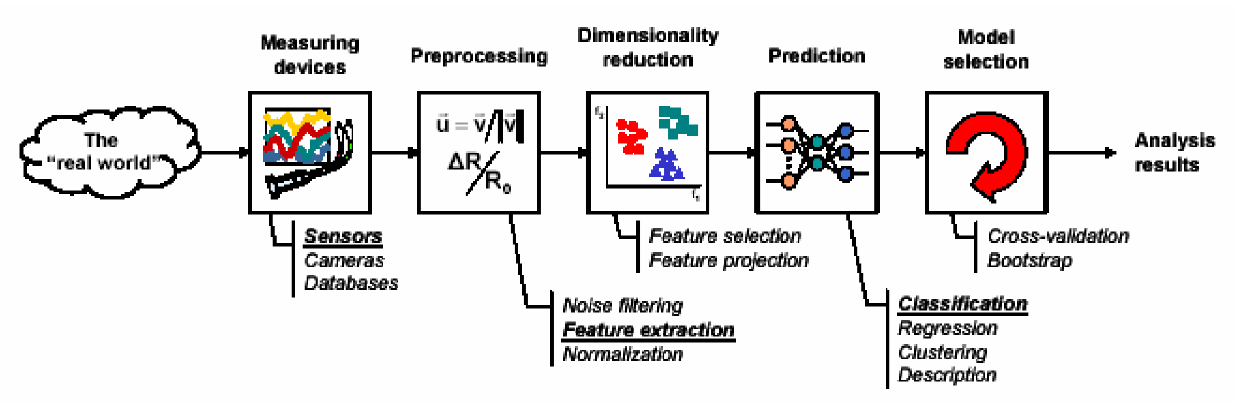
Selain itu, ada beberapa bidang dalam sains dan rekayasa yang terkait dengan pengenalan pola, diantaranya adalah

- Peng. Sinyal Adaptif
- Pembelajaran Mesin

- Jaringan Syaraf Buatan
- Robotika dan Vision
- Sains Kognitif
- Matematika Statistik
- Optimisasi Tak-Linier
- Analisa Data *Exploratory*
- Sistem Fuzzy dan Genetika
- Teori Deteksi dan Estimasi
- Bahasa Formal
- Pemodelan Struktural
- Sibernetika Biologis
- Neurosains Komputasional

5.2 Komponen Sistem Pengenalan Pola

Secara mendasar, suatu sistem pengenalan pola terdiri dari komponen-komponen berikut: sensor, mekanisme pre-processing, mekanisme penyari fitur (manual/otomatis), algoritma pemilah dan sekumpulan contoh pelatihan yang telah dipilah. Diagram blok dari sistem pengenal pola dapat digambarkan sebagai berikut:



Gambar 5.5 Komponen sistem pengenal pola

- Sensor: berfungsi untuk menangkap objek dari dunia nyata menjadi sinyal-sinyal listrik (dan selanjutnya dalam bilangan-bilangan setelah proses dijitasi).
- Preprocessing: berfungsi untuk menonjolkan sinyal informasi dan menekan derau dalam sinyal.
- Penyari fitur: mengambil besaran komponen tertentu dari sinyal yang mewakili sifat utama sinyal, sekaligus mengurangi dimensi sinyal menjadi sekumpulan bilangan yang lebih sedikit tetapi representatif.
- Algoritma pemilah: melakukan assignment fitur ke kelas yang sesuai
- Sekumpulan contoh pelatihan: dipakai untuk mendapatkan repersentasi kelas

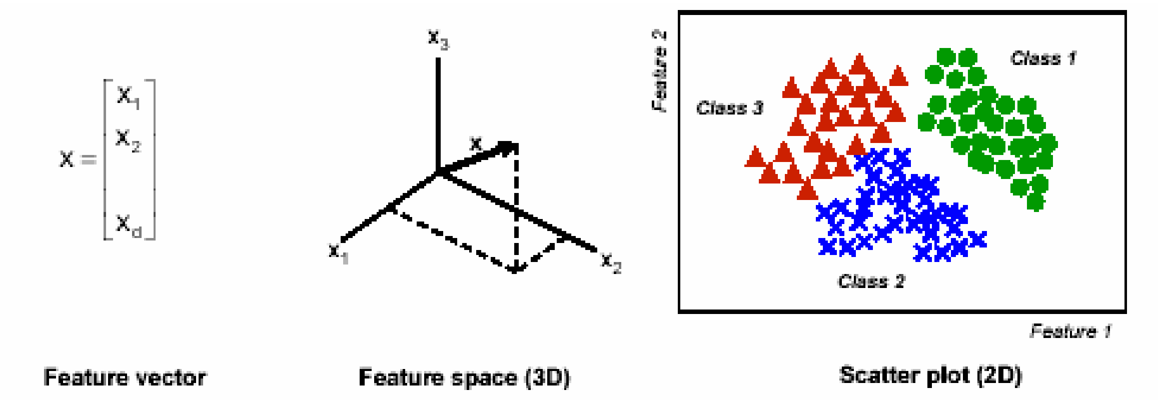
Ada beberapa bidang terkait pengenalan pola yang sepintas mirip tetapi sebenarnya berlainan, yaitu:

- **Pemilahan.** Adalah pengelompokkan obyek kedalam kelas-kelas. Keluaran dari sistem pengenal pola adalah bilangan bulat atau label. Contoh: klasifikasi produk menjadi "bagus" dan "jelek"
- **Regresi.** Adalah perampatan (generalisasi) dari klasifikasi dimana keluarannya berupa bilangan riil. Contoh: peramalan nilai saham suatu perusahaan berdasarkan kinerja masa lalu di bursa saham
- **Klastering.** Adalah permasalahan mengorganisasi objek kedalam kelompok-kelompok yang berarti. Keluaran dari pengklasteran akan berupa kelompok-kelompok objek (kadangkala hirarkis). Contoh: mengorganisasikan makhluk hidup kedalam taksonomi dari spesies
- **Deskripsi.** Adalah permasalahan merepresentasikan objek sebagai sederetan primitive. Keluarannya berupa deskripsi struktural atau lnguistik. Contoh: pelabelan atau klasifikasi sinyal ECG

5.3 Fitur (Feature) dan Pola (Pattern)

Fitur adalah segala jenis aspek pembeda, kualitas atau karakteristik. Fitur bisa berwujud simbolik (mis. warna) atau numerik (mis. tinggi). Ada beberapa definisi terkait dengan fitur sebagai berikut:

- Kombinasi dari *d*-buah fitur dinyatakan sbg vektor kolom dimensi-*d* dan disebut **vektor fitur**.
- Ruang dimensi-*d* yang dibentuk oleh vektor fitur disebut sebagai **ruang fitur.**
- Objek dinyatakan sebagai sebuah titik di dlm ruang fitur. Penggambaran demikian disebut sbg **diagram hambur** (scatter plot)

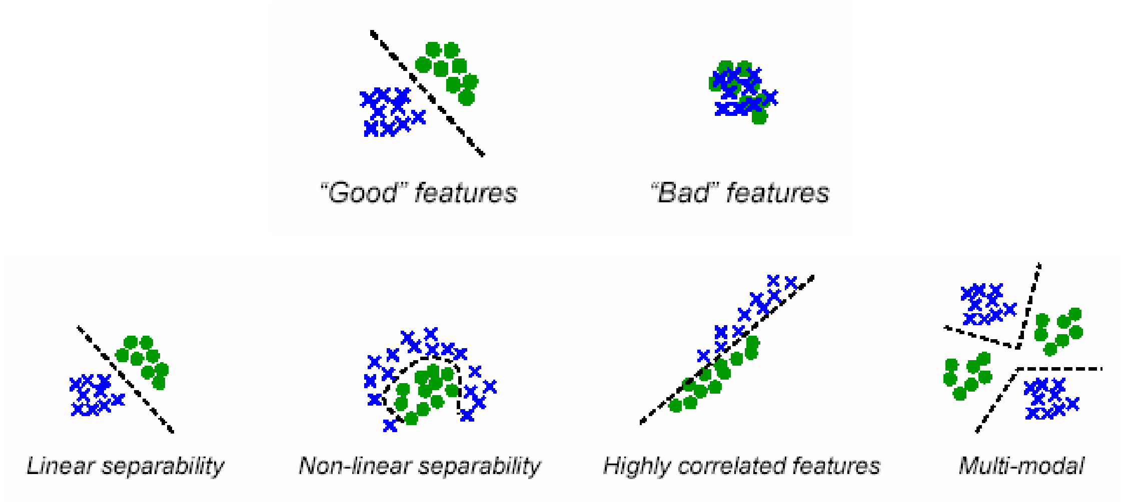


Gambar 5.6 Aspek fitur: (a) vektor, (b) ruang fitur dan (c) diagram hambur fitur

Pola adalah komposit/gabungan dari fitur yang merupakan sifat dari sebuah objek. Didalam klasifikasi, pola berupa sepasang variabel (**x**, ω), dimana

- **x** adalah sekumpulan pengamatan atau fitur (vektor fitur)
- ω adalah konsep dibalik pengamatan (label)

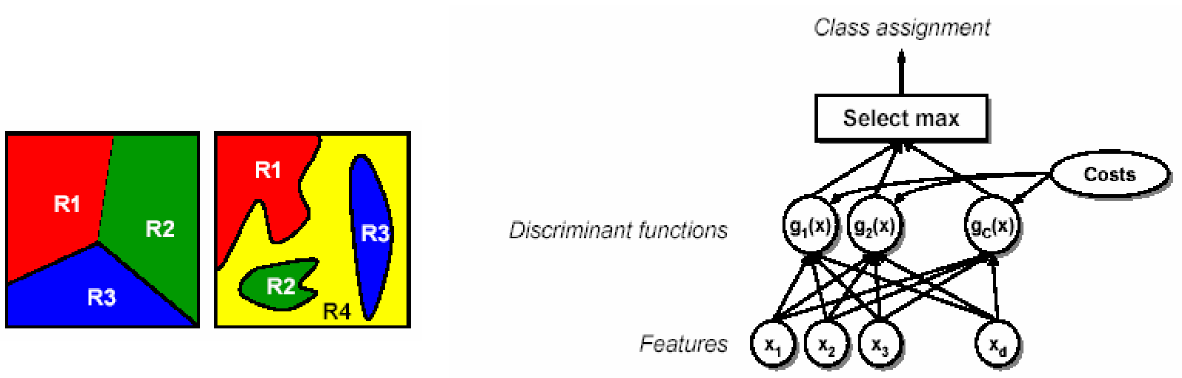
Kualitas dari suatu vektor fitur dilihat dari kemampuannya membedakan objek yang berasal dari kelas yang berbeda-beda. Objek dalam kelas yang sama harus punya nilai vektor fitur yang sama dan objek yang berada dalam kelas yang berbeda harus punya nilai vektor fitur yang berlainan pula. Sifat-sifat dari fitur, linear/non-linear separability, korelasi dan modalitas, sangat penting untuk menentukan sistem pengenal yang cocok. Berbagai sifat fitur yang penting dapat digambarkan dalam diagram hambur berikut ini.



Gambar 5.7 Sifat keterpisahan fitur dan jenis-jenis fitur

5.4 Pemilah (Classifier)

Tugas dari pemilah adalah untuk menyekat ruang fitur kedalam daerah-daerah yang dilabeli dengan kelas (Gb.XX). Garis batas antar daerah keputusan disebut sebagai perbatasan keputusan. Pemilahan vektor fitur x meliputi penentuan daerah keputusan yang sesuai dan pengelompokan x kedalam kelas ini Pemilah dapat diwujudkan sebagai sekumpulan fungsi diskriminan. Dalam contoh di atas, pemilah menentukan vektor fitur x ke dalam kelas ω_i jika $g_i(x) > g_j(x)$, $\forall i \neq j$.

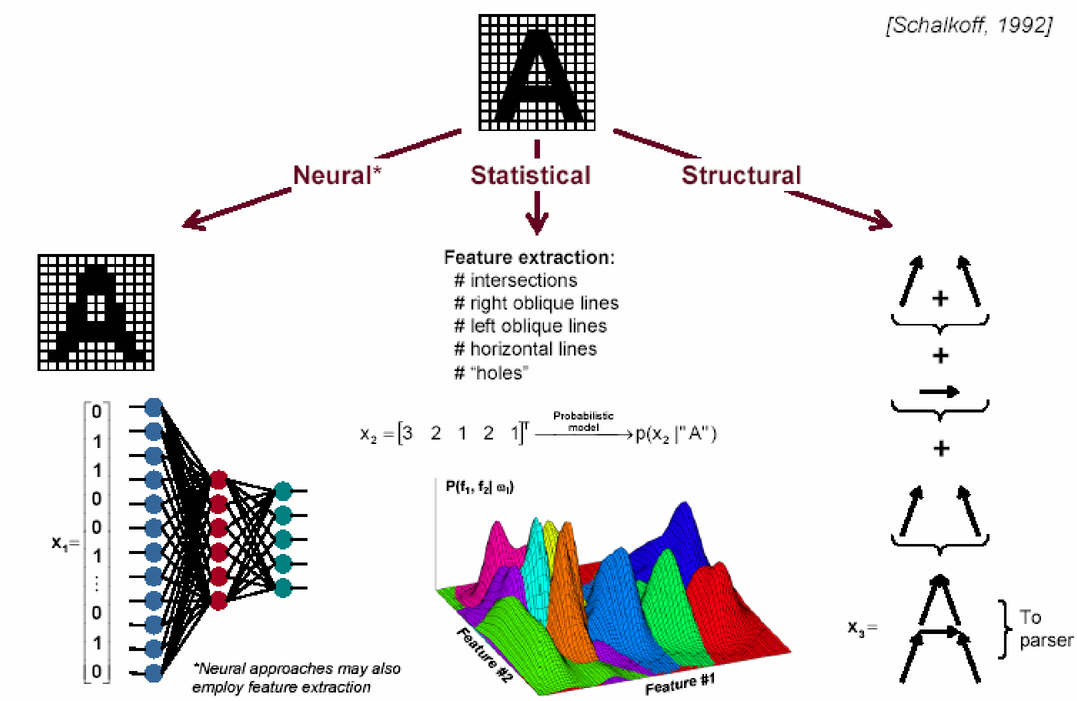


Gambar 5.8 Daerah kelas dan pemilah dinyatakan sebagai fungsi diskriminan

Metode Pengenalan Pola

Ada beberapa metode penting dalam pengenalan pola yang perlu kita kenal, yaitu: metode pola statistik, jaringan syaraf tiruan dan metode sintaktik. Didalam metoda statistik (StatPR), pola dipilah berdasarkan model statistik dari fitur. Model statistik didefinisikan sebagai sebuah keluarga dari fungsi kerapatan peluang bersyarat kelas $\Pr(x \mid c_i)$, yakni peluang vektor fitur x jika diberikan kelas c_i. Dalam metode jaringan syaraf /Neural (NeurPR), pemilahan dilakukan berdasarkan tanggapan suatu neuron jaringan pengolah sinyal (neuron) terhadap stimulus masukan (pola). 'Pengetahuan' disimpan dalam sambungan antar neuron dan kekuatan pembobot sinaptik. NeuroPR dapat dilatih, non-algoritmik, dan memakai strategi *black box*. NeuroPR sangat menarik karena:

- pengetahuan apriori yang diperlukan hanya sedikit
- dengan jumlah lapisan dan neurun secukupnya, JST dapat membentuk semua jenis daerah keputusan yang rumit sekalipun



Gambar 5.9 Perbandingan tiga teknik pengenalan pola: neural, statistik dan struktura/ sintaktik

Dalam metoda sintaktik (SyntPR) atau metoda struktural, pola dipilah berdasarkan keserupaan ukuran struktural, 'pengetahuan' direpresentasikan secara *formal grammar* atau deskripsi relasional (graf). SyntPR dipakai tidak hanya untuk pemilahan, tetapi juga untuk deskripsi. Biasanya, SyntPR memformulasikan deskripsi hirarkis dari pola kompleks yang tersusun dari pola bagian yang lebih sederhana. Perbedaan ketiga metode ini dalam menangani masalah pengenalan karakter dapat digambarkan sebagai berikut.

5.5 Siklus Perancangan Sistem Pengenal Pola

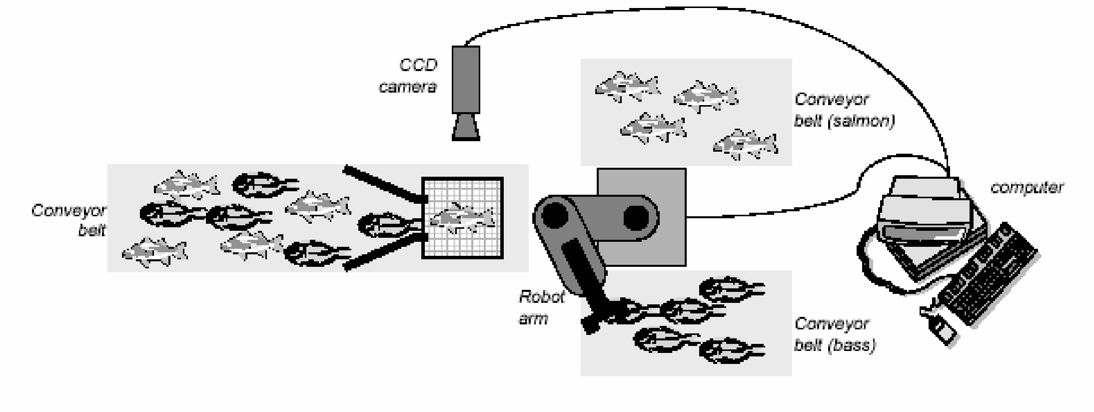
Pembuatan suatu sistem pengenal pola yang baik mengikuti beberapa tahapan, dimulai dari pengumpulan data, pemilihan fitur, pemilihan teknik pemilah atau model yang sesuai, pelatihan dan pengujian.

- **Pengumpulan data.** Barangkali bagian paling intensif dalam proyek pengenalan pola. Pertanyaan mendasar dalam tahap ini adalah: Seberapa banyak yang diperlukan hingga dianggap cukup ?
- **Pemilihan fitur**. Ini adalah bagian paling kritis dari keberhasilan sistem PR. Fitur yang baik akan memberikan hasil baik, demikian pula sebaliknya ('garbage in, garbage out'). Di sini juga diperlukan pengetahuan prior dasar dari objek.
- **Pemilihan model**. Memilih model pengenal yang sesuai, apakah dengan pendekatan statistik, neural dan struktural. Hal yang harus diperhatikan pula adalah seting parameter
- **Pelatihan**. Diberikan kumpulan fitur dan model kosong, adaptasikan model untuk menjelaskan data. Disini dikenal ajar supervisi, tak-supervisi dan *reinforcement*.
- **Pengujian**. Mengevaluasi seberapa bagus kinerja dari model yang sudah dilatih? Harus diperhitungkan juga masalah overfitting vs generalisasi.

Studi Kasus

Untuk memperjelas permasalah perencanaan pengenalan pola, kita akan meninjau satu kasus berikut ini. Suatu Perusahaan pengalengan ikan ingin melakukan otomasi pemilahan ikan menurut spesies-nya yang hanya ada dua, yakni ikan salmon dan ikan *sea bass*. Sistem yang dirancang terdiri dari (lihat Gb.10)

- satu *conveyor belt* untuk mengangkut produk yg datang
- dua conveyor belt untuk mengangkut produk yang telah dipilah
- sebuah tangan robot untuk mengambil dan meletakkan ikan
- sistem vision dengan kamera pangamat CCD untuk mengambil citra ikan
- sebuah komputer untuk menganalisa citra ikan dan mengendalikan tangan robot



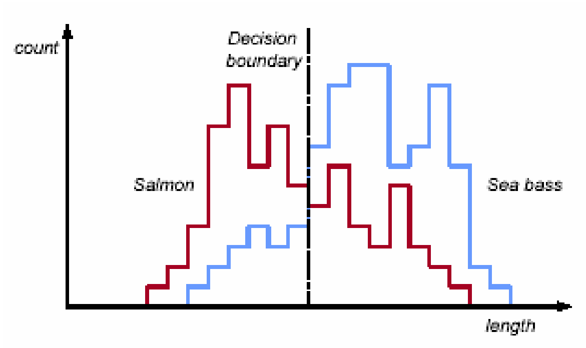
Gambar 5.10 Studi kasus: pemilahan ikan kedalam dua jenis

Suatu kamera dijital CCD dipakai untuk mendapatkan citra dijital dari ikan yang datang. Selanjutnya, suatu pengolah citra awal dipakai untuk memperjelas citra ikan dengan mengatur intensitas dan juga sekaligus melakukan segmentasi untuk memisahkan citra ikan dari citra latar belakangnya.

Diketahui bahwa ukuran rata-rata *sea bass* lebih besar daripada salmon, ini adalah informasi prior yang kita ketahui. Selanjutnya panjang ikan diestimasi dari citra ikan yang telah disegmentasi, yang merupakan tahap ekstraksi fitur.

Dalam tahap selanjutnya, yaitu pemilahan, dilakukan proses-proses berikut ini:

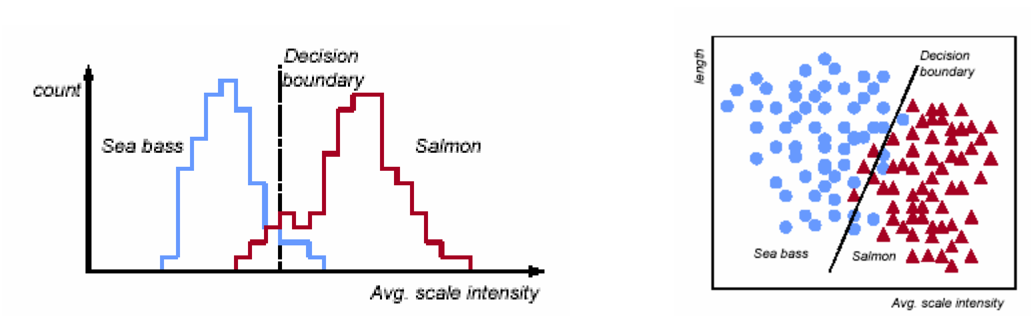
- o Beberapa contoh dari kedua jenis ikan dikumpulkan
- o Distribusi panjang dari kedua jenis ikan ditentukan
- o Selanjutnya kita cari perbatasan keputusan (ambang) yang meminimumkan kesalahan pemilahan.



Gambar 5.11 Sebaran kedua jenis: salmon dan *sea-bass*

Ternyata peluang salah dari pemilah diestimasi dan diperoleh hasil yang mengecewakan, yakni sebesar 40 %. Apa yang harus kita lakukan untuk meningkatkan kinerja sistem kita?

Pabrik pengalengan menginginkan kinerja minimum sebesar 95% laju pengenalan. Oleh karena itu kita coba berbagai macam fitur yang lebih sesuai, misalnya: lebar ikan, luasnya, posisi mata terhadap mulut, tetapi ternyata fitur-fitur tersebut tak memiliki informasi pembeda (yang signifikan). Akhirnya kita temukan fitur yang cukup baik, yaitu intensitas (rata-rata) dari citra ikan. Selanjutnya kita kombinasikan panjang ikan dan rata-rata intensitas citra ikan untuk memperbaiki keterpisahan kelas. Kita hitung fungsi diskriminan linier untuk menyekat kedua kelas, dan diperoleh laju klasifikasi 95.7% sesuai dengan yang diinginkan.



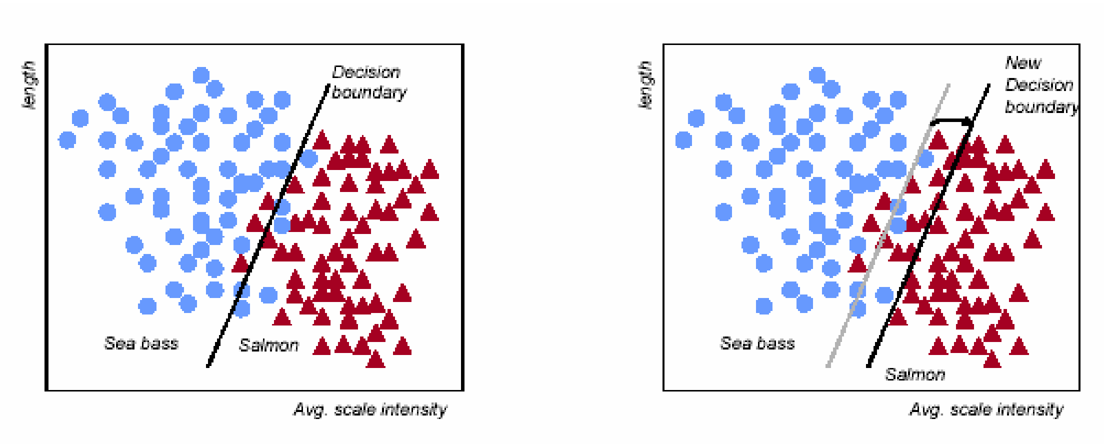
(a) (b)

Gambar 5.11 Fitur kedua, intensitas rata-rata:
(a) sebaran dan (b) diagram hambur fitur 2D

Pemilah linier yang kita implementasikan didesain untuk meminimumkan laju salah pilah. Apakah ini fungsi objektif dari pabrik pengolah ikan kita? Kita valuasi dua kasus salah pilah berikut ini:

- Biaya salah pilah dari salmon menjadi sea bass adalah, pelanggan kadangkala menemukan rasa lezat ikan salmon padahal ia membeli sea bass
- Biaya salah pilah sea bass menjadi salmon adalah pelanggan kita kadangkala kecewa saat memesan salmon karena merasakan sea bass

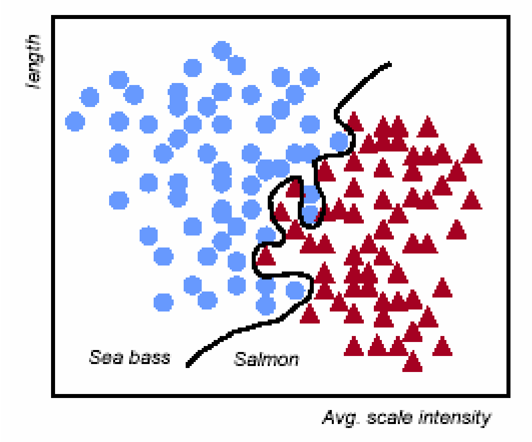
Secara intuitif hal ini bisa dipecahkan dengan menala perbatasan keputusan untuk meminimumkan fungsi biaya ini.



Gambar 5.12 Penalaan perbatasan keputusan untuk meningkatkan kinerja

Laju pengenalan dari pemilah linier kita (95.7%) memenuhi persyaratan perancangan, tetapi kita masih berpikir untuk memperbaiki kinerja sistem dengan cara merancang suatu jaringan syaraf tiruan dengan lima lapisan tersembunyi, kombinasi dari fungsi aktivasi logistik dan tangen hiperbolik, pelatihan dengan algoritma Levenberg-Marquardt, dan mendapatkan laju pilah yang mengagumkan sebesar 99.9975% dengan perbatasan keputusan berikut.

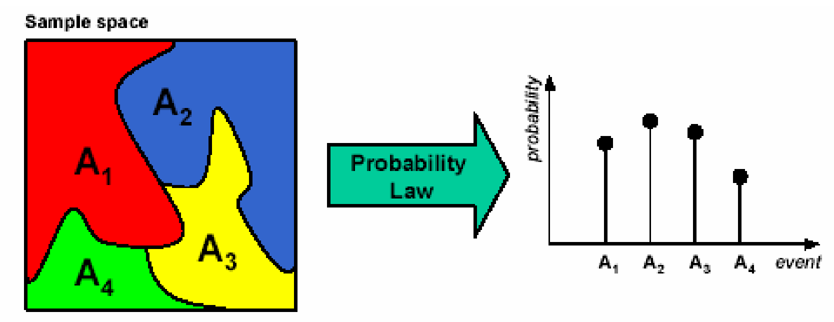
Puas dengan sistem pemilah, kita mengintegrasikan sistem dan memasang di pabrik pengolah ikan. Beberapa hari kemudian ternyata manajer pabrik mengeluh bahwa terjadi kesalahan pilah ikan, rata-rata sebesar 25%. Apanya yang salah? Ini berkaitan dengan masalah *trade-off* generalisasi yang akan dibahas lebih lanjut pada pelajaran mengenai jaringan syaraf tiruan.



Gambar 5.13 Pemilah bekerja optimal jika mampu membuat perbatasan yang sesuai

5.6 Dasar Teori Peluang dan Inferensi

Peluang adalah sebuah bilangan yang menyatakan seberapa mungkinkah suatu peristiwa akan muncul saat percobaan acak dilakukan.Hukum peluang untuk percobaan acak adalah suatu aturan yang memberikan nilai peluang kepada suatu peristiwa dalam percobaan tsb. Dalam peluang dikenal istilah ruang sampel (*sample space*). Ruang sampel S dari suatu percobaan acak adalah himpunan semua peristiwa yang mungkin terjadi.



Gambar 5.14 Ruang sampel dan pemetaannya ke fungsi distribusi peluang

Teori peluang didasarkan pada tiga aksioma berikut ini:

- Aksioma I: $0 \leq P(A_i)$
- Aksioma II: $P(S) = 1$
- Aksioma III: jika $A_i \cap A_j = \varnothing$, maka $P(A_i \cup A_j) = P(A_i) + P(A_j)$

dimana A_i adalah kumpulan kejadian dalam ruang sampel S. Aksioma pertama menyatakan bahwa nilai peluang adalah tak negatif. Aksioma kedua menyatakan bahwa keseluruhan ruang sampel memilki nilai peluang satu, sedangkan aksioma ketiga menyatakan bahwa untuk dua peristiwa (kumpulan kejadian) berbeda dan tak beririsan didalam suatu ruang sampel, peluang gabungan kedua peristiwa itu merupakan penjumlahan peluang masing-masing peristiwa. Peluang memiliki sifat-sifat berikut ini:

- Sifat 1: $P(A^C) = 1 - P(A)$, dengan A^C peristiwa komplementer dari A
- Sifat 2: $P(A) \leq 1$
- Sifat 3: $P(\varnothing) = 0$
- Sifat 4: untuk sekumpulan peristiwa $\{A_1, A_2, \dots, A_N\}$ jika $\{A_i \cap A_j = \varnothing, \forall i \neq j\}$, maka berlakulah

$$P\left(\bigcup_{k=1}^{N} A_k\right) = \sum_{k=1}^{n} P(A_k) \tag{5.1}$$

- Sifat 5: $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
- Sifat 6:

$$P\left(\bigcup_{k=1}^{N} A_k\right) = \sum_{k=1}^{N} P(A_k) - \sum_{j=k}^{N} P\left(A_j \cap A_k\right) + \dots + (-1)^{N+1} P\left(A_1 \cap A_2 \cap \dots \cap A_N\right) \tag{5.2}$$

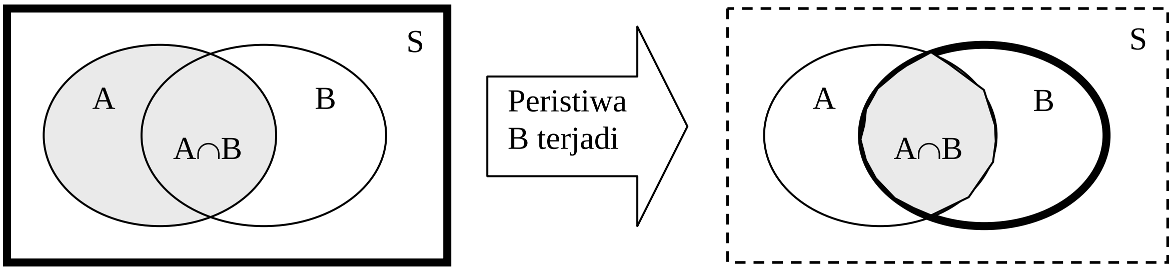
- Sifat 7: Jika $A_1 \subset A_2$, maka $P(A_1) \leq P(A_2)$

Peluang Bersyarat

Jika A dan B dua buah peristiwa dalam ruang sample S, peluang peristiwa A jika diketahui peristiwa B muncul didefinisikan sebagai peluang bersyarat

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ untuk } P(B) > 0 \tag{5.3}$$

Peluang bersyarat $P(A|B)$ dibaca sebagai “peluang bersyarat A dikondisikan pada B” atau “peluang A jika diberikan B”. Konsep peluang bersyarat bisa diilustrasikan dengan diagram Venn berikut ini.



Gambar 5.15 Penyusutan ruang sampel pada peluang bersyarat

Diagram Venn ini bisa ditafsirkan sebagai:

- Mula-mula peristiwa A adalah sebesar lingkaran yang diarsir pada gambar sebelah kiri
- Bukti baru "B terjadi" memberikan efek bahwa ruang sampel S (seluruh kotak) tereduksi menjadi B (lingkaran tebal paling kanan) dan peristiwa A menciut menjadi A∩B

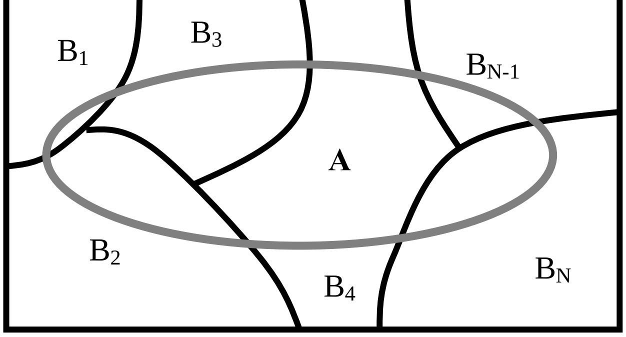
Dengan demikian P(B) menormalisasikan peluang peristiwa yang terjadi secara bersamaan dengan peristiwa B.

5.7 Teorema Peluang Total

Konsep penting lainnya dalam teori peluang adalah teorema peluang total yang dinyatakan secara sederhana sebagai berikut:

- Misalkan B_1, B_2, ... , B_N peristiwa yang saling bebas (*mutually exclusive*) yang gabungannya sama dengan ruang sampel S. Himpunan ini kita namakan <u>partisi</u> dari S.
- Maka, sebuah peristiwa A di dalam S dapat dinyatakan sebagai:

$$A = A \cap S = A \cap \left(B_1 \cup B_2 \cup ... \cup B_N\right) = \left(A \cap B_1\right) \cup \left(A \cap B_2\right) \cup ... \cup \left(A \cap B_N\right) \qquad (5.4)$$



Gambar 5.16 Partisi ruang sampel

Karena B_1, B_2, ..., B_N saling bebas, maka $P(A) = P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_N)$, dan akibatnya

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... + P(A|B_N)P(B_N) \\ &= \sum_{k=1}^{N} P(A|B_k)P(B_k) \end{aligned} \tag{5.5}$$

Teorema Bayes dan Pengenalan Pola Secara Statistik

Inferensi statistik pada suatu sistem pengenal pola statistik didasarkan pada teorema Bayes sebagai berikut. Diberikan B_1, B_2, ... , B_N yang merupakan partisi dari ruang sampel S. Jika suatu peristiwa A terjadi, berapakah peluang terjadinya peristiwa B_j? Dengan memakai definisi peluang bersyarat dan teorema peluang total akan diperoleh

$$P(B_j | A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{k=1}^{N} P(A|B_k)P(B_k)} \tag{5.6}$$

Rev. Thomas Bayes (1702-1761)

Persamaan diatas disebut juga sebagai aturan Bayes dan merupakan salah satu persamaan paling penting dalam teori peluang dan statistik. Tentu saja, persamaan Bayes adalah persamaan paling fundamental dalam pengenalan pola statistik.

Untuk tujuan pemilahan pola, teorema Bayes dapat dinyatakan sebagai

$$P(\omega_j | \mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{\sum_{k=1}^{N} P(\mathbf{x}|\omega_k)P(\omega_k)} = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})} \tag{5.7}$$

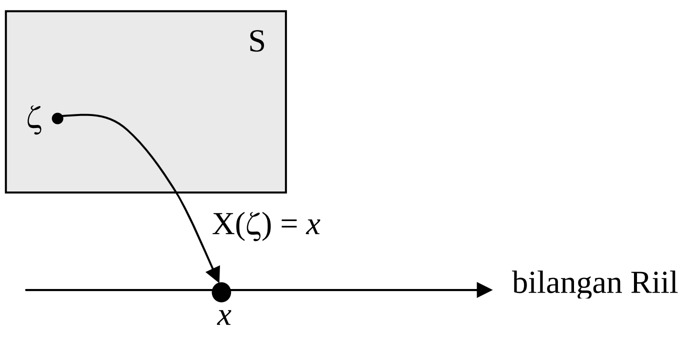
dimana ω_j adalah kelas ke-i dan **x** adalah vektor fitur. Aturan keputusan (pelabelan kelas pada vector masukan) adalah dengan memilih kelas ω_i yang memberikan $P(\omega_i|\mathbf{x})$ terbesar. Secara intuitif, ini berarti kita akan memilih kelas yang paling 'mungkin' untuk vektor fitur **x**. Setiap suku didalam Teorema Bayes memiliki nama khusus sebagai berikut

- $P(\omega_j)$: Peluang prior (dari kelas ω_j)
- $P(\omega_j|\mathbf{x})$: Peluang posterior (dari kelas ω_j jika diberikan pengamatan **x**)
- $P(\mathbf{x}|\omega_j)$: Tingkat kemungkinan (likelihood, peluang bersyarat pengamatan **x** jika diberikan kelas ω_j)
- $P(\mathbf{x})$: Konstanta normalisasi yang tidak mempengaruhi keputusan.

5.8 Peubah Acak

Pada saat melakukan suatu percobaan acak biasanya orang tertarik untuk mengkuantifikasi hasil percobaan. Misalnya:

- Saat melakukan sampling populasi, kita ingin mengetahui berat (seseorang)
- Saat menguji kinerja dua komputer, kita ingin tahu kecepatan masing-masing komputer
- Saat mengenali pesawat musuh, kita ingin mengukur parameter yang berhubungan dengan bentuknya atau penampang lintang radar (RCS-*Radar Cross Section*).
- dan seterusnya



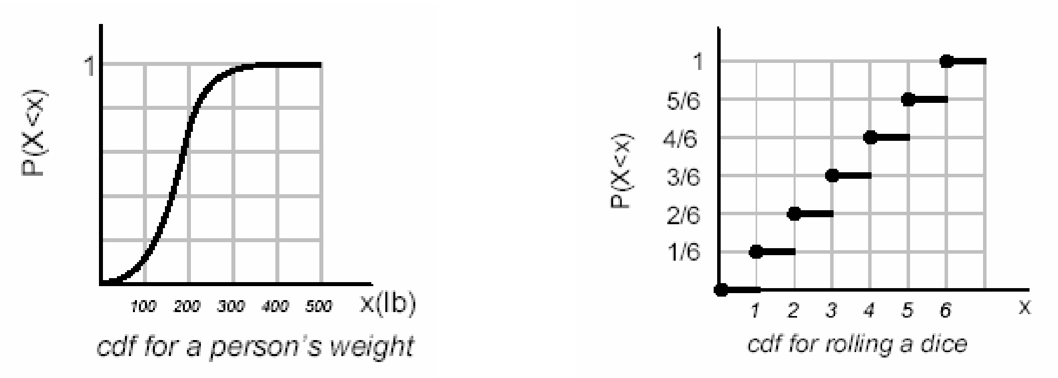
Gambar 5.17 Pemetaan oleh peubah acak

Contoh-contoh tersebut menunjukkan adanya konsep *peubah acak*. Peubah acak X adalah suatu fungsi yang memetakan setiap hasil eksperimen ζ didalam ruang sampel ke suatu bilangan riil x=X(ζ). Jadi sebenarnya peubah acak bukanlah peubah, melainkan fungsi yang memetakan hasil eksperimen ke suatu harga dengan cara yang deterministik, misalnya "*hitung jumlah muka dalam pelemparan koin sebanyak tiga kali*". Keacakan nilai yang diamati adalah akibat dari keacakan *argumen* dari fungsi X, yang disebut sebagai hasil eksperimen ζ. Peubah acak dapat bernilai diskrit, misalnya nilai yang keluar pada pelemparan dadu atau kontinyu, misalnya berat seseorang

5.9 Fungsi Distribusi Kumulatif (cdf : *cumulative distribution function*)

Fungsi distribusi kumulatif (cdf) $F_X(x)$ dari peubah acak X didefinisikan sebagai peluang dari kejadian $\{X \le x\}$

$$F_X(x) = P(X \le x) \text{ untuk } -\infty < X < +\infty \qquad (5.8)$$



(a) (b)

Gambar 5.18 Fungsi distribusi kumulatif: (a) kontinyu dan (b) diskrit

Fungsi distribusi kumulatif memiliki sifat-sifat berikut ini:

- $0 \le F_X(X) \le 1$
- $\lim_{x\to\infty} F_X(x) = 1$
- $\lim_{x\to 0} F_X(x) = 0$
- $F_X(a) \le F_X(b)$ jika $a \le b$
- $F_X(b) = \lim_{h\to 0} F_X(b+h) = F_X(b^+)$

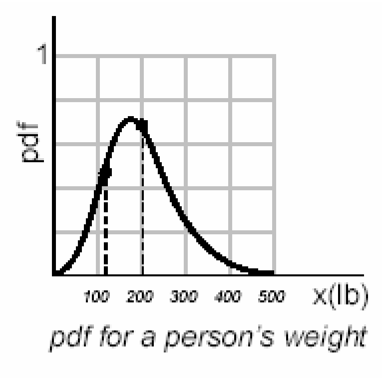
Fungsi Kerapatan Peluang (pdf: *probability density function*)

Fungsi kerapatan peluang (pdf) dari peubah acak kontinyu X, jika ada, didefinisikan sebagai turunan dari fungsi distribusi peluang $F_X(x)$.

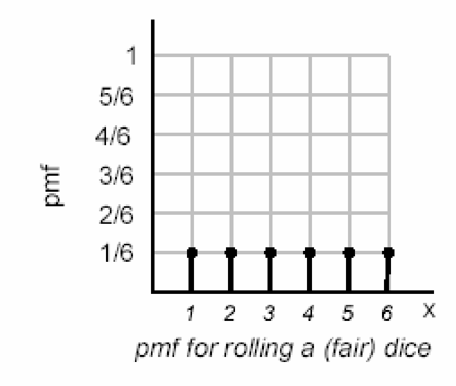
$$f_X(x) = \frac{dF_X(x)}{dx} \tag{5.9}$$

Untuk peubah acak diskrit, fungsi yang ekivalen dengan pdf adalah fungsi distribusi massa

$$f_X(x) = \frac{\Delta F_X(x)}{\Delta x} \tag{5.10}$$



(a)



(b)

Gambar 5.19 Fungsi kerapatan peluang: (a) kontinyu dan (b) diskrit

Fungsi rapat peluang memiliki sifat-sifat berikut:

- $f_X(x) > 0$
- $P(a < x < b) = \int_a^b f_X(x)\,dx$
- $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx$

- $1=\int_{-\infty}^{\infty} f_X(x)\,dx$
- $f_X(x \mid A)=\frac{d}{dx}F_X(x \mid A)$ dimana $F_X(x \mid A)=\frac{P(\{X<x\}\cap A)}{P(A)}$ jika $P(A)>0$

Ada hal-hal berkaitan dengan fungsi rapat peluang yang harus diperhatikan. Dari pdf di sebelah kiri pada gambar di atas, berapakah peluang seseorang punya berat 200 lb? Berdasarkan kurva, harganya sekitar 0.62 dan bilangan ini nampak masuk akal. Sekarang, berapakah peluang seseorang memiliki berat 124.876? Menurut kurva pdf harganya sekitar 0.43. Akan tetapi, secara intuitif kita tahu peluangnya haruslah sangat kecil (hampir nol). Bagaimana menjelaskan paradoks ini?

Sebenarnya pdf tidak mendefinisikan nilai peluang, melainkan rapat peluang. Untuk memperoleh nilai peluang yang sesungguhnya, kita harus mengintegrasikannya pada suatu selang tertentu. Jadi, pertanyaan yang tepat seharusnya adalah: Berapakah peluang seseorang memiliki berat 124.876 lb plus-minus 2 lb?

Di sisi lain, fungsi peluang massa (pmf-*probability mass function*) membeikan nilai peluang yang sebenarnya (inilah alasan penyebutan 'massa' dan bukan 'kerapatan'). Pmf ini memberi indikasi bahwa peluang sebarang bilangan pada dadu saat dilantunkan adalah sama, dan sama dng 1/6; suatu jawaban yg sangat masuk akal. Dengan demikian pmf tidak perlu diintegrasikan untuk mendapatkan nilai peluang (pada *x* tertentu).

5.10 Karakterisasi Statistik dari Peubah Acak

Untuk mengkarakterisasi suatu peubah acak, cdf atau pdf sudah sepenuhnya mencukupi. Namun demikian, peubah acak dapat pula dikarakterisasi secara parsial dengan *measure* yang lain, yakni:

- Ekspektasi/ nilai harap: nilai harap dapat dipakai untuk merenyatakan pusat masa dari suatu kerapatan.

$$E(X)=\mu=\int_{-\infty}^{\infty} x f_X(x)\,dx \tag{5.11}$$

- Variansi: variansi menyatakan sebaran disekitar nilai rata-rata

$$VAR(X)=E\left(\left(X-E(X)\right)^2\right)=\int_{-\infty}^{\infty}(x-\mu)^2\, f_X(x)\,dx \tag{5.12}$$

- Simpangan baku: akar kuadrat dari variansi

$$STD(X)=\sqrt{VAR(X)} \tag{5.13}$$

- Momen ke-N

$$E\left(X^N=\int_{-\infty}^{\infty} x^N f_X(x)\,dx\right) \tag{5.14}$$

Vektor Acak

Vektor acak adalah perluasan dari konsep peubah acak. Lebih spesifik lagi, vektor dari suatu peubah acak **X** adalah fungsi yang memetakan hasil percobaan ζ dari ruang sampel **S**, ke vektor bilangan riil. Vektor acak diberi notasi sebagai suatu vektor kolom. Selanjutnya cdf dan pdf digantikan dengan fungsi distribusi kumulatif gabungan (*joint cdf*) dan fungsi rapat peluang gabungan (*joint pdf*). Untuk suatu vektor acak $\mathbf{X}=\left[X_1\, X_2 \ldots X_N\right]^T$, cdf dan pdf gabungan didefinisikan sebagai:

- Cdf gabungan: $F_X(X)=P_X\left(\{X_1 \le x_1\}\cap\{X_2 \le x_2\}\cap\ldots\cap\{X_N \le x_N\}\right)$ (5.15)
- Pdf gabungan: $f_X(x)=\dfrac{\partial^N F_X(x)}{\partial x_1 \partial x_2 \ldots \partial x_N}$ (5.16)

Istilah *marginal pdf* dipakai untuk menyatakan pdf dari himpunan bagian semua dimensi vektor acak. M*arginal pdf* diperoleh dng mengintegrasikan peubah selain yang 'diinginkan', misalnya, untuk vektor acak dimensi dua **X** = $[x_1\ x_2]^T$, maka *marginal pdf* untuk x_1 untuk *joint pdf* $f_{X_1X_2}(x_1x_2)$ adalah

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1X_2}(x_1x_2)\,dx_2 \tag{5.17}$$

Vektor acak juga sepenuhnya dikarakterisasikan oleh *joint* cdf mapun *joint pdf.* Demikian pula vektor acak bisa digambarkan secara parsial dengan *measure* yang serupa, yakni:

- Vektor mean (rata-rata)

$$E(X) = \left[E(X_1)\ E(X_2) \ldots E(X_N)\right]^T = \left[\mu_1\ \mu_2 \ldots \mu_N\right] = \boldsymbol{\mu} \tag{5.18}$$

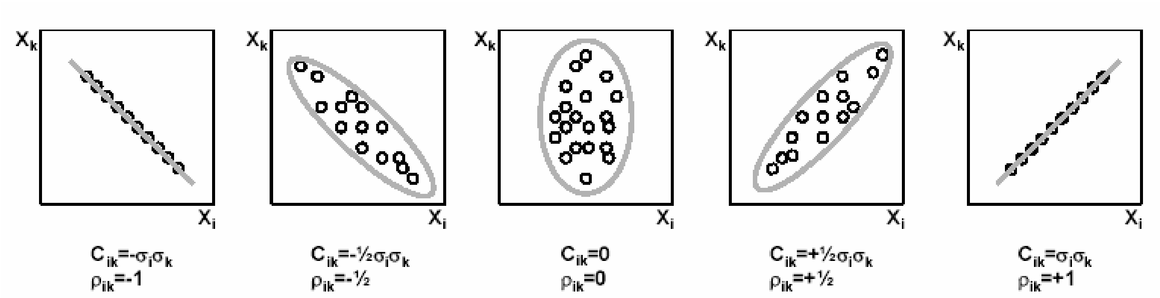
- Matriks kovariansi

$$\begin{aligned} COV(\mathbf{X}) = \boldsymbol{\Sigma} = E\left((\mathbf{X}-\mu)(\mathbf{X}-\mu)^T\right) \\ = \begin{pmatrix} E\left((X_1-\mu_1)(X_1-\mu_1)\right) & \ldots & E\left((X_1-\mu_1)(X_N-\mu_N)\right) \\ \ldots & \ldots & \ldots \\ E\left((X_N-\mu_N)(X_1-\mu_1)\right) & \ldots & E\left((X_N-\mu_N)(X_N-\mu_N)\right) \end{pmatrix} \\ = \begin{pmatrix} \sigma_1^2 & \ldots & C_{1N} \\ \ldots & \ldots & \ldots \\ C_{N1} & \ldots & \sigma_N^2 \end{pmatrix} \end{aligned} \tag{5.19}$$

Matriks kovariansi menyatakan tendensi dari setiap pasang fitur untuk berubah (vary) secara bersamaan (jadi, *co-vari*). Kovariansi memiliki berbagai sifat penting berikut ini

- Jika x_i dan x_k cenderung naik bersamaan, maka c_{ik}>0
- Jika x_i cenderung turun saat x_k naik, maka c_{ik}<0
- Jika x_i dan x_k tak berkorelasi, maka c_{ik}=0
- $|c_{ik}| \leq \sigma_i\sigma_k$, dimana σ_i simpangan baku dari x_i
- $c_{ii} = \sigma_i^2 = \mathrm{VAR}(x_i)$

Suku kovariansi dapat dinyatakan sebagai $c_{ii} = \sigma_i^2$ dan $c_{ik} = \rho_{ik}\sigma_i\sigma_k$ dimana ρ_{ik} disebut sebagai koefisien korelasi. Berbagai sifat kovariansi diatas dilukiskan pada gambar berikut ini.



Gambar 5.20 Ilistrasi berbagai sigat kovariansi

Matriks kovariansi juga bisa ditulis sebagai

$$\mathbf{\Sigma} = E\left((\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right) = E\left(\mathbf{X}\mathbf{X}^T\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T = \mathbf{S} - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

$$\text{dengan } \mathbf{S} = E\left(\mathbf{X}\mathbf{X}^T\right) = \begin{pmatrix} E(\mathbf{X}_1\mathbf{X}_1) & \ldots & E(\mathbf{X}_1\mathbf{X}_N) \\ \ldots & \ldots & \ldots \\ E(\mathbf{X}_N\mathbf{X}_1) & \ldots & E(\mathbf{X}_N\mathbf{X}_N) \end{pmatrix} \tag{5.20}$$

dengan S matriks autokorelasi yang mengandung informasi sama banyaknya dengan matriks kovariansi. Disamping bentuk di atas, matriks kovariansi dapat pula ditulsikan sebagai

$$\mathbf{\Sigma} = \mathbf{\Gamma R \Gamma} = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \sigma_N \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1N} \\ \rho_{12} & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{1N} & \ldots & \ldots & \sigma_N \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \sigma_N \end{pmatrix} \tag{5.21}$$

Bentuk di atas adalah formulasi yang lebih baik karena Γ mengandung skala fitur dan **R** mempertahankan informasi esensial yang berisi hubungan antar fitur. Disini **R** adalah matriks korelasi.

Korelasi juga berhubungan juga dengan kebebasan data. Dua buah peubah acak x_i dan x_k tak-berkorelasi jika E[x_1 x_k] =E[x_1] E[x_k]. Peubah yang tak berkorelasi disebut juga sebagai peubah yang bebas linier. Dua peubah acak x_i dan x_k disebut saling bebas jika jika P(x_i,x_k) =P(x_i) P(x_k).

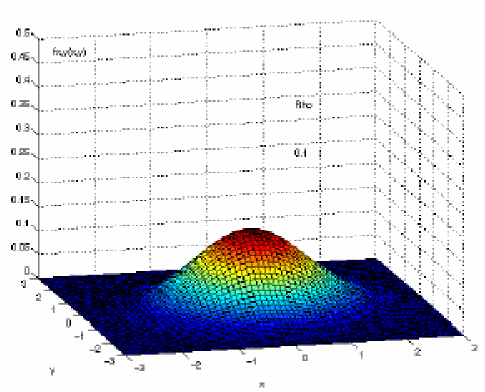
5.11 Distribusi Normal atau Distribusi Gauss

Distribusi Normal (Gauss) multivariate N(μ, Σ) didefinisikan sebagai

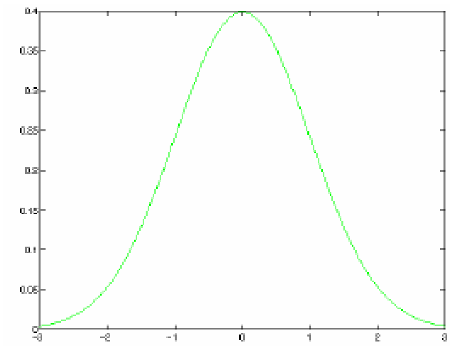
$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right) \tag{5.22}$$

Untuk kasus satu dimensi, bentuk ini disederhanakan menjadi

$$f_X(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right) \tag{5.23}$$



(a)



(b)

Gambar 5.21 Distribusi Gaussian: (a) dua dimensi dan (b) satu dimensi

Distribusi Gauss sangat populer karena perimbangan-pertimbangan berikut:

- Cukup dengan parameter (μ,Σ), fungsi distribusi tergambarkan secara unik
- Jika x_i tak saling berkorelasi (c_{ik}=0), maka mereka juga saling bebas. Dengan demikian matriks kovariansinya akan berbentuk matriks diagonal dengan masing-masing variansi berada pada diagonal utama
- *Central Limit Theorem*
- Kerapatan marginal dan kondisionalnya juga berbentuk Gaussian

- Sebarang transformasi linier dari N peubah acak Gaussian akan menghasilkan N buah peubah acak yang juga Gaussian. Jika $\mathbf{X} = [x_1\ x_2\ \ldots\ x_N]^T$ yang secara bersamaan Gaussian, dan A matriks N×N yang *invertibel*, maka $\mathbf{Y} = \mathbf{AX}$ juga akan Gaussian secara bersamaan.

Gambaran fungsi distribusi (kiri) dan distribusi marjinalnya

$$f_Y(y) = \frac{f_X\left(\mathbf{A}^{-1}\mathbf{y}\right)}{|\mathbf{A}|} \tag{5.24}$$

Central Limit Theorem (CLT). Teorem ini menyatakan bahwa jika ada distribusi dengan rata-rata μ dan variansi σ_i^2, pencuplikan sebaran dari rata-rata akan mendekati distribusi normal dengan rata-rata μ dan variansi σ_i^2/N, ketika jumlah sampel N meningkat. Apapun bentuk sebaran asal, distribusi pencuplikan dari rata-ratanya akan menuju ke distribusi normal. Ingat bahwa N adalah besarnya sampel untuk setiap rata-rata, bukan banyaknya sample. Ide diatas digambarkan dengan kasus distribusi seragam berikut: Dilakukan 500 kali percobaan

- Untuk N=1, satu sampel diambil dan rata-rata nya (untuk masing-masing 500 eksperimen) dicatat. Tentu hasilnya distribusi seragam
- Untuk N=4, mak 4 buah sampel diambil dan rata-rata nya (untuk masing-masing 500 eksperimen) dicatat. Bentuk kurva Gaussian akan mulai muncul
- Begitru seterusnya untuk N=7 dan N=10. Seiring naiknya N, histogram semakin menunjukkan distribusi Normal