Pemisahan Berdasarkan Teori Keputusan Bayes

Octa Heriana, 05906-TE
Sumarna, 06248-TE
Jurusan Teknik Elektro FT UGM,
Yogyakarta

2.1 PENGANTAR

Bersumber dari variasi sifat-sifat statistik pada suatu pola dan derau pada sensor-sensor pengukuran memungkinkan untuk merancang suatu model penggolongan dalam sistem pengenalan pola dengan pedekatan berdasarkan pada pengaturan argumen probabilistik sifat statistik pada ciri-ciri yang digenerasi. Salah satunya adalah penggolongan pola yang tidak diketahui ke dalam suatu kemungkinan terbaik (*most probable*) pada kelas-kelas tertentu.

Misalkan penggolongan ke dalam sejumlah M kelas (ω_1, ω_2, ω_3, ... , ω_M) dari sebuah pola yang tidak diketahui yang direpresentasikan oleh satu vektor ciri $\boldsymbol{x}$. Langkah awal adalah menyusun probabilitas bersyarat pada M dengan $P(\omega_i \mid \boldsymbol{x})$ dengan i = 1, 2, 3, ..., M. Setiap nilai $P(\omega_i \mid \boldsymbol{x})$ menggambarkan satu probabilitas bahwa pola-pola yang tidak diketahui termasuk pada masing-masing kelas ω_i dengan ketentuan bahwa vektor ciri yang sesuai bernilai $\boldsymbol{x}$. Dasar pengelompokan (penggolongan ke dalam suatu kelas) yang dipertimbangkan dapat berupa : nilai M terbesar, atau adanya kesamaan suatu nilai, atau ketepatan maksimum terhadap fungsi yang telah lebih dahulu ditetapkan. Pola yang tidak diketahui kemudian dimasukkan ke dalam satu kelas yang sesuai.

2.2 TEORI KEPUTUSAN BAYES

Misalkan ω_1 dan ω_2 merupakan dua kelas yang memuat pola-pola yang tidak diketahui. Diasumsikan bahwa probabilitas apriori $P(\omega_1)$ dan $P(\omega_2)$ diketahui. Jika N jumlah total dari pola-pola yang tersedia, kemudian N_1 dan N_2 masing-masing termasuk dalam ω_1 dan ω_2, maka

$$P(\omega_1) = \frac{N_1}{N} \quad \text{dan} \quad P(\omega_2) = \frac{N_2}{N}.$$

Kuantitas statistik lain yang dikenal adalah fungsi kerapatan probabilitas kelas bersyarat adalah $p(\boldsymbol{x} \mid \omega_i)$ dengan i = 1, 2 yang mendiskripsikan distribusi vektor ciri pada masing-masing kelas. Vektor ciri dapat bernilai sembarang pada ruang ciri yang berdimensi l. Pada kasus vektor ciri yang diskrit, fungsi kerapatan $p(\boldsymbol{x} \mid \omega_i)$ menjadi probabilitas dan dinyatakan sebagai $P(\boldsymbol{x} \mid \omega_i)$. Aturan Bayes menyebutkan bahwa:

$$P(\omega_i \mid \boldsymbol{x}) = \frac{p(\vec{x}|\omega_i)P(\omega_i)}{p(\vec{x})} \qquad (2\text{-}1)$$

dengan

$$p(\boldsymbol{x}) \;=\; p(\vec{x}) \;=\; \sum_{i=1}^{2} p(\vec{x}|\omega_i)P(\omega_i) \qquad (2.2)$$

$$=\; p(\vec{x}|\omega_1)P(\omega_1) \;+\; p(\vec{x}|\omega_2)P(\omega_2)\,.$$

Aturan pengelompokan Bayes sekarang dapat dinyatakan sebagai :

jika $P(\omega_1 \mid \boldsymbol{x}) > P(\omega_2 \mid \boldsymbol{x})$, maka $\boldsymbol{x}$ dikelompokkan ke ω_1 dan

jika $P(\omega_1 \mid \boldsymbol{x}) < P(\omega_2 \mid \boldsymbol{x})$, maka $\boldsymbol{x}$ dikelompokkan ke ω_2.

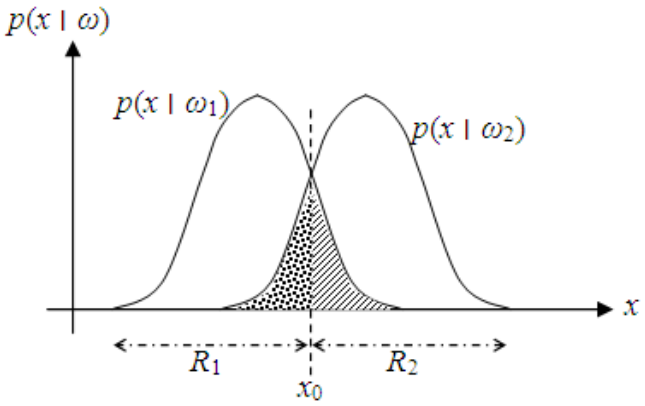
Hal yang mengganggu adalah jika $P(\omega_1 \mid \boldsymbol{x}) = P(\omega_2 \mid \boldsymbol{x})$, tetapi pola dapat dimasukkan ke dalam salah satu kelas. Selain itu, pengelompokan yang menggunakan persamaan (2-1) juga dapat didasarkan pada kesamaan atau ketidak-samaan.

$$p(\vec{x}|\omega_1)P(\omega_1) \;\neq\; p(\vec{x}|\omega_2)P(\omega_2)$$

maka $p(\vec{x})$ tidak akan diperhitungkan, karena sama untuk setiap kelas.

Jika $P(\omega_1) = P(\omega_2) = \frac{1}{2}$, maka $p(\vec{x}|\omega_1) \neq p(\vec{x}|\omega_2)$.

Selanjutnya, pencarian nilai maksimum pada fungsi kerapatan probabilitas bersyarat yang dievaluasi pada $\boldsymbol{x}$. Ditinjau dua kelas yang berpeluang sama dan memiliki variasi pada $p(\boldsymbol{x} \mid \omega_i)$ sebagai fungsi x pada kasus sederhana dengan ciri tunggal ($l = 1$).



Gambar 2.1 *Contoh dari dua wilayah R_1 dan R_2 yang dibentuk oleh pemisah Bayesian*

Garis patah-patah pada x_0 merupakan batas partisi ruang ciri dua daerah R_1 dan R_2. Berdasar aturan keputusan Bayes, untuk semua nilai x pada R_1 masuk di dalam kelas ω_1 dan untuk semua nilai x pada R_2 masuk di dalam kelas ω_2. Tetapi ada peluang x yang terletak pada R_2 dan pada saat yang sama termasuk pada kelas ω_1, maka keputusannya menjadi salah dan sebaliknya. Dengan demikian total peluang memasukkan keputusan yang salah diberikan sebagai :

$$2P_e = \int_{-\infty}^{x_0} p(\vec{x}|\omega_2)dx + \int_{x_0}^{+\infty} p(\vec{x}|\omega_1)dx$$

yang sama dengan luas total daerah irisan (terarsir) kedua kurva.

Minimalisasi Probabilitas Kesalahan Klasisikasi

Klasifikasi berdasarkan Bayesian adalah optimal dengan minimalisasi probabilitas kesalahan klasifikasi. Dengan menggerakkan garis batas menjauh dari x_0 selalu menambah luas arsiran yang berkorespondensi pada daerah irisan kedua kurva. Misalkan R_1 merupakan daerah ruang ciri tempat penentuan pilihan ω_1 dan R_2 adalah daerah yang berkorespondensi dengan ω_2. Suatu kesalahan terjadi jika $\boldsymbol{x} \in R_1$ walaupun termasuk pada ω_2, atau jika $\boldsymbol{x} \in R_2$ walaupun termasuk pada ω_1.

$$\begin{aligned} P_e &= P(\boldsymbol{x} \in R_2, \omega_1) + P(\boldsymbol{x} \in R_1, \omega_2) \\ &= \int_{R_2} P(\omega_1|\vec{x})p(\vec{x})dx + \int_{R_1} P(\omega_2|\vec{x})p(\vec{x})dx \end{aligned}$$

di mana $P(\text{-} , \text{-})$ merupakan *joint-probability* dari dua peristiwa. Terlihat bahwa kesalahan diminimalisasi jika daerah partisi R_1 dan R_2 dari ruang ciri dipilih sedemikian hingga :

$$R_1 : P(\omega_1 \mid \boldsymbol{x}) > P(\omega_2 \mid \boldsymbol{x})$$
$$R_2 : P(\omega_2 \mid \boldsymbol{x}) > P(\omega_1 \mid \boldsymbol{x}).$$

Karena gabungan daerah R_1 dan R_2 mencakup seluruh ruang, berdasarkan definisi fungsi rapat probabilitas berlaku :

$$\int_{R_1} P(\omega_1|\vec{x})p(\vec{x})dx + \int_{R_2} P(\omega_2|\vec{x})p(\vec{x})dx = P(\omega_1)$$

dengan demikian dapat dituliskan

$$P_e = P(\omega_1) - \int_{R_1} \{P(\omega_1|\vec{x}) - P(\omega_2|\vec{x})\}p(\vec{x})dx$$

yang berarti bahwa probabilitas kesalahan diminimalisasi jika R_1 merupakan daerah dari ruang di mana $P(\omega_1 \mid \boldsymbol{x}) > P(\omega_2 \mid \boldsymbol{x})$. Kemudian R_2 menjadi daerah yang sebaliknya benar. Generalisasi pada kasus banyak kelas (*multiclass*), untuk total M kelas (ω_1, ω_2, ω_3, ... , ω_M), satu pola tidak diketahui, direpresentasikan dengan vektor ciri $\boldsymbol{x}$, ditempatkan pada kelas ω_i jika :

$$P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x}) \forall\, j \neq i\,. \qquad (2\text{-}13)$$

Minimalisasi Resiko Rerata

Penggolongan berdasarkan probabilitas kesalahan tidak selalu merupakan kriteria terbaik yang dipilih untuk minimalisasi, karena menetapkan kepentingan yang sama untuk seluruh kesalahan. Banyak kasus di mana satu kesalahan berdampak lebih serius dari pada yang lain. Dalam kasus tersebut lebih cocok untuk memasukkan suku pinalti untuk memboboti setiap kesalahan. Misalkan pada persoalan kelas sebesar M di mana R_j (j = 1, 2, 3, ... , M) merupakan daerah ruang ciri yang masing-masing ditempati kelas ω_j. Diasumsikan vektor ciri $\vec{x}$ milik kelas ω_k terletak pada R_i ($i \neq k$). Kemudian salah mengklasifikasi vektor tersebut ke dalam ω_i dan satu kesalahan ditetapkan. Suku pinalti λ_{ki}, disebut *loss* (rugi), terkait dengan keputusan yang salah. Suatu matriks $\boldsymbol{L}$, dengan (k,i) lokasi suku pinalti yang bersesuaian, disebut sebagai *loss matrix*. Resiko atau *loss* yang terkait dengan ω_k didefinisikan sebagai

$$r_k = \sum_{i=1}^{M} \lambda_{ki} \int_{R_i} p(\vec{x}|\omega_k)dx\,.$$

Integral tersebut merupakan probabilitas keseluruhandari satu vektor ciri pada kelas ω_k yang dikelompokkan ke dalam ω_i. Probabilitas ini diberi bobot λ_{ki}. Tujuannya sekarang adalah memilih daerah partisi R_j sedemikian hingga resiko reratanya

$$r = \sum_{k=1}^{M} r_k P(\omega_k)$$

$$= \sum_{i=1}^{M} \int_{R_i} \left(\sum_{k=1}^{M} \lambda_{ki}\ p(\vec{x}|\omega_k)P(\omega_k)\right) dx$$

diminimalisasi.

Hal ini tercapai jika setiap integral diminimalisasi, yang mana ini ekivalen dengan pemilihan daerah partisi sedemikian hingga $\vec{x} \in R_i$ jika

$$l_i \equiv \sum_{k=1}^{M} \lambda_{ki} p(\vec{x}|\omega_k)P(\omega_k) < l_i \equiv \sum_{k=1}^{M} \lambda_{kj} p(\vec{x}|\omega_k)P(\omega_k) \quad \forall\, j \neq i \qquad (2\text{-}16)$$

Untuk kasus dua-kelas diperoleh :

$l_1 = \lambda_{11} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\vec{x}|\omega_2)P(\omega_2)$

$l_2 = \lambda_{12} p(\vec{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\vec{x}|\omega_2)P(\omega_2)$.

$\vec{x}$ dimasukkan ke ω_1 jika $l_1 < l_2$, yaitu

$(\lambda_{21} - \lambda_{22}) p(\vec{x}|\omega_2)P(\omega_2) < (\lambda_{12} - \lambda_{11}) p(\vec{x}|\omega_1)P(\omega_1)$ dengan asumsi $\lambda_{ij} > \lambda_{ii}$.

Dengan asumsi tersebut, maka kriteria pengambilan keputusan (2-16) pada kasus dua-kelas menjadi

$$\vec{x} \in \omega_1 \text{ jika } l_{12} \equiv \frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}$$

atau

$$\vec{x} \in \omega_2 \text{ jika } l_{12} \equiv \frac{p(\vec{x}|\omega_1)}{p(\vec{x}|\omega_2)} < \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}$$

Rasio l_{12} disebut sebagai **rasio kemungkinan** dan uji yang mendahului disebut sebagai **uji rasio kemungkinan**.

2.3 FUNGSI DISKRIMINAN DAN DECISION SURFACE

Baik minimalisasi resiko ataupun probabilitas kesalahan adalah ekivalen dengan partisi ruang ciri ke dalam M daerah pada persoalan dengan M kelas. Jika daerah R_i, R_j kejadian yang bersebelahan, maka mereka dipisahkan dengan *decision surface* dalam ruang ciri multidimensi. Untuk kasus probabilitas kesalahan minimum dideskripsikan dengan persamaan

$$P(\omega_i|\vec{x}) - P(\omega_j|\vec{x}) = 0.$$

Dari satu sisi permukaan selisih itu positif dan dari sisi yang lain adalah negatif. Di samping bekerja secara langsung dengan probabilitas (atau fungsi resiko), kadang lebih cocok, dari pandangan matematis, untuk bekerja dengan fungsi yang ekivalen, misalnya

$$g_i(\vec{x}) \equiv f(P(\omega_i|\vec{x}))$$

dengan *f*(.) merupakan sebuah fungsi yang makin besar secara monotonik, $g_i(\vec{x})$ disebut sebagai **fungsi diskriminan**. Uji keputusan persamaan (2-13) sekarang dapat dinyatakan sebagai menggolongkan $\vec{x}$ ke dalam ω_i jika $g_i(\vec{x}) > g_j(\vec{x}) \quad \forall j \neq i$.

Decision surface, pembatas daerah-daerah yang bersebelahan, dideskripsikan sebagai

$g_{ij}(\vec{x}) \equiv g_i(\vec{x}) - g_j(\vec{x}) = 0$ dengan $i, j = 1, 2, \ldots, M$ dan $i \neq j$.

2.4 PENGGOLONGAN BAYESIAN UNTUK DISTRIBUSI NORMAL

Salah satu fungsi kerapatan probabilitas yang sering dijumpai dalam praktek adalah fungsi kerapatan normal atau Gaussian. Sekarang diasumsikan bahwa fungsi kemungkinan ω_i terhadap $\vec{x}$ di dalam ruang ciri berdimensi-*l* mengikuti kerapatan normal multivariasi umum sebagai

$$p(\vec{x}|\omega_i) = \frac{1}{(2\pi)^{l/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\}, \qquad i = 1, 2, \ldots, M$$

di mana $\vec{\mu}_i = \boldsymbol{E}[\vec{x}]$ merupakan nilai rerata dari kelas ω_i dan Σ_i matriks kovarian *l* x *l* yang didefinisikan sebagai

$$\Sigma_i = \boldsymbol{E}[(\vec{x}-\vec{\mu}_i)(\vec{x}-\vec{\mu}_i)^T]$$

$|\Sigma_i|$ menyatakan determinan Σ_i dan $\boldsymbol{E}[.]$ adalah nilai rerata (nilai harap) dari suatu variabel acak. Kadang digunakan simbol $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ untuk menyatakan pdf Gaussian dengan nilai rerata $\boldsymbol{\mu}$ dan kovarian Σ.

Desain klasifikasi Bayesian, karena bentuk eksponensial dari kerapatan yang tercakup, lebih baik bekerja dengan fungsi diskriminan yang memuat fungsi logaritmik (monotonik) ln(.) sebagai

$$g_i(\vec{x}) = \ln(p(\vec{x}|\omega_i)P(\omega_i)) = \ln p(\vec{x}|\omega_i) + \ln P(\omega_i)$$

atau

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i) + \ln P(\omega_i) + c_i \qquad (2\text{-}26)$$

di mana c_i suatu konstanta yang bernilai $c_i = -(l/2)\ln 2\pi - (1/2)\ln|\Sigma_i|$.

Ekspansi persamaan (2-26) dapat diperoleh

$$g_i(\vec{x}) = -\frac{1}{2}\,\vec{x}^T\,\Sigma_i^{-1}\,\vec{x} + \frac{1}{2}\,\vec{x}^T\,\Sigma_i^{-1}\,\vec{\mu}_i - \frac{1}{2}\,\vec{\mu}_i^T\,\Sigma_i^{-1}\,\vec{\mu}_i + \frac{1}{2}\,\vec{\mu}_i^T\,\Sigma_i^{-1}\,\vec{x} + \ln P(\omega_i) + c_i \qquad (2\text{-}27)$$

Pada umumnya persamaan (2-27) berbentuk kuadratik nonlinier. Sebagai contoh ditinjau kasus $l = 2$ dan diasumsikan bahwa

$$\Sigma_i = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{bmatrix}.$$

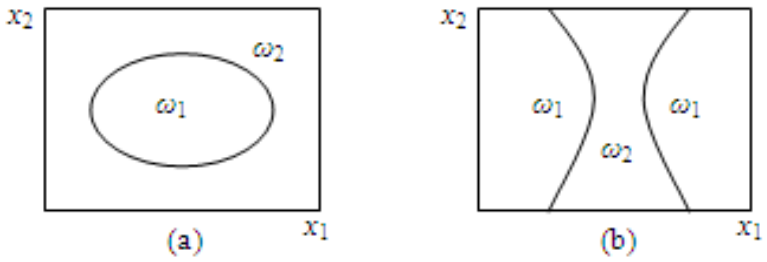
Persamaan (2-27) menjadi

$$g_i(\vec{x}) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln P(\omega_i) + c_i$$

dan dengan jelas kurva keputusan yang terkait $g_i(\vec{x}) - g_j(\vec{x}) = 0$ adalah kuadrik (misalnya ellipsoid, parabola, hiperbola, dan pasangan garis). Untuk $l > 2$ *decision surface* merupakan *hyperquadrics*. Gambar berikut menunjukkan kurva keputusan yang sesuai $P(\omega_1) = P(\omega_2)$, $\vec{\mu}_1 = [0, 0]^T$ dan $\vec{\mu}_2 = [1, 0]^T$. Matrik kovarian dua kelas adalah untuk :

gambar (a) $\Sigma_1 = \begin{bmatrix} 0{,}1 & 0{,}0 \\ 0{,}0 & 0{,}15 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0{,}2 & 0{,}0 \\ 0{,}0 & 0{,}25 \end{bmatrix}$

dan gambar (b) $\Sigma_1 = \begin{bmatrix} 0{,}1 & 0{,}0 \\ 0{,}0 & 0{,}15 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0{,}15 & 0{,}0 \\ 0{,}0 & 0{,}1 \end{bmatrix}$.



Gambar 2.2 *Kurva keputusan Kuadrik*

Keputusan Hyperplanes

Kontribusi kuadratik hanya datang dari suku $\vec{x}^T\Sigma_i^{-1}\vec{x}$. Jika diasumsikan matrik kovarian sama untuk semua kelas, yakni $\Sigma_i = \Sigma$, maka suku kuadratik akan sama untuk semua fungsi diskriminan. Sehingga hal

itu tidak masuk pembandingan dalam menghitung nilai maksimum dan karenanya dibuang dari persamaan decision surface. Hal yang sama berlaku untuk konstanta c_i. Sehingga mereka dapat dihilangkan dan $g_i(\vec{x})$ didefinisikan kembali sebagai

$$g_i(\vec{x}) = \vec{w}_i^T + w_{i0} \qquad (2\text{-}29)$$

dengan $\vec{w}_i = \Sigma^{-1}\vec{\mu}_i$ dan $w_{i0} = \ln P(\omega_i) - \frac{1}{2}\vec{\mu}_i^T \Sigma^{-1} \vec{\mu}_i$. Karena itu, $g_i(\vec{x})$ merupakan fungsi linier terhadap $\vec{x}$ dan masing-masing decision surface adalah hyperplanes.

Selanjutnya ditinjau pada matriks kovarian diagonal dengan elemen-elemen sama, dengan asumsi bahwa ciri individual, penyusun vektor ciri, adalah tidak berkorelasi timbal-balik dan dengan variansi sama ($E[(x_i - \mu_i)(x_j - \mu_j)] = \sigma^2\delta_{ij}$). Sehingga $\Sigma = \sigma^2 I$, di mana I adalah matriks identitas berdimensi-l, dan persamaan (2-29) menjadi

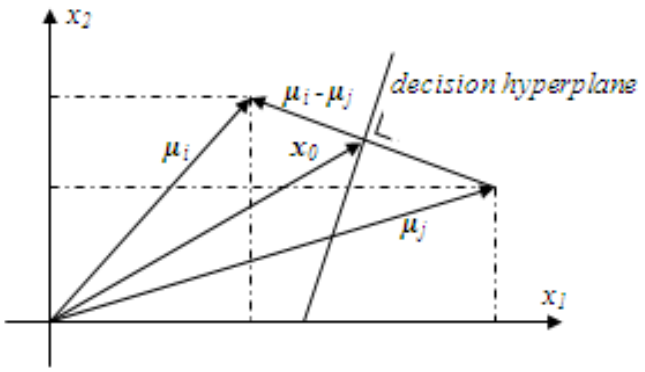
$$g_i(\vec{x}) = \frac{1}{\sigma^2}\vec{u}_i^T + w_{i0}.$$

Sehingga decision hyperplanes yang sesuai dapat dituliskan sebagai

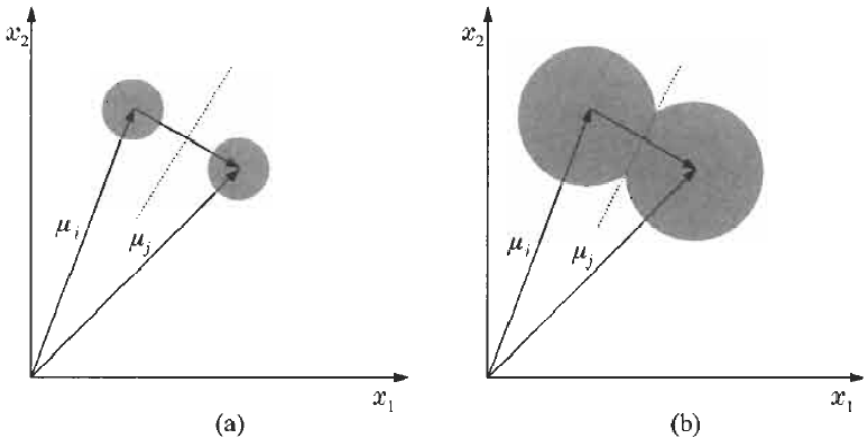
$$g_{ij}(\vec{x}) \equiv g_i(\vec{x}) - g_j(\vec{x}) = \vec{w}^T(\vec{x} - \vec{x}_0) = 0$$

dengan $\vec{w} = \vec{\mu}_i - \vec{\mu}_j$, dan $\vec{x}_0 = \frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\vec{\mu}_i - \vec{\mu}_j}{\left\|\vec{\mu}_i - \vec{\mu}_j\right\|^2}$

di mana $\|\vec{x}\| = \sqrt{x_1^2 + x_2^2 + ... + x_l^2}$ adalah Euclidean norm dari $\vec{x}$. Sehingga decision surface-nya merupakan suatu hyperplane yang melalui titik $\vec{x}_0$. Jika $P(\omega_i) = P(\omega_j)$, maka $\vec{x}_0 = \frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j)$ dan hyperplane tersebut melalui rerata dari $\vec{\mu}_i$, $\vec{\mu}_j$. Untuk kasus dua-dimensi, dengan $\Sigma = \sigma^2 I$, geometri decision line digambarkan sebagai berikut :



Gambar 2.3 *Garis keputusan untuk dua kelas dan vektor distribusi normal dengan* $\Sigma = \sigma^2 I$



Gambar 2.4 *Garis keputusan (a) kelas yang compact dan (b) kelas yang uncompact*

Tampak bahwa *decision hyperplane* (garis lurus) tegak lurus terhadap $\vec{\mu}_i - \vec{\mu}_j$. Sembarang titik $\vec{x}$ yang terletak pada *decision hyperplane*, maka vertor $\vec{x} - \vec{x}_0$ juga terletak pada *hyperplane* tersebut, dan

$$g_{ij}(\vec{x}) = 0 \Rightarrow \vec{w}^T(\vec{x} - \vec{x}_0) = (\vec{\mu}_i - \vec{\mu}_j)^T(\vec{x} - \vec{x}_0) = 0.$$

Maka $(\vec{\mu}_i - \vec{\mu}_j)$ tegak lurus terhadap *decision hyperplane*. Hal lain yang perlu ditekankan adalah bahwa *hyperplane* terletak lebih dekat ke $\vec{\mu}_i$ jika $P(\omega_i) < P(\omega_j)$ atau lebih dekat ke $\vec{\mu}_j$ jika $P(\omega_i) > P(\omega_j)$. Selanjutnya jika σ^2 kecil terhadap $\|\vec{\mu}_i - \vec{\mu}_j\|$, maka lokasi *hyperplane* agak tidak sensitif terhadap nilai $P(\omega_i)$, $P(\omega_j)$. Hal ini diharapkan karena variansi yang kecil menunjukkan bahwa vektor-vektor acak tercakup di dalam radius kecil di sekitar nilai reratanya. Sehingga pergeseran kecil *decision hyperplane* berakibat kecil pada hasilnya.

Berikutnya ditinjau pada matriks kovarian nondiagonal dengan *hyperplanes* yang dideskripsikan sebagai

$$g_{ij}(\vec{x}) = \vec{w}^T(\vec{x} - \vec{x}_0) = 0$$

dengan $\vec{w} = \Sigma^{-1}(\vec{\mu}_i - \vec{\mu}_j)$, dan $\vec{x}_0 = \frac{1}{2}(\vec{\mu}_i + \vec{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\vec{\mu}_i - \vec{\mu}_j}{\|\vec{\mu}_i - \vec{\mu}_j\|^2_{\Sigma^{-1}}}$

di mana $\|\vec{x}\|_{\Sigma^{-1}} = (\vec{x}^T\ \Sigma^{-1}\ \vec{x})^{1/2}$ yang juga disebut Σ^{-1} *norm* dari $\vec{x}$. Pada kasus ini *decision hyperplane* tidak selamanya tegak lurus terhadap vektor $\vec{\mu}_i - \vec{\mu}_j$ tetapi terhadap transformasi liniernya $\Sigma^{-1}(\vec{\mu}_i - \vec{\mu}_j)$.

Penggolongan Jarak Minimum

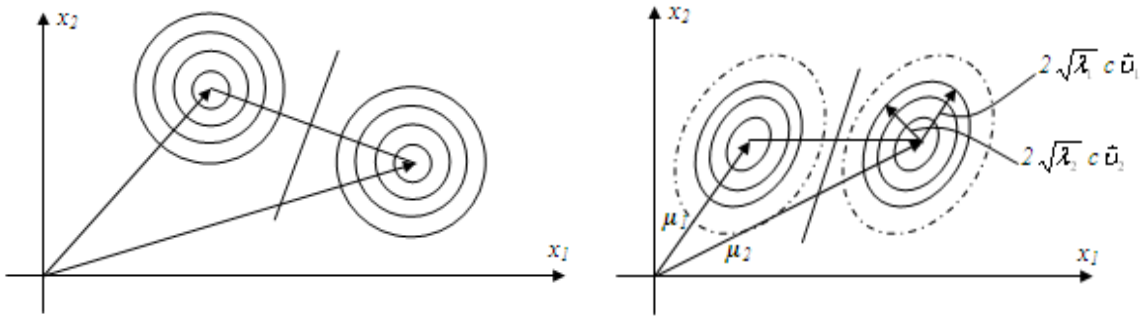
Diasumsikan kelas-kelas *equiprobable* (berkeboleh-jadian setara) dengan matriks kovarian sama, maka $g_i(\vec{x})$ pada persamaan (2-26) disederhanakan menjadi

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)$$ konstantanya diabaikan.

Pada kasus $\Sigma = \sigma^2 I$ maksimum dari $g_i(\vec{x})$ adalah minimum dari jarak *Euclidean* (d_ϵ).

$$d_\epsilon = \left\|\vec{x}-\vec{\mu}_i\right\|.$$

Maka vektor-vektor ciri dimasukkan ke dalam kelas-kelas sesuai dengan jarak *Euclidean*-nya dari titik-titik rerata masing-masing. Gambar berikut menunjukkan kurva-kurva berjarak sama $d_\epsilon = c$ dari titik-titik rerata untuk setiap kelas. Semuanya berupa lingkaran dengan jejari c (dalam kasus yang lebih umum merupakan *hyperspheres*).



(a) (b)

Gambar 2.5 *Kurva (a) jarak Euclidean dan (b) jarak Mahalanobis dari titik-titik mean tiap kelas*

Berikutnya, pada kasus maksimalisasi $g_i(\vec{x})$ adalah ekivalen dengan minimalisasi Σ^{-1} *norm*, yang dikenal sebagai jarak *Mahalanobis* (d_m) sebagai :

$$d_m = \left((\vec{x}-\vec{\mu}_i)^T \Sigma^{-1}(\vec{x}-\vec{\mu}_i)\right)^{1/2} \qquad (2\text{-}41)$$

Pada kasus ini kurva-kurva $d_m = c$ berjarak konstan adalah ellips (*hyperellipses*). Sebenarnya matriks kovarian adalah simetris selalu dapat didiagonalisasi dengan transformasi *unitary* sebagai

$$\Sigma = \Phi \Lambda \Phi^T \qquad (2\text{-}42)$$

di mana $\Phi^T = \Phi^{-1}$ dan Λ adalah matriks diagonal yang elemen-elemennya adalah nilai-nilai *eigen* dari Σ. Φ dengan kolom-kolomnya suatu korespondensi (ortonormal) vektor-vektor *eigen* dari Σ

$$\Sigma = (\vec{\upsilon}_1, \vec{\upsilon}_2, \ldots, \vec{\upsilon}_l).$$

Kombinasi persamaan (2-41) dan (2-42) diperoleh :

$$(\vec{x} - \vec{\mu}_i)^T \, \Phi\Lambda^{-1}\Phi^T \, (\vec{x} - \vec{\mu}_i) = c^2. \qquad (2\text{-}44)$$

Mendefinisikan $\vec{x}' = \Phi^T \vec{x}$. Koordinat $\vec{x}'$ adalah setara dengan $\vec{\upsilon}_k^T \vec{x}$, $k = 1, 2, \ldots, l$, yaitu proyeksi $\vec{x}$ pada vektor-vektor *eigen*. Dengan kata lain, semua adalah koordinat $\vec{x}$ terhadap sistem koordinat baru dengan sumbu-sumbu yang ditentukan oleh $\vec{\upsilon}_k$, k = 1, 2, ... , *l*. Persamaan (2-44) sekarang dapat dituliskan sebagai

$$\frac{(x_1' - \upsilon_{i1}')^2}{\lambda_1} + \frac{(x_2' - \upsilon_{i2}')^2}{\lambda_2} + \ldots + \frac{(x_l' - \upsilon_{il}')^2}{\lambda_l} = c^2.$$

Persamaan tersebut adalah *hyperellipsoid* dalam sistem koordinat baru. Gambar berikut menunjukkan kasus $l = 2$. Pusat massa *ellipse* pada $\vec{\mu}_i$, dan sumbu utamanya disekutukan / disejajarjan dengan vektor-vektor *eigen* yang cocok dan masing-masing memiliki panjang $2\sqrt{\lambda_k}\, c$. Sehingga, semua titik yang memiliki jarak *Mahalanobis* yang sama dari satu titik spesifik berada pada sebuah *ellipse*.

2.5 ESTIMASI FUNGSI KERAPATAN PROBABILITAS YANG TAK DIKETAHUI

Pembahasan sejauh ini berasumsi bahwa fungsi kerapatan probabilitas telah diketahui. Hal ini bukanlah peristiwa yang biasa terjadi. Dalam banyak persoalan yang mendasari pdf telah dapat diestimasi dari data yang tersedia. Ada banyak cara pendekatan terhadap suatu persoalan. Sering telah diketahui jenis pdf (seperti Gaussian, Rayleigh) tetapi tidak diketahui parameternya, seperti nilai rerata dan varian. Sebaliknya, dalam kasus yang berbeda tidak memiliki informasi tentang jenis pdf tetapi diketahui parameter statistiknya, seperti nilai rerata dan varian. Pendekatan yang berbeda dapat dipilih tergantung dari informasi yang tersedia.

2.5.1 Estimasi Parameter-kemungkinan Maksimum

Dibahas persoalan *M*-kelas dengan vektor-vektor ciri terdistribusi sesuai dengan $p(\vec{x}|\omega_i)$, dengan $i = 1, 2, 3, \ldots, M$. Diasumsikan bahwa fungsi kemungkinan itu diberikan dalam bentuk parametrik dan bahwa bentuk parameter yang bersesuaian dengan vektor $\boldsymbol{\theta}_i$ yang tidak diketahui. Untuk menunjukkan

ketergantungan pada $\boldsymbol{\theta}_i$ dituliskan $p(\vec{x}|\omega_i;\vec{\theta}_i)$. Tujuannya adalah untuk mengestimasi parameter yang tidak diketahui menggunakan sekumpulan vektor ciri yang diketahui pada setiap kelas. Jika diasumsikan lebih jauh bahwa dari satu kelas tidak mempengaruhi estimasi parameter yang lain, maka dapat dirumuskan persoalan yang bebas terhadap kelas-kelas dan menyederhanakan notasi.

Misalnya $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$ adalah sampel acak yang digambar dari pdf $p(\vec{x};\vec{\theta})$. Kemudian membentuk *joint* pdf $p(\vec{X};\vec{\theta})$, di mana $\vec{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ adalah himpunan dari sampel tersebut. Asumsi kebebasan secara statistik antara sampel-sampel yang berbeda dapat dituliskan

$$p(\vec{X};\vec{\theta}) \equiv p(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N; \vec{\theta}) = \prod_{k=1}^{N} p(\vec{x}_k;\vec{\theta}).$$

Ini merupakan fungsi $\vec{\theta}$ dan dikenal sebagai fungsi kemungkinan dari $\vec{\theta}$ terhadap $\vec{X}$. Metode kemungkinan-maksimum (*ML : maximum likelihood*) mengestimasi $\vec{\theta}$ sehingga fungsi kemungkinan tersebut mengambil nilai maksimumnya, yaitu

$$\hat{\theta}_{ML} = arg \max \prod_{k=1}^{N} p(\vec{x}_k;\vec{\theta}).$$

Satu kondisi yang diperlukan bahwa $\hat{\theta}_{ML}$ harus memenuhi agar menjadi maksimum yaitu bahwa turunan pertama (gradien) terhadap $\vec{\theta}$ dari fungsi kebolehjadian adalah nol, yakni

$$\frac{\partial}{\partial\vec{\theta}}\left[\prod_{k=1}^{N} p(\vec{x}_k;\vec{\theta})\right] = 0 \qquad (2\text{-}49)$$

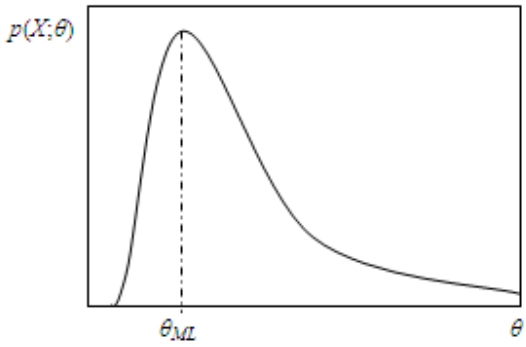
Karena monotonisitas dari fungsi logaritmik, maka didefinisikan fungsi *loglikelihood* sebagai

$$\boldsymbol{L}(\vec{\theta}) \equiv \ln \prod_{k=1}^{N} p(\vec{x}_k;\vec{\theta})$$

dan persamaan (2-49) ekivalen dengan

$$\frac{\partial}{\partial\vec{\theta}}[\boldsymbol{L}(\vec{\theta})] = \sum_{k=1}^{N} \frac{\partial}{\partial\vec{\theta}}[\ln p(\vec{x};\vec{\theta})] = \sum_{k=1}^{N} \frac{1}{p(\vec{x}_k;\vec{\theta})} \frac{\partial}{\partial\vec{\theta}}[p(\vec{x};\vec{\theta})] = 0.$$

Metode untuk kasus parameter tunggal yang tidak diketahui ditunjukkan pada gambar berikut. Estimasi *ML* bersesuaian dengan puncak dari (log) fungsi kebolehjadian.



Gambar 2.6 *Estimator kemungkinan maksimum*

Estimasi kebolehjadian maksimum memiliki banyak sifat yang dikehendaki. Jika $\vec{\theta}_0$ adalah nilai benar dari parameter yang tidak diketahui dari $p(\vec{x};\vec{\theta})$, maka dapat ditunjukkan bahwa di bawah kondisi valid secara umum yang berikut ini adalah benar :

Estimasi *ML* adalah tidak bias secara asimptotik, yang berdasarkan definisi berarti bahwa

$$\lim_{N\to\infty} E[\hat{\theta}_{ML}] = \vec{\theta}_0$$

Secara alternatif dikatakan bahwa estimasi konvergen pada rerata menuju nilai yang benar. Estimasi $\hat{\theta}_{ML}$ dengan sendirinya vektor acak, sebab untuk himpunan sampel yang berbeda $\boldsymbol{X}$ menghasilkan estimasi yang berbeda. Estimasi dikatakan tidak bias jika reratanya merupakan nilai yang benar dari parameter yang tidak diketahui. Dalam kasus *ML* ini benar hanya secara asimptotik ($N\to\infty$).

Estimasi *ML* adalah konsisten secara asimptotik, yaitu memenuhi

$$\lim_{N\to\infty} \text{prob}\,\{\| \hat{\theta}_{ML} - \vec{\theta}_0 \| \le \epsilon\} = 1$$

di mana ϵ adalah konstanta sembarang yang kecil. Secara alternatif dikatakan bahwa estimasi konvergen menuju probabilitas. Dengan kata lain untuk *N* yang besar berkemungkinan tinggi bahwa hasil estimasi akan berubah-ubah mendekati nilai yang benar. Kondisi kuat untuk konsistensi yang juga benar adalah

$$\lim_{N\to\infty} E[\| \hat{\theta}_{ML} - \vec{\theta}_0 \|^2] = 0$$

Dalam kasus demikian dikatakan bahwa estimasi konvergen menuju akar rerata. Dengan kata lain, untuk *N* besar, variansi estimasi *ML* cenderung menuju nol. Konsistensi sangat penting untuk sebuah estimator, sebab tidak terbias tetapi hasil estimasi menunjukkan variansi yang besar di sekitar rerata. Dalam kasus demikian mempunyai kepercayaan yang kecil terhadap hasil dari himpunen ***X*** tunggal.

Estimasi *ML* adalah efisien secara asimptotik, yaitu mencapai ikatan lebih rendah Cramer-Rao. Ini merupakan nilai variansi terendah yang dapat dicapai oleh estimasi sembarang.

Suatu pdf dari estimasi *ML* ketika $N\to\infty$ mendekati distribusi Gaussian dengan rerata $\vec{\theta}_0$. Sifat ini merupakan suatuketurunan dari (a) teorema limit pusat, dan (b) kenyataan bahwa estimasi *ML* terkait jumlah variabel acak, yaitu $\partial \ln(p(\vec{x}_k;\vec{\theta}))/\partial\,\vec{\theta}$.

Sebagai kesimpulan, estimator *ML* adalah tidak bias, terdistribusi secara normal, dan mempunyai kemungkinan variansi minimum. Tetapi semua sifat baik nin valid hanya untuk nilai *N* yang besar.

2.5.2 Maksimum Estimasi Probabilitas Posteriori

Untuk turunan estimasi kemungkinan maksimum, ditentukan θ sebagai parameter yang tidak dikenal. Pada bagian ini akan dihitung sebagai vektor acak, dan akan di perkirakan nilainya dalam kondisi sampel-sampelnya yaitu $x_1,...,x_N$. Tetapkan $X=\{x_1,...,x_N\}$ dengan titik awalnyaa adalah p(θ|X). Dari teorema Bayes diperoleh

$$p(\theta)p(X|\theta) = p(X)p(\theta|X) \tag{2.58}$$

atau

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)}$$

Maximum A Posteriori Probability (MAP) memperkirakan $\hat{\theta}_{\text{MAP}}$ didefinisikan pada titik dimana p(θ|X) dalam kondisi maksimum.

$$\hat{\theta}_{MAP}: \frac{\partial}{\partial\theta} P(\theta|x) = 0 \text{ atau } \frac{\partial}{\partial\theta}\big(p(\theta)p(X|\theta)\big) = 0 \tag{2.60}$$

Perbedaan diantara estimasi ML dan MAP terlihat dalam pergerakan p(θ) pada kasus sebelumnya. Apabila diasumsikan bahwa hal tersebut merupakan distribusi seragam, maka sifatnya konstan untuk semua θ, dari seluruh estimasi memberikan hasil yang identik. Hal ini juga menunjukkan bahwa p(θ) berada pada nilai variasi yang kecil. Gambar 2.7a dan 2.7b menggambarkan dua kasus tersebut.

Contoh 2.4 dari contoh sebelumnya 2.3 vektor mean μ diketahui untuk didistibusi normalkan sebagai berikut.

$$p(\mu) = \frac{1}{(2\pi)^{1/2}\sigma_\mu^l} exp\left(-\frac{1}{2}\frac{||\mu-\mu_0||^2}{\sigma_\mu^2}\right)$$

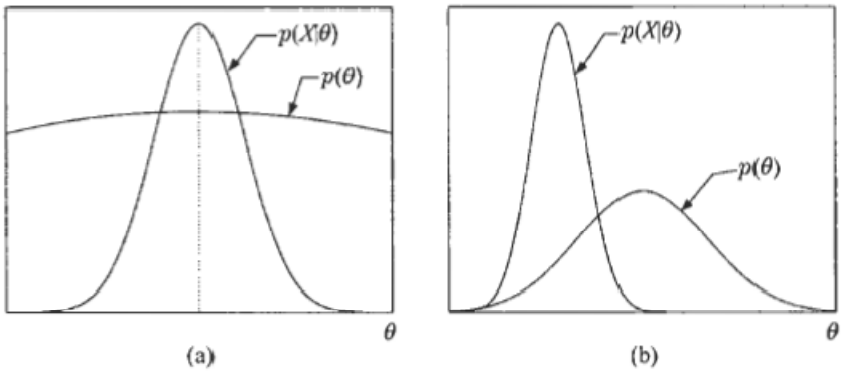
Estimasi MAP diberikan melalui persamaan

$$\frac{\partial}{\partial \mu} ln\left(\prod_{k=1}^{N} p(x_k|\mu)p(\mu)\right) = 0$$

Atau untuk $\sum=\sigma^2 I$

$\sum_{k=1}^{N} \frac{1}{\sigma^2}(x_k - \hat{\mu}) - \frac{1}{\sigma_\mu^2}(\hat{\mu} - \mu_0) = 0$

$$\hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_\mu^2}{\sigma^2}\sum_{k=1}^{N} x_k}{1 + \frac{\sigma_\mu^2}{\sigma^2}N}$$



Gambar 2.7 *Perkiraan ML dan MAPyang diaproksimasikan sama (a) dan berbeda (b)*

Dapat diobservasi bahwa $\frac{\sigma_\mu^2}{\sigma^2}$>>1, artinya varian sangat besar dan koresponden Gaussian sangat lebar dengan variasi yang kecil pada kisaran yang terjadi, maka

$$\hat{\mu}_{MAP} \approx \hat{\mu}_{ML} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

2.5.3 Kesimpulan Bayesian (Bayesian inference)

Seluruh metoda yang diperhitungkan pada sub bagian sebelumnya digunakan untuk menghitung estimasi spesifik terhadap paramater vektor θ yang belum diketahui. Pada metode sekarang ini diambil jalur yang berbeda. Diberikan X pada N vektor pelatihan dan informasi utama mengenai pdf p(θ), tujuannya adalah untuk menghitung kondisi pdf p(x|X). Persamaan berikut ini adalah persamaan penyelesaian dari hubungan-hubungan diatas.

$$p(x|X) = \int p(x|\theta)p(\theta|X)d\theta \qquad (2.61)$$

Dengan

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(Xx|\theta)p(\theta)d\theta} \quad (2.62)$$

$$p(X|\theta) = \prod_{k=1}^{N} p(x_k|\theta) \quad (2.63)$$

Persamaan 2.63 perhitungan statistik independen terhadap sampel-sampel pelatihan.

Keterangan

- Jika p(θ|X) dalam perhitungan 2.62 memuncak tajam pada $\hat{\theta}$ yang merupakan fungsi delta, persamaan 2.61 menjadi $p(x|X) \approx p(x|\hat{\theta})$; yaitu estimasi parameter yang diperkirakan sesuai dengan estimasi MAP. Sebagai contoh, jika p(X|θ) diartikan seputar puncak tajam dan p(θ) cukup luas di sekitar puncak ini. Kemudian estimasi hasil kurang lebih seperti ML.
- Pengertian lebih jauh mengenai metode ini dapat ditingkatkan dengan memfokuskan pada contoh berikut ini. Tentukan p(x|μ) menjadi variasi Gaussian N(μ, σ_0^2) dengan parameter yang tidak diketahui, yang juga diasumsikan mengikuti Gaussian N(μ, σ_0^2), hal ini merupakan aljabar sederhana (Problem 2.22), diberikan sejumlah sampel N, p(μ|X) dialihkan benjadi Gaussian dengan mean

$$\mu_N = \frac{N\sigma_\mu^2\bar{x}+\sigma^2\mu_0}{N\sigma_0^2+\sigma^2} \quad (2.64)$$

Dan varian

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2} \quad (2.65)$$

Dimana$\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$, N bervariasi dari 1 sampai ∞, dihasilkan runtutan Gaussian $N(\mu_N, \sigma_N^2)$ dimana nilai-nilai digeser dari μ_0 dan dijaga tetap pada batasnya sampai mean sampel $\bar{x}$, dan varian-varian tersebut menurun pada σ^2/N dengan N besar. Karena itu untuk nilai-nilai yang besar dari N, p(μ|X) menjadi memuncak tajam sekitar $\bar{x}$.

2.5.4 Estimasi Entropi Maksimum

Konsep dari entropi dikenal dari teori informasi *Shannon*, yang merupakan ukuran dari sifat acak dari pesan yang menjadi output dari sistem. Jika p(x) adalah fungsi densiti, entropi gabungan H diberikan oleh

$$H = -\int_x p(x)\ln p(x)\,dx \quad (2.66)$$

Asumsikan bahwa p(x) tidak diketahui, tetapi angka terkait lainnya diketahui (nilai mean, varian, dll). Perkiraan entropi maksimum dari pdf yang tidak diketahui adalah entropi yang di maksimisasi, berdasarkan penekanan yang diberikan. Berdasarkan pada prinsip entropi maksimum, yang ditetapkan oleh Jaynes [Jayn 82], merupakan hubungan estimasi terhadap distribusi yang menunjukkan kemungkinan random yang paling tinggi, berdasarkan pada penekanan yang tersedia.

Contoh 2.5 variabel random x adalah nonzero untuk x1<x<x2 dan yang lainnya adalah nol. Hitung estimasi entropi maksimum dari pdf nya

Dari persamaan 2.66

$$\int_{x_1}^{x_2} p(x)dx = 1 \quad (2.67)$$

Dengan menggunakan Lagrange multipliers, ekuivalen untuk memaksimalkan

$$H_L = -\int_{x_1}^{x_2} p(x)\,(\ln p(x) - \lambda)\,dx \tag{2.68}$$

Diambil turunan dengan memperhatikan p(x), didapatkan

$$\frac{\partial H_L}{\partial p(x)} = -\int_{x_1}^{x_2}\{(\ln p(x) - \lambda) + 1\}dx \tag{2.69}$$

Dengan menjadikan nol, didapatkan

$$\hat{\mathrm{p}}(\mathrm{x}) = \exp(\lambda - 1) \tag{2.70}$$

Untuk menghitung λ, substitusikan persamaan tersebut pada persamaan (2.67) dan didapatkan $\exp(\lambda - 1) = \frac{1}{x_2 - x_1}$, maka

$$\hat{\mathrm{p}}(\mathrm{x}) = \begin{cases} \frac{1}{x_2 - x_1} & jika\ x_1 \le x \le x_2 \\ 0 & untuk\ yang\ lainnya \end{cases} \tag{2.71}$$

Estimasi entropi maksimum dari pdf yang tidak dikenal merupakan distribusi seragam. Hasil estimasi adalah sesuatu yang memaksimalkan kerandoman dan semua poin dapat dimungkinkan. Berarti bahwa nilai mean dan varian diberikan sebagai penekanan yang kedua dan ketiga. Hasil Estimasi entropi maksimum dari pdf adalah berkisar -∞<x<+∞, merupakan Gaussian (Problem 2.25).

2.5.5 Model-model Campuran

Cara alternatif untuk memodelkan p(X) yang tidak diketahui adalah dengan kombinasi linear dari fungsi densiti dalam bentuk

$$p(x) = \sum_{j=1}^{J} p(x|j)Pj \tag{2.72}$$

Dimana

$$\sum_{j=1}^{J} P_j, \int_x\ p(x|j)dx = 1 \tag{2.73}$$

Dengan kata lain dapat diasumsikan bahwa distrbusi *J* berperan pada pembentukan p(x). Maka model ini mengasumsikan secara implisit bahwa poin x dapat digambarkan dari beberapa distribusi model distribusi *J* dengan probabilitas P_j,j=1,2....,J. Langkah pertama adalah menentukan set dari komponen densiti p(x|j) dalam bentuk parametrik, yaitu p(x|j:θ), kemudian komputasi dari parameter yang tidak diketahui, θ dan P_j,j=1,2,....,J, berdasarkan set dari sampel pelatihan yang tersedia x_k. Maksimum tipikal menyerupai Formulasi, memaksimalkan fungsi $\prod_k p(x_k; \theta, P_1, P_2, \dots, P_j)$ dengan memperhaitkan pada θ dan P_j. Kesulitan yang ditemukan disini terlihat dari keadaan bahwa parameter yang tidak diketahui, memasuki bagian maksimisasi dalam model nonlinear.

Algoritma Expectation Maximization (EM)

Algoritma ini idealnya cocok untuk kasus-kasus apabila data set tersedia secara komplit. Denotasikan suatu y sampel data komplit, dengan$y \in Y \subseteq R^m$, dan hubungkan pdf menjadi $p_y(y;\theta)$, dimana θ adalah

vektor parameter yang tidak diketahui. Sampel y tidak dapat diobservasi secara langsung. Yang diobservasi adalah sampel-sampel. $x = g(y) \in X_{0b} \subseteq R^l, l < m$. Diperhatikan bahwa keterkaitan pdf $p_x(x; \theta)$. Disebut sebagai *many-to-one-mapping*. Tentukan $Y(x) \subseteq Y$ sebagai subset dari korenponding y terhadap x yang spesifik. Kemudian pdf dari data yang belum komplit ditentukan dengan

$$p_x(x;\theta) = \int_{Y(x)} p_y(y;\theta)dy \quad (2.74)$$

Estimasi maksimum dari 0 ditentukan dari

$$\hat{\theta}_{ML}: \sum_k \frac{\partial \ln(p_y(y_k;\theta))}{\partial \theta} = 0 \quad (2.75)$$

Keadaan y tidak tersedia. Maka algoritma EM memaksimalkan ekspektasi dari fungsi log, yang dikondisikan pada sampel yang diobservasi dan estimasi iterasi dari 0. Dua langkah algoritma tersebut adalah:

E-step: pada langkah (t+1) dari iterasi, dengan $\theta(t)$ tersedia, hitung nilai yang diharapkan dari

$$Q(\theta;\theta(t)) \equiv E\left[\sum_k \ln(p_y(y_k;\theta|X;\theta(t))\right] \quad (2.76)$$

Langkah ini disebut juga dengan langkah yang diharapkan dari algoritma.

M-step: Hitung estimasi (t+1) berikutnya dari θ dengan memaksimalkan $Q(\theta; \theta(t))$, yaitu

$$\theta(t+1): \frac{\partial Q(\theta:\theta(t))}{\partial \theta} = 0 \quad (2.77)$$

Disebut juga langkah maksimum, yang dapat didiferensialkan dengan jelas.

Aplikasi permasalahan pemodelan campuran

Pada kasus ini set data komplit terdiri dari hubungan $(x_k,j_k),k=1,2,....,N$ dan j_k mengambil nilai-nilai integer dalam interval [1,j] dan di denotes campuran dari komponen yang ditimbulkan dari x_k.

$$p(x_k, j_k;\theta) = p(x_k|j_k;\theta)P_{jk} \quad (2.78)$$

Dengan mengasumsikan hubungan terpisah dari sampel-sampel set data, bentuk fungsi menjadi

$$L(\theta) \equiv \sum_{k=1}^{N} \ln(p(x_k|j_k;\theta)P_{jk}) \quad (2.79)$$

Tentukan $P=[P_1,P_2,...,P_j]^T$. vektor parameter adalah $\Theta^T=[\Theta^T,P^T]^T$.dengan mengambil data yang unobserved kondisi pada sampel-sampel pelatihan dan estimasi yang ada, $\Theta(t)$, dari paraeter yang tak diketahui didapat

E-step:

$$\begin{aligned} Q(\Theta;\Theta(t)) &= E\left[\sum_{k=1}^{N} \ln(p(x_k|j_k;\theta)P_{jk})\right] \\ &= \sum_{k=1}^{N} E[\ln(p(x_k|j_k;\theta)P_{jk})] \end{aligned} \quad (2.80)$$

$$\sum_{k=1}^{N}\sum_{j_k=1}^{j} P(j_k|x_k;\Theta(t)) \ln(p(x_k|j_k;\theta)P_{jk}) \quad (2.81)$$

Notasi tersebut dapat disederhanakan dengan menurunkan indeks k dari j_k. Karena untuk setiap k, dapat dijumlahkan seluruh kemungkinan nilai J dari j_k dan juga sama halnya untuk semua k. Algoritma untuk kasus campuran Gaussian dengan matriks kovarian diagonal dari bentuk $\sum_j = \sigma_j^2 I$ yaitu

$$(p(x_k|j_k;\theta) = \frac{1}{2\pi\sigma_j^{2\,l/2}} exp\left(-\frac{||x_k-\mu_j||^2}{2\pi\sigma_j^2}\right) \quad (2.82)$$

Asumsikan bahwa disamping probabilitas atas P_j, nilai mean respektif μj dengan varian $\sigma_j^2, j = 1,2,\dots,J$, yang diketahui dari Gaussian. Maka θ adalah vektor dimensi J(l+1). Dengan mengkombinasikan persamaan (2.81) dan (2.82), dan menghilangkan konstanta didapat persamaan:

E-step:

$$Q(\Theta;\Theta(t)) = \sum_{k=1}^{N}\sum_{j_k=1}^{j} \mathrm{P}(j_k|x_k;\Theta(t))\left(-\frac{1}{2}\ln\sigma_j^2 - \frac{1}{2\sigma_j^2}||x_k-\mu_j||^2 + \ln P_j\right) \quad (2.83)$$

M-step

$$\mu_j(t+1) = \frac{\sum_{k=1}^{N}\mathrm{P}(j_k|x_k;\Theta(t))x_k}{\sum_{k=1}^{N}\mathrm{P}(j_k|x_k;\Theta(t))} \quad (2.84)$$

$$\sigma_j^2(t+1) = \frac{\sum_{k=1}^{N}\mathrm{P}(j_k|x_k;\Theta(t))||x_k-\mu_{j(t+1)}||^2}{l\sum_{k=1}^{N}\mathrm{P}(j_k|x_k;\Theta(t))} \quad (2.85)$$

$$P_j(t+1) = \frac{1}{N}\sum_{j_k=1}^{j}\mathrm{P}(j_k|x_k;\Theta(t)) \quad (2.86)$$

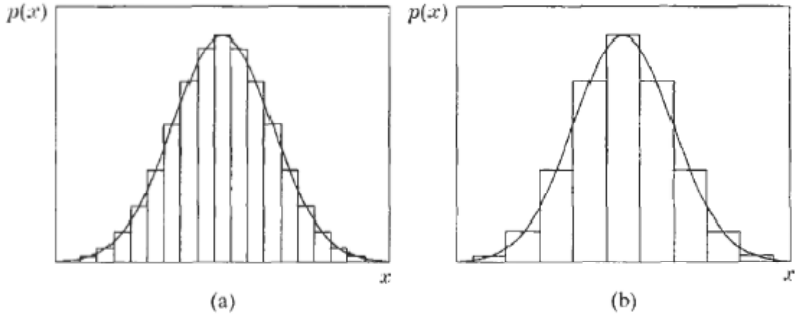
Untuk melengkapi iterasi, dibutuhkan perhitungan$\mathrm{P}(j_k|x_k;\Theta(t))$.

$$\mathrm{P}(j_k|x_k;\Theta(t)) = \frac{\mathrm{p}(x_k|j;\Theta(t))P_j(t)}{\mathrm{p}(x_k;\Theta(t))} \quad (2.87)$$

$$\mathrm{P}(j_k|x_k;\Theta(t)) = \sum_{j_k=1}^{j}\mathrm{p}(x_k|j;\Theta(t))P_j(t) \quad (2.88)$$

2.5.6 Estimasi Nonparametrik

Sebuah contoh dari kasus dari dimensi sederhana, gambar 2.8 menunjukkan dua contoh pdf dan aproksimasinya dengan metode histogram. X-aksis (ruang dimensi satu) dibagi menjadi beberapa bagian dengan lebar h. Kemudian probabilitas dari sampel x akan ditempatkan dalam wadah untuk diestimasikan.



Gambar 2.8 *Kemungkinan fungsi densitidengan metode histogram (a) interval kecil dan (b) interval besar*

Jika N adalah jumlah total sampel dan kemungkinan k_N yang diaproksimasikan dengan rasio frekuensi

$$P \approx k_N/N \tag{2.89}$$

Nilai koresponding pdf diasumsikan konstan di seluruh wadah dan diaproksimasi oleh

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{1}{h}\frac{k_N}{N}, |x - \hat{x}| \leq \frac{h}{2} \tag{2.90}$$

Dimana x adalah titik tengah dari wadah yang menunjukkan amplitude dari kurva histogram seluruh wadah. Hal itu adalah aproksimasi yang mungkin untuk p(x) yang kontinyu dan h yang cukup kecil, maka asumsi dari p(x) konstan dalam wadah dapat berubah. Dapat ditunjukkan bahwa p(x) konvergen terhadap nilai nyata p(x) pada N->

- $h_N \to 0$
- $k_N \to \infty$
- $\frac{k_N}{N} \to 0$

Dengan h_N digunakan untuk menunjukkan dependen dari N. Pada setiap titik dimana $p(x) \neq 0$, dapat merubah ukuran dari h_N, sekecil apapun, probabilitas P dari titik-titik yang dihitung dalam wadah ini dapat terbatase.dan $k_N \approx P_N$ dan k_N mendekati takterbatas seperti halnya perkembangan N menjadi takterbatas. Pada prakteknya, jumlah N data poin adalah terbatas. Kondisi sebelmnya menunjukkan cara bahwa parameter yang bevariasi harus dipilih. N harus cukup besar, h_N cukup kecil, dan jumlah setiap titik yang jatuh pada wadah harus cukup besar pula. Seberapa kecil dan seberapa besar bergantung pada jenis fungsi pdf dan derajat aproksimasi yang menguntungkan.

Jendela Parzen. Dalam kasus multidimensional, mengenai wadah dari ukuran h, ruang dimensi-l dibagi menjadi banyak kubus dengan lebar sisi h dan volume h^l. Tetapkan x_i, i = 1,2,...,N menjadi vektor ciri yang tersedia. Definisikan fungsi (x) maka

$$\emptyset(x_i) = \begin{cases} 1 & untuk|x_{ij}| \leq {}^1/_2 \\ 0 & untuk\ yang\ lainnya \end{cases} \tag{2.91}$$

Dimana x_{ij},j = 1,....,l merupakan komponen dari x_i. Dengan arti, fungsi tersebut sama dengan 1 untuk semua titik didalam kubus sisi yang dipusatkan pada origin dan 0, diluarnya. Persamaan 2.90 dapat ditulis menjadi

$$\hat{p}(x) \equiv \frac{1}{h}\left(\frac{1}{N}\sum_{i=1}^{N} \emptyset\left(\frac{x_i - x}{h}\right)\right) \quad (2.92)$$

Kemudian Parzen menggeneralisir persamaan 2.92 dengan menggunakan fungsi yang halus dalam $\emptyset(.)$, yang dapat ditunjukkan sebagai berikut

$$\emptyset(x) \geq 0 \text{ dan} \quad (2.93)$$

$$\int_x \emptyset(x)dx = 1 \quad (2.94)$$

Kemudian diambil nilai mean dari persamaan 2.92

$$E[\hat{p}(x)] \equiv \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N} E\left[\emptyset\left(\frac{x_i - x}{h}\right)\right]\right)$$

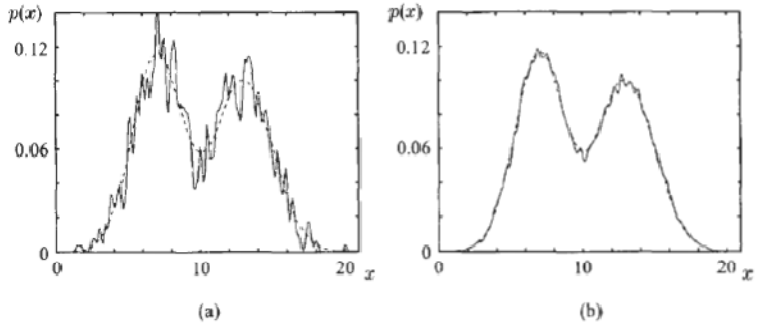
$$\equiv \int_{x`} \frac{1}{h^l}\emptyset\left(\frac{x` - x}{h}\right)p(x`)dx` \quad (2.95)$$

Nilai mean adalah vesi yang dihaluskan dari pdf nyata p(x). Maka dari itu sebagaimana

$h \to 0$ fungsi $\frac{1}{h^l}\emptyset\left(\frac{x` - x}{h}\right)$ berdasar pada fungsi delta $\delta(x_i - x)$ Amplitude berubah menjadi takterbatas, lebarnya mendekati 0 dan integral dari persamaan 2.94 kembali sama dengan 1.

Keterangan

- Untuk N yang tetap, h yang kebih kecil dan varian yang lebih tinggi, diindikasikan oleh keadaan berderau dari estimasi hasil pdf, contohnya adalah pada gambar 2.9a dan 2.10a. Hal ini dikarenakan p(x) diaproksimasikan oleh jumlah terbatas dari fungsi δ, yang dipusatkan pada titik sampel pelatihan.



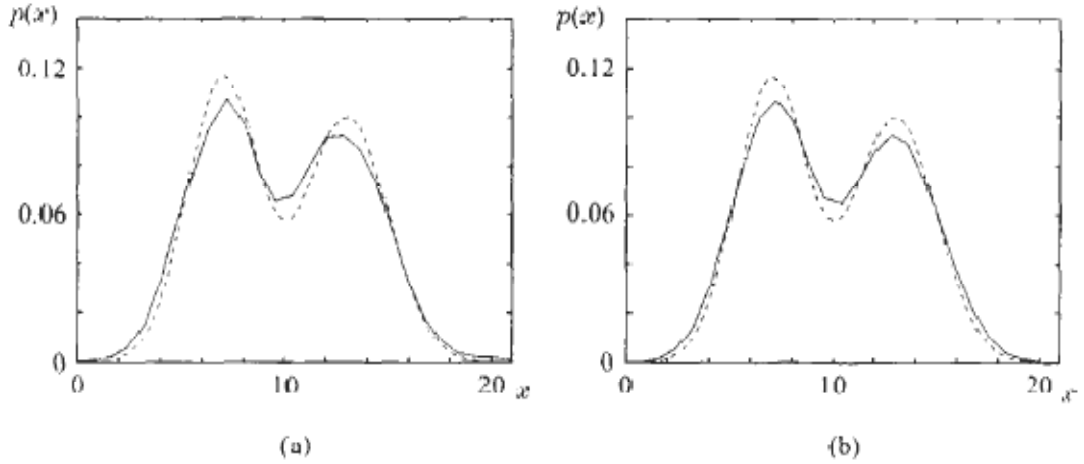
Gambar 2.9 *Perkiraan pdf dengan jendela Parzen (a) h=0.1 dan 1000 sampel pelatihan (b) h=0.1 dengan 20000 sampel*

Maka jika sesuatu menggerakan x dalam ruang respon dari p(x) akan sangat tinggi mendekati titik pelatihan dan akan menurunkan sangat rapid seperti saat dipindahkan, mengikuti penampilan derau.

- Untuk h yang tetap, varian diturunkan sebagai angka N poin sampel. Hal ini dikarenakan ruangan menjadi padat dalam titik, dan fungsi yang berlawanan ditempatkan, seperti pada Gambar 2.9(b), lebih lanjut lagi untuk jumlah sampek yang banyak, h yang lebih kecil akurasi yang lebih baik dari hasil estimasi, sebagai contoh Gambar 2.9(b) dan 2.10(b).
- Dapat ditunjukkan sebagai contoh [Parz 62, Fuku 90] bahwa dibawah bebeapa kondisi pada ∅(.), yang valid untuk kebanyakan fungsi densiti, jika h mendekati 0hasil estimaasi adalah semuanya unbias dan konsisten asimtotik.

Keterangan

- Dalam prakteknya, dimana hanya ada sejumlah terbatas dari sampel yang mungkin, perjanjian antara h dan N harus ditentukan. Pilihan dari nilai yang sesuai untuk h adalah krusial dan beberapa peningkatan sudah pernah diajukan dalam literatur [Wand 95].



Gambar 2.10 *Perkiraan pdf dengan jendela Parzen (a) h=0.8 dan 1000 sampel pelatihan (b) h=0.8 dengan 20000 sampel*

- Biasanya N yang besar dibutuhkan untuk performa yang dapat diterima.

Aplikasi untuk pengklasifikasian :

$$berikan\ x\ ke\ \omega_1(\omega_2)\ jika\ l_{12} \approx \left(\frac{\frac{1}{N_l h^l}\sum_{i=l}^{N_1}\emptyset\left(\frac{x_i-x}{h}\right)}{\frac{1}{N_l h^l}\sum_{i=l}^{N_2}\emptyset\left(\frac{x_i-x}{h}\right)}\right) > (<) \frac{P(\omega_2)\lambda_{21}-\lambda_{22}}{P(\omega_1)\lambda_{12}-\lambda_{11}} \qquad (2.96)$$

Dimana N_1, N_2 adalah vektor pelatihan pada kelas w_1, w_2, Untuk N_1, N_2 yang besar, komputasi ini membutuhkan persediaan waktu proses dan memori yang cukup.

k Estimasi Nearest Neighbor Density. Dalam estimasi Parzen dari pdf dalam 2.92). volume di sekitar poin x dibuat tetap (h`) dan jumlah poin k_N yang jatuh dalam volume, ditinggalkan untuk keadaan acak dari pon ke poin. Jumlah dari poin k_N = k akan diperbaiki dan ukuran volume di sekitar x akan diatur pada tiap waktu, untuk memasukkan poin k. jadi dalam area densiti rendah, volume akan menjadi besar dan daerah dengan densiti yang tiggi akan menjadi kecil. Estimator tersebut dapat ditulis sebagai

$$\hat{p}(x) = \frac{k}{NV(x)} \quad (2.97)$$

Secara pandangan praktis, dari vektor dengan ciri yang tidak diketahui dari x, dapat dihitung jarak d, sebagai contoh Euclidean, dari seluruh vektor pelatihan dengan kelas yang bervariasi, sebagai contoh w_1,w_2. Tetukan r_1 sebagai radius segi banyak, dipusatkan pada x, yang mengandung titik k dari w_1 dan r_2. Radius dari segibanyak yang mengandung titik k dari kelas w2 (k tidak akan dibutuhkan, sama untuk semua kelas). Jika ditunjukkan oleh V_1, V_2 volume segi banyak berturut-turut, kemungkinan tes perbandingan menjadi

$$berikan\ x\ ke\ \omega_1(\omega_2)\ jika\ l_{12} \approx \frac{kN_2V_2}{kN_1V_1} > (<) \frac{P(\omega_2)\lambda_{21}-\lambda_{22}}{P(\omega_1)\lambda_{12}-\lambda_{11}}$$

$$\frac{V_2}{V_1} > (<) \frac{P(\omega_2)\lambda_{21}-\lambda_{22}}{P(\omega_1)\lambda_{12}-\lambda_{11}} \quad (2.89)$$

2.6 ATURAN TETANGGA TERDEKAT

Variasi dari hasil teknik estimasi densiti kNN dalam suboptimal, telah ppular dalam prakteknya,pemisahan nonliear. Berikut adalah algoritma untuk aturan tetangga terdekat. Berikan vektor x dengan ciri yang tidak diketahui dan sebuah ukuran jarak, kemudian:

- Keluarkan vektor latihan N, identifikasi tetangga terdekat k, terlepas dari label kelas, dipilih k untuk diganjilkan pada maslah dua kelas, dan pada umumnya bukan perkalian dari jumlah kelas M
- Keluarkan sampel k tersebut, identifikasi beberapa vektor, ki, yang menjadi bagian kelas w_1, i = 1,2,...,M. Yaitu $\sum_i\ k_i = k$
- Berikan x ke kelas wi dengan jumlah maksimum ki sampel,

Beberapa ukuran jarak dapat digunakan, termasuk jarak Euclidean dan Mahalanobis. Dalam [Hast 96] metrik efektif dianjurkan untuk menguasai informasi lokal pada setiap titik. Versi paling sederhana dari algoritma tersebut adalah untuk k = 1, yang diketahui sebgai aturan Nearest Neighbor (NN). Dengan kata lain ciri vektor x diberikan pada kelas tetangga terdekatnya. Isediakan jumlah dari sampel platihan cukup besar, aturan sederhana ini menunjukkan performa yang bagus. P_{NN} di tentukan oleh

$$P_B \le P_{NN} \le P_B\left(2 - \frac{M}{M-1}P_B\right) \le 2P_B \quad (2.99)$$

Dimana P_B adalah kesalahan Bayesian yang optimal. Jadi kesalahan yang dihasilkan oleh pemisah NN adalah (seacara asimmtot) lebih sering dua kali dari pemisah yang optimaal. Performa asimtot dari kNN lebih baik daripada NN itu sendiri, dan jumlah ikatan yang menarik diperoleh. Sebagai contoh, untuk kasus dua kelas dapat ditunjukkan [Devr 96] bahwa

$$P_B \le P_{kNN} \le P_B + \frac{1}{\sqrt{ke}}\ atau\ P_B \le P_{kNN} \le P_B + \sqrt{\frac{2P_{NN}}{k}} \quad (2.100)$$

Untuk nilai kesalahan Bayyesian yang kecil, perkiraan berikut adalah valid [Devr 96]:

$$P_{NN} \approx 2\ P_B \quad (2.101)$$

$$P_{3NN} \approx P_B + 3(P_B)^2 \quad (2.102)$$

Jadi, untuk N besar dan kesalahan Bayesian yang kecil, diharapkan agar pemisah 3NN dapat memberikan performasi yang hampir identik terhadap pemisah Bayesian. Sebagai contoh, dikatakan bahwa kemungkinan kesalahan dari pemisah Bayesiam adalah pada kisaran 1%; kemudian kesalahan yang dihasilkan dari pemisah 3NN akan berkisar 1.03 %. Perkiraan meningkat pada nilai k yang lebiih tinggi. Berdasar dugaan dari N besar, kisaran dari segi banyak (jarak Euclidean) dipusatkan pada x dan mengandung tetangga terdekat dari k, dipertahankan tetap 0 [Devr 96]. Hal ini adalah alami, karena untuk N yang sangat besar dapat diharapkan ruang untuk diisi penuh oleh sampel. Maka k (bagian terkecil dari N) tetangga dari x akan diltempatkan sangat dekat terhadap x, dan probabilitas kelas kondisional pada setiap titik didalam segi banyak di sekitar x akan dipekirakan sama dengan P(wi|x)(mengasumsikan kontinuitas). Lebih jauh lagi untuk k yang besar (Potongan yang sangat kecil dari N) sebagian besar titik-ttik dari wilayah tesebut akan menjadi bagian dari korespinding kelas terhadap kemungkinan kondisi maksimum. Maka aturan kNN titetapkan terhadap pemisah Bayesian. Pada kasus sampel yang terbatas pada contoh Problem 2.29, dimana hasil dari kNN pada kemungkinan kesalahan ang lebih tinggi daripada NN. Sebagai konklusi, dapat ditetapkan bahwa teknik *nearest neighbor* adalah kandidat yang serius untuk diadaptasi sebagai pemisah dalam beberapa aplikasi.Studi komparatif tentang beberapa pemisah statistikal dipertimbangkan pada bagian pembahasan ini, dapat dilihat di [Aebe 94].

Keterangan

- Hal yang berkenaan dengan teknik (k)NN adalah kompleksitas dari tetangga terdekat diantara sampel pelatihan N yang tersedia.
- Pada k = 1 aturan tetangga tedekat digunakan, dan vektor-vektor ciri pada pelatihan xi, i = 1,2,....,N mendefinisikan pemisahan ruang dimensi l ke wilayah N, Ri, tiap wilayah ini didefinisikan oleh

$$R_i = \{x: d(x, x_i) < d(x, x_j), i \neq j\} \quad (2.103)$$

Yaitu, Ri meliputi seluruh titik dlaam ruang yang lebih dekat terhadap xi daripada titik yang lainnya pada set pelatihan, dengan memperhatikan jarak d. Pemisahan dari ruang ciri dikenal dengan *Voronoi teeselation* untuk kasus dari l = 2 dan jarak Euclidean.