Ekplorasi data

Oleh Mulaab

Atribut Data numerik

Dalam bab ini, kita membahas metode statistik dasar untuk analisis ekploarasi data atribut numerik. Kita membahas ukuran kecenderungan pusat (central tendency), ukuran dispersi atau sebaran, dan ukuran ketergantungan linier atau hubungan antara atribut. Kita menekankan hubungan antara probabilistik dan geometris dan aljabar dari sudut pandang data matriks

Analisa univariat

Analisis univariat dilakukan pada atribut tunggal (X); dengan demikian matriks data D bisa dianggap sebagai matriks $n \times 1$, atau sebagai vektor kolom, yang dianyatakan dengan

$$D = \begin{pmatrix} \frac{X}{x_1} \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

dimana X adalah atribut numerik yang dimaksudkan, dengan $x_i \in \mathbb{R}$. X diasumsikan adalah variabel random, dengan setiap titik $x_i (1 \leq i \leq n \, x_i (1 \leq i \leq n)$, merupakan variabel acak. Kita asumsikan bawa data pengamatan adalah. Kami berasumsi bahwa data yang diamati adalah sampel acak yang diambil dari X, artinya, setiap variabel x_i adalah saling bebas dan berdistribus sama (iid). Dalam sudut pandang vektor, kami memperlakukan sampel sebagai vektor n-dimensi, dan menulis $X \in \mathbb{R}^n$

Secara umum, fungsi padat probabilitas atau fungsi mass $f(x)$ dan fungsi distribusi kumulatif $F(x)$, untuk atribut X keduanya tidak diketahui. Akan tetapi, kita dapat mengestimasi distribusi ini langsung dar data sample, juga juga memungkinkan kita untuk menghitung beberapa parameter penting populasi.

Fungsi distribusi Kumulatif Empiris

Fungsi distribusi kumulatif empiris (CDF) dari X dinyatakan dengan

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

dimana

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

adalah variabel indikator biner yang menyatakan variabel indikator biner yang menunjukkan apakah kondisi yang diberikan terpenuhi atau tidak.

Fungsi distribusi kumulatif Invers

Definisi fungsi distribusi kumulatif invers atau fungsi quantile untuk variabel acak X sebagai berikut :

$$F^{-1}(q) = \min\{x|\hat{F}(x) \geq q\} \quad \text{for } q \in [0,1]$$

Fungsi distribusi kumulatif Invers empiris dapat diperoleh dari persamaan (2)

Fungsi massa Probabilitas Empiris

Fungsi massa probabilitas empiris dari X dinyatakan dengan

$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$$

dimana

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

Fungsi massa probabilitas empiris juga menempatkan massa probabitas $\frac{1}{n}$ pada setipa titik x_i

Mengukur kecenduran terpusat

Ukuran ini memberikan indikasi tentang konsentrasi massa probabilitas , nilai tengah dan lainnya.

Mean

Mean juga disebut dengan nilai harapan dari variabel acak X adalah rata rata aritmetika dari nilai X. Itu merupakan salah satu dari kecenderungan terpusat dari X.

Mean atau nilai harapan dari variabel acak X didefinisikan dengan

$$\mu = E[X] = \sum_{x} xf(x)$$

diman $f(x)$ adalah fungsi massa probabilitas dari X.

Nilai harapan dari variabel acak kontinu X dinyakan dengan

$$\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

dimana $f(x)$ adalah fungsi padat probabilitas dari X.

Sample Mean. Sample mean adalah statistik, yaitu fungsi $\hat{\mu} : \{x_1, x_2, \ldots, x_n\} \to \mathbb{R}$, didefinisikan sebagai nilai rata-rata dari x_i :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

nilai adalah sebagai pengestimasi nilai mean yang tidak diketahui dari X. Nilai tersebut diperoleh dengan memasukkan dalam fungsi massa probabilitas empiris dalam persamaan (7)

$$\hat{\mu} = \sum_{x} x\hat{f}(x) = \sum_{x} x(\frac{1}{n}\sum_{i=1}^{n} I(x_i = x)) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample mean adalah tidak bias . Estimator $\hat{\theta}$ disebut dengan unbiased estimatore (stimator tidak bias) untuk parameter θ jika $E[\hat{\theta}] = \theta$ untuk setiap kemungkinan nilai dari θ . Sample mean $\hat{\mu}$ adalah unbiased estimator untuk mean populasi μ sehingga

$$E[\hat{\mu}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

dimana kita gunakan fakta bahwa variabel acak x_i adalah IID sesuai dengan X, yang berarti bahwa mereka memiliki rata-rata μ yang sama seperti X , yaitu,$E[x_i] = \mu$ untuk semua x_i. Kita juga menggunakan fakta bahwa fungsi ekpektasi E adalah linier operator yaitu untuk suatu dua bilangan acak X dan Y dan bilangan real a dan b , kita memiliki $E[aX + bY] = aE[X] + bE[Y]$

Robustnes Kita mengatakan bahwa statistik adalah robust jika tidak dipengaruhi oleh suatu nilai ekstrim (misal outlier/pencilan) dalam data. Rata-rata sampel sayangnya tidak kuat karena ada satu nilai besar (outlier) dapat mejadikan rata-rata yang tidak sebenarnya. Ukuran yang lebih robust adalah trimmed mean yang didapatkan setalah mengabaikan sebagian kecil dari nilai nilai ekstrim pada salah satu ujungnya.

Median

Median dari suatu variabel acak didefinisikan dengan nilai m sehingga

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

Degan kata lain, median m adalah nilai paling tengan (middle-most). Dalam istliah (invers) cumulatif distribution function , median m dinyatakan dengan

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

Sample median dapat diperoleh dari Fungsi distribusi kumulatif invers atau fungsi distribusi kumulatif invers empiris dengan dihitung

$$\hat{F}(m) = 0.5 \text{ atau } m = \hat{F}^{-1}(0.5)$$

Pendekatan paling sederhan untuk menghitung sample median adalah pertama kai dari mengurutkan semua nilai x_i $(i \in [1, n])$ dengan urutan naik. Jika n adalah ganjil , media adalah nilai pada posisi $\frac{n+1}{2}$. Jika n adalah genap, nilai padan posisi $\frac{n}{2}$ dan $\frac{n}{2} + 1$ adalah keduanaya median. idak seperti mean, media adalah robust, sehingga ia tidak dipengaruhi oleh banyak nilai extrim. Juga nilai tersebut terjadi dalam sample dan nilai yang bisa diasumsikan oleh variabel acak.

Mode

Nilai *mode* dari variabel acak adalah nilai dimana fungsi massa probabilitas atau fungsi padat probabilitas mencapai nilai maximumnya, bergantung pada apakah X adalah diskrit atau kontinu.

Sample mode adalah nila untuk fungsi probabilitas empiris mencapai nilai maksimum, dinyatakan dengan

$$mode(X) = \arg\max_{x} \hat{f}(x)$$

Mode ini mungkin bukan ukuran kecenderungan sentral yang sangat berguna untuk sampel, karena kemungkinan elemen yang tidak representatif menjadi elemen yang paling sering muncul. Selanjutnya, jika semua nilai dalam sampel berbeda, maka masing-masing akan menjadi mode

Contoh. **(Sample Mean, Median, dan Mode)**. Perhatikan atribut sepal length (Xi) dalam data iris. Data iris, dimana nilainya seperti yang ditunjukkan dalam tebel 1.2 . Sample mean dinyatakan dengan

$$\hat{\mu} = \frac{1}{150}(5.9 + 6.9 + \cdots + 7.7 + 5.1) = \frac{876.5}{150} = 5.843$$

Gambar 2.1 menunjukkan semua dari 150 nilai sepal length dan sample mean. Gambar 2.2a menunjukkan fungsi distribusi kumulatif empiri dan gambar 2.2b menunjukkan fungsi distribusi kumulatif empiris untuk sepal length

Karena $n = 150$ adalah genap, sample median adalah nilai pada posisi $\frac{n}{2} = 75$ dan $\frac{n}{2} + 1 = 76$ setelah diurutkan. Untuk sepal length kedua nilainya adalah 5.8, kemudian sample media adalah 5.8 . Dari fungsi distribusi kumulatif invers dalam gambar 2.2b, kita dapat melihat bahwa

$$\hat{F}(5.8) = 0.5 \text{ or } 5.8 = \hat{F}^{-1}(0.5)$$

Sample mode untuk sepal length adalah 5. yang dapat dilihat dari frequency dari 5 dalam gambar 2.1. Massa probabilitas empiris pada $x = 5$ adalah

$$\hat{f}\,(5) = \frac{10}{150} = 0.067$$

Mengukur sebaran (dispersion)

Mengukur dispersi memberikan indikasi tentang sebaran atau variasi pada nilai nilai variabel acak.

Jangkauan

Jangkauan nilai atau secara sederhana **jangkauan** (range) variabel acak X adalah perbedaan antara nilai maximum dan nilai minimum dari X dinyatakan dengan

$$r = \max\{X\} - \min\{X\}$$

Sample range adalah statistik, dinyatakan dengan

$$\hat{r} = \underset{i=1}{\overset{n}{max}}\,\{x_i\} - \underset{i=1}{\overset{n}{min}}\,\{x_i\}$$

Dengan definisi, jangkauan adalah sensitif terhadap nilai extrime sehingga tidak robust.

Jangkauan antar interquartile

Quartile adalah nilai khusus dari fungsi quantile persaman (2.2) yang membagi data kedalam empat bagian. Furthermore quartile terkati dengan nilai-nilai quantile 0.25, 0.5, dan 0.74 dan 1.0. Quantile pertama adalah nilai $q_1 = F^{-1}(0.25)$ 25% dari sebelah kiri rentang titik, kuartile ke dua adalah sama dengan nilai median $q_2 = F^{-1}(0.5)$, 50 % dari sebelah kiri data dan $q_3 = F^{-1}(0.75)$ adalah nilai 75% dari sebelah kiri dan quantile ke empat adalah nilai maximum dari X, 100 % sebelah kiri dari rentang data.

Ukuran yang lebih robust dari seberan X adalah jangkauan interquartile (IQR) dinyatakan dengan

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25)$$

Variansi dan standar deviasi

Variansi dari variabel acak X memberikan pengukuran berapa banyak nilai nilai dari penyimpangan X dari rata-rata atau nilai harapan dari X. Lebih tepatnya variansi adalah nilai harapan dari penyimpangan dari mean yang dikuadratkan yang didefinisikan dengan

$$\sigma^2 = \operatorname{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{jika } X \text{ adalah diskrit} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{jika } X \text{ adalah kontinu} \end{cases}$$

Standar deviasi σ didefinisikan sebagai akar kuadrat positif dari variansi σ^2. Kita dapat juga menulis variansi sebagai selisih antara ekpektasi X^2 dan akar dari ekpektasi X :

Variansi Sampel

Variansi sampel didefinisikan dengan

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Standar deviasi adalah akar dari variansi sample yang dinyatakan dengan

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

Analisa Bivariate

Dalam analisa bivariate, kita memandang dua atribut pada waktu yang sama. Kita fokus untuk memahami keterkaitan atau kebergantunga antara dua variabel atau atribut tersebut, jika ada. Kita lalu membatasi pada dua variabel X_1 dan X_2 , dengan D dinyatakan sebagai matrik dengan ukuran $n \times 2$

$$D = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Secara geometri, kita dapat memandang D dalam dua cara. Itu dapat dianggap sebagai n titik atau vektor dalam 2-ruang dimensi terhadap atribut X_1 dan X_2 yaitu $x_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$ .Selain itu dapat dilihat sebagai 2 titik atau vektor dalam n-ruang dimensi yang berisi titik, yaitu setiap kolom adalah vektor dalam $\mathbb{R}^n$ sebagai berikut :

$$X_1 = (x_{11}, x_{21}, \ldots, x_{n1})^T$$
$$X_2 = (x_{12}, x_{22}, \ldots, x_{n2})^T$$

Dalam sudut pandang probabilistik, vektor kolom $X = (X_1, X_2)^T$ dianggapa variabel acak bivariate dan titik titik $x_i (1 \leq i \leq n)$ dinyatakan sebagai sampel acak yang diperoleh dari X, yaitu x_i dianggap independent and identically distributed (iid) seperti X.

Fungsi Massa Probabilitas Gabungan Empiris

Fungsi Massa Probabilitas Gabungan Empiris untuk X dinyatakan dengan

$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$$

$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n}\sum_{i=1}^{n} I(x_{i1} = x_1, x_{i2} = x_2)$$

dimana I adalah variabel indikator yang bernilai 1 jika argumen argumennya benar

$$I(x_i = x) = \begin{cases} 1 & \text{jika } x_{i1} = x_1 \text{ dan } x_{i2} = x_2 \\ 0 & \text{untuk yang lainnya} \end{cases}$$

Seperti dalam kasus univariate, fungsi probabilitas menempatkan massa probabilitas $\frac{1}{n}$ pada setiap objek dalam data sampel.

Mengukur Dispersi

Mean

Rata rata bivariate didefinisikan sebagai nilai harapan dari variabel acak vektor X, didefinisikan sebagai berikut :

$$\mu = E[X] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Dengan kata lain, rata-rata bivariate adalah nilai harapan dari masing masing atribut.

Rata-rata sampel dapat diperoleh dari $\hat{f}_{x_1}$ dan $\hat{f}_{x_2}$, fungsi massa probabilitas empiris dari X_1 dan X_2 , menggunakan persamaan (2.5). Dapat juga dihitung dari gabungan fungsi massa probabilitas empiris dalam persamaan (2.17)

$$\hat{\mu} = \sum_x x\hat{f}(x) = \sum_x x\left(\frac{1}{n}\sum_{i=1}^{n} I(x_i = x)\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variansi

Kita dapat menghitung variansi masing masing atribut, yaitu σ_1^2 untuk X_1 dan σ_2^2 untuk X_2 mengggunkan persamaan (2.8). Variansi secara keseluruhan (1.4) dinyatakan dengan

$$var(D) = \sigma_1^2 + \sigma_2^2$$

Variansi sampel $\hat{\sigma}_1^2 + \hat{\sigma}_2^2$ dapat diestimasi dengan menggunakanpersamaan (2.10) dan jumlah variansi sample adalah $\sigma_1^2 + \sigma_2^2$

Mengukur keterkaitan

Covarian

Kovarian antara dua atribut X_1 dan X_2 mengukur keterkaitan antara kebergantungan linier diantaranya dan didefinisikan dengan

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

Dengan linieraritas dari harapan, kita miliki

$$\begin{aligned} &\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] \\ &= E[X_1X_2] - \mu_2E[X_1] - \mu_1E[X_2] + \mu_1\mu_2 \\ &= E[X_1X_2] - \mu_1\mu_2 \\ &= E[X_1X_2] - E[X_1]E[X_2] \end{aligned}$$

Persamaan (2.21) dapat dianggap sebagai generalisasi dari variansi univariate persamaan (2.9) pada kasus bivariate.

Jika X_1 dan X_2 adalah variabel acak saling bebas, maka kita dapat simpulkan bahwa covariannya adalah nol. Ini karena jika X_1 dan X_2 adalah saling bebas, maka kita memiliki

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

yang pada akhirnya menyiratkan bahwa

$$\sigma 12 = 0$$

Namaun sebaliknya tidak benar. Yaitu jika $\sigma_{12} = 0$, kita tidak dapat mengklaim bahwa X_1 dan X_2 adalah saling bebas. Semuanya kita katakan bahwa tidak adalah kebergantung linier antara keduanya. Kovarian sampel antra $X1$ dan X_2 dinyatakan dengan

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$

Korelasi

Korelasi antara variabel X_1 dan X_2 adalah standarisasi kovarian, yang didapatkan dengan menormalisasi kovarian dengan standar deviasi masing masing variabel dinyatakan dengan

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

Korelasi sample untuk atribut X_1 dan X_2 dinyatakan dengan

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{m} (x_{i2} - \hat{\mu}_2)^2}}$$

Matrik Kovarian

Variansi dari untuk dua atribut X_1 dan X_2 dapat diringkas dalam matrik covarianse bujursangkar denga ukuran 2×2 dinyatakan dengan

$$\begin{aligned}
&\Sigma = E[(X - \mu)(X - \mu)^T] \\
&= E\left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} (X_1 - \mu_1 \quad X_2 - \mu_2) \right] \\
&= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}
\end{aligned}$$

Karena $\sigma_{12} = \sigma_{21}$, Σ adalah matrik simetris. Matrik vovarian merekam variansi tertentu atribut pada diagonal utamanya, dan informasi covarian pada elemen element bukan diagonal. Total variance dari dua atribut dinyatakan sebagai jumlah elemen elemen diagonal dari Σ , yang juga disebut *trace* dari Σ dinyatakan dengan

$$\text{var}(D) = \text{tr}(\Sigma) = \sigma_1^2 + \sigma_2^2$$

Kita segera memiliki $tr(\Sigma) \geq 0$

Secara umum covarian adalah non-negatif, karena

$$|\Sigma| = \det(\Sigma) = \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 = \sigma_1^2 \sigma_2^2 - \rho_{12}^2 \sigma_1^2 \sigma_2^2 = (1 - \rho_{12}^2) \sigma_1^2 \sigma_2^2$$

dimana kitu gunakan persamaan (2.23), yaiut $\rho_{12} \sigma_1 \sigma_2$. dengan $|\Sigma|$ adalah determinan dari matrik kovarian. Perhatikan bahwa $|\rho_{12}| \leq 1$ menyebabkan $\rho_{12}^2 \leq 1$ sehingga det $(\Sigma) \geq 1$ furthermore determinannya adalah non-negative.

Matrik kovarian sampel dinyatakan dengan

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$$

Matrik kovarian sampe $\hat{\Sigma}$ memilki karakteristik sama seperti Σ , yaitu simetris dan $|\hat{\Sigma}| \geq 0$ dan itu dapat digunakan untum memudahkan mendapatkan total sampel dan variansi secara umum

Contoh (Rata rata Sampel dan Covarian)

Perhatikan atribut sepal length dan sepal width untuk data iris, seperti yang diplot dalam gambar 2.4. Ada n=150 data dalam $d = 2$ ruang dimensi. Rata rata sampel adalah

$$\hat{\mu} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

Matrik covarian dinyatakan dengan

$$\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

Variansi untuk sepal length adalah $\hat{\sigma}_1^2 = 0.681$ dan sepal width adalah $\hat{\sigma}_2^2 = 0.187$. Covarian antara dua atribut adalah $\hat{\sigma}_{12} = -0.039$ dan korelasi antara dua atribut tersebut adalah

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Lalu, ada korelasi yang sangat lemah antara dua atribut tersebut

Total variansi sampel dinyatakan dengan

$$\text{tr}(\hat{\Sigma}) = 0.681 + 0.187 = 0.868$$

dan variansi secara umum dinyatakan dengan

$$\hat{\Sigma}| = \det(\hat{\Sigma}) = 0.681 \cdot 0.187 - (-0.039)^2 = 0.126$$

Analisa Multivariate

Dalam analisa multivariate, kita melihat atribut numerik dengan d dimensi $X_1, X_2, \ldots X_d$. Data dinyatakan degan matrik $n \times d$ seperti berikut

$$D = \begin{pmatrix} \begin{array}{cccc} X_1 & X_2 & \cdots & X_d \end{array} \\ \hline \begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \end{pmatrix}$$

Jika dilihat dari baris data memiliki n objek atatu vektor dalam d ruang dimensi atribut

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T \in \mathbb{R}^d$$

Jika dilihat dari sudut pandang kolom, data diangga sebagai d objek atau vektor dalam n dimensi ruang dengan titik-titik data

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in R^n$$

Jika dilihat dari sudut pandang probabilitas, d atribut dimodelkan dengan variabel acak vektor $X = (X_1, X_2, \ldots X_d)^T$ dan titik titik x_i dianggap sebagai sampel acak yang diperoleh dari X, atribut atribut tersebut independent and identfically distributed dari X (i.i.d X)

Mean

Generalisasi persamaan (2.18) rata-rata vektor multivariate diperoleh dari masing-masing atribut yang dinyatakan dengan

$$\mu = E[X] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

Generalisasi persamaan (2.19) rata-rata sampel dinyatakan dengan

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Matrik Kovarian

Generalisasi persamaan (2.26) untuk d dimensi, kovarian multicovariate di dinyatakan dengan matrik kovarian simetris $d \times d$yang menyatakan kovarian untuk setiap pasangan atribut

$$\Sigma = E[(X - \mu)(X - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Elemen diagonal σ_i^2 menyatakan variansi atribut X_i, dimana elemen-elemen bukan diagonal $\sigma_{ij} = \sigma_{ji}$ menyatakan kovarian antara atribut pasangan X_i dan X_j. Matrik kovarian adalah positif semidefinite

Contoh Rata-rata sample dan matrik covarian.

Perhatikan semua atribut numerik untuk data iris, namanya sepal length, petal length, dan petal width. Rata rata multivarean dinyatakan dengan

$$\hat{\mu} = (5.843 \quad 3.054 \quad 3.759 \quad 1.199)^T$$

dan matrik covarian nya adalah

$$\hat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 & 1.265 & 0.513 \\ -0.039 & 0.187 & -0.320 & -0.117 \\ 1.265 & -0.320 & 3.092 & 1.288 \\ 0.513 & -0.117 & 1.288 & 0.579 \end{pmatrix}$$

Jumlah variansi adalah

$$\text{var}(D) = \text{tr}(\hat{\Sigma}) = 0.681 + 0.187 + 3.092 + 0.579 = 4.539$$

Contoh Perkalian dalam dan perkalian luar. Untuk mengdeskripsikan komputasi perkalian dalam dan perkalian luar dari matrik covarian, perhatikan data 2-dimensi

$$D = \left(\begin{array}{cc} A_1 & A_2 \\ \hline 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{array} \right)$$

Rata-rata vektor adalah sebagai berikut

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} 15/3 \\ 8.7/3 \end{pmatrix} = \begin{pmatrix} 5 \\ 2.9 \end{pmatrix}$$

dan matrik data terpusat dinyatakan

$$Z = D - 1 \cdot \mu^T = \begin{pmatrix} 1 & 0.8 \\ 5 & 2.4 \\ 9 & 5.5 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 5 & 2.9 \end{pmatrix} = \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$

Pendekatan perkalian dalam [pers. 2.30] untuk menghitung matrik kovarian adalah

$$\hat{\Sigma} = \tfrac{1}{n} Z^T Z = \tfrac{1}{3} \begin{pmatrix} -4 & 0 & 4 \\ -2.1 & -0.5 & 2.6 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \\ 0 & -0.5 \\ 4 & 2.6 \end{pmatrix}$$
$$= \tfrac{1}{3} \begin{pmatrix} 32 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}$$

Pendekatan lain yaitu dengan perkalian luar [pers. 2.31] dibyatakan dengan

$$\hat{\Sigma} = \frac{1}{n} \sum_{j=1}^{n} z_i \cdot z_i^T$$

$$= \frac{1}{3} \left[\begin{pmatrix} -4 \\ -2.1 \end{pmatrix} \cdot \begin{pmatrix} -4 & -2.1 \end{pmatrix} + \begin{pmatrix} 0 \\ -0.5 \end{pmatrix} \cdot \begin{pmatrix} 0 & -0.5 \end{pmatrix} + \begin{pmatrix} 4 \\ 2.6 \end{pmatrix} \cdot \begin{pmatrix} 4 & 2.6 \end{pmatrix} \right]$$

$$= \tfrac{1}{3} [\begin{pmatrix} 16.0 & 8.4 \\ 8.4 & 4.41 \end{pmatrix} + \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 0.25 \end{pmatrix} + \begin{pmatrix} 16.0 & 10.4 \\ 10.4 & 6.76 \end{pmatrix}]$$
$$= \tfrac{1}{3} \begin{pmatrix} 32.0 & 18.8 \\ 18.8 & 11.42 \end{pmatrix} = \begin{pmatrix} 10.67 & 6.27 \\ 6.27 & 3.81 \end{pmatrix}$$

dimana data terpusat z_i adalah baris dari Z

Atribut Kategorikal

Kita asumsikan bahwa data terdiri dari satu atribut X. Domain dari X terdiri dari m nilai simbolis $dom(X) = a_1, a_2, \ldots a_m$. Data D adalah $n \times 1$ matrik data simbolis yang dinyatakan dengan

$$D = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

dimana setiap nilai $x_i \in dom(X)$

Variabel Bernouli

Marilah kita lihat kasus ketika atribut kategorikal X memililik domain $\{a_1, a_2\}$ dengan $m = 2$. Kita dapat memodelkan X sebagai variabel acak Bernouli, yang didasarkan pada dua nilai berbeda yaitu 1 dan 0, sesuai dengan pemetaan

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

Fungsi massa probabilitas (PMF) dari X dinyatakan dengan

$$P(X = x) = f(x) = \begin{cases} p_1 & \text{if } x = 1 \\ p_0 & \text{if } x = 0 \end{cases}$$

dimana p_1 dan p_0 adalah parameter distribusi, yang harus memenuhi kondisi

$$p_1 + p_0 = 1$$

Karena hanya ada satu parameter bebas, biasanya menotasikan $p_1 = p$ maka $p_0 = 1 - p$. Fungsi Massa Probabilitas dari variabel acak Bernouli X dapat kemudian ditulis dengan

$$P(X = x) = f(x) = p^x (1 - p)^{1-x}$$

Kita dapat melihat bahwa $P(X = 1) = p^1 (1 - p)^0 = p \text{ and } P(X = 0) = p^0 (1 - p)^1 = 1 - p$ seperti yand diharapkan

Mean dan Variansi

Nilai harapan dari X dinyatakan dengan

$$\mu = E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

dan variansi dari X dinyatakan dengan

$$\begin{aligned} \sigma^2 &= \text{var}(X) = E[X^2] - (E[X])^2 \\ &= (1^2 \cdot p + 0^2 \cdot (1 - p)) - p^2 = p - p^2 = p(1 - p) \end{aligned}$$

Rata-rata sampel dan Variansi

Untuk mengestimasi parameter dari variabel Bernouli X, kita asumsikan bahwa setiap simbol dipetakan ke nilai biner. Sehingga, sekumpulan nilai $x_1, x_2, \ldots x_n$ diasumsikan menjadi sampel acak yang diperoleh dari X (yaitu setiap x_i adalah IID dengan X.

Rata-rata sampel dinyatakan dengan

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{n_1}{n} = \hat{p}$$

dimana n_1 adalah banyaknya titik dengan $x_1 = 1$ dalam sampel acak (sama dengan banyak kejadian dari simbol a_1)

Misal $n_0 = n - n_1$ menyatakan banyak titik dengan $x_i = 0$ dalam sampel acak. Variansi sample dinyatakan dengan

$$\begin{aligned} \hat{\sigma}^2 &= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \\ &= \tfrac{n_1}{n} (1 - \hat{p})^2 + \tfrac{n - n_1}{n} (-\hat{p})^2 \\ &= \hat{p}(1 - \hat{p})^2 + (1 - \hat{p})\hat{p}^2 \\ &= \hat{p}(1 - \hat{p})(1 - \hat{p} + \hat{p}) \\ &= \hat{p}(1 - \hat{p}) \end{aligned}$$

Variansi sampel dapat juga diperoleh langsung dari persamaan(3.1) dengan mensubsitusikan $\hat{p}$ untuk p .

Contoh

Perhatikan atribut sepal length (X) untuk dataset iris dalam tabel 1.1. Marilah kita definisikan bunga iris dengan *Long* jika bunga itu sepal length dalam range $[7, \infty]$, dan *short* jika sepal length dalam range $[-\infty, 7]$. Kemudian X_1 dapat dinyatakan dengan atribut kategorikan dengan domain {Long,Short}. Dari sampel yang diamati ukuran $n = 150$, kita menemukan 13 iris long. Rata-rata sampel dari X_1 adalah

$$\hat{\mu} = \hat{p} = 13/150 = 0.087$$

dan variansinya adalah

$$\hat{\sigma}^2 = \hat{p}(1 - \hat{p}) = 0.087(1 - 0.087) = 0.087 \cdot 0.913 = 0.079$$

Ditribusi binomial : banyaknya kejadian

Diberikan variabel Bernoulli X, misal $\{x_1, x_2, \ldots x_n\}$ menyatakan sampel acak dari ukuran n yang diperoleh dari X. Misal N adalah variabel acak yang menyatakan numlah kejadi dari simbol a_1(nilai $X = 1$) dalam sampe. N adalah distribusi binomial yang dinyatakan dengan

$$f(N = n_1 | n, p) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1}$$

Dalam kenyataannya, N adalah jumlah dari n variabel acak Bernoulli x_i yang saling bebas dan (IID) dengan X yaitu $N = \sum_{i=1}^{n} x_i$. Dengan liniearitas dari ekpektasi, mean atau jumlah harapan dari kejadian simbol a_i dinyatakan dengan

$$\mu_N = E[N] = E\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} E[x_i] = \sum_{i=1}^{n} p = np$$

Karena x_i adalah semuanya saling bebas, variansi dari N dinyatakan dengan

$$\sigma_N^2 = \text{var}(N) = \sum_{i=1}^{n} \text{var}(x_i) = \sum_{i=1}^{n} p(1 - p) = np(1 - p)$$

Contoh 3.2. Dengan meneruskan contoh 3.1, kita dapat menggunakan parameter yang telah diestimasi $\hat{p} = 0.087$ untuk menghitung banyaknya kejadian yang diharapkan N long dari sepal length. distribusi binomial Iris

$$E[N] = n\hat{p} = 150 \cdot 0.087 = 13$$

Dalam kasus ini, karena p dihitung dari sample melalui $\hat{p}$, tidak mengherankan bahwa jumlah kejadian diharapkan dari Long Iris sama dengan kejadian yang sebenarnya. Akan tetapi yang lebih menarik adalah kita dapat menghitung variansi jumlah kejadian

$$\text{var}(N) = n\hat{p}(1 - \hat{p}) = 150 \cdot 0.079 = 11.9$$

Meningkatnya ukuran sample, distribusi binomial seperti yang diberikan dapalam persamaan 3.3 cenderung ke distribusi normal dengan $\mu = 13$ dan $\sigma = \sqrt{11.9} = 3.45$. Sehingga dengan kepercaan lebih besar dari 95%, kita dapat mengklam bahwa jumlah kejadian dari a_i akan terletak dalam rentang $\mu \pm 2\sigma = [9.55, 16.45]$ yang mengikuti dari fakta bahwa untuk distribusi normal 95,45% dari massa probabilitas terletak dalam dua standar deviasi dari rata-rata.

Variable multivariate Bernoulli

Sekarang kita memandang kasus umum ketika X adalah atribut kategorical dengan domain $\{a_1, a_2, \ldots a_m\}$. Kita dapat memodelkan X sebagai variabel acak Bernoulli m -dimensi $X = (A_1, A_2, \ldots, A_m)^T$ dimana setiap A_i adalah variabel Bernoulli dengan parameter p_i yang menotasikan probabilitas dari pengamatan simbol a_i. Akan tetapi karena X dapat mengasumsikan hanya satu dari nilai simbolik pada suatu waktum jika $X = a_i$ maka $A_i = 1$ dan $A_j = 0$ untuk semua $j \neq i$. Variabel acak $X \in 0, 1^m$, dan jika $X = a_i$, maka $X = e_i$, dimana e_i adalah standar vektor basis ke i, $e_i \in \mathbb{R}^m$ yang dinyatakan dengan

$$e_i = (\overbrace{0, \ldots, 0}^{i-1}, 1, \overbrace{0, \ldots, 0}^{m-i})^T$$

Pada e_i hanya elemen ke i adalah 1 ($e_{ii} = 1$) , sedangkan semua elemen yang lain adalah nol, ($e_{ij} = 0, \forall j \neq i$).

Disini, definis yang lebih tepat dari variabel Bernoulli multivariate , yaitu generalisasi dari variabel Bernoullii dari dua hasil ke m hasil. Kita kemudian memodelkan atribut kategorical X sebagai variabel Bernoulli multivariate X didefinisikan dengan

$$X(v) = e_i \text{ if } v = a_i$$

Rentang dari X terdiri dari m nilai vektor berbeda $\{e_1, e_2, \ldots e_m\}$ dengan fungsi massa probabilitas dari X dinyatakan dengan

$$P(X = e_i) = f(e_i) = p_i$$

dimana p_i adalah probabilitas dari nilai pengamatan a_i. Parameter ini harus memenuhi kondisi

$$\sum_{i=1}^{m} p_i = 1$$

Fungsi massa prababilitas dapat ditulis secara utuh sebagai berikut

$$P(X = e_i) = f(e_i) = \prod_{j=1}^{m} p_j^{e_{ij}} Ka$$

Kareana $e_i i = 1$ dan $e_i j = 0$ funtuk $j \neq i$, kita dapat melihat bahwa, seperti yang diharapkan, kita miliki

$$f(e_i) = \prod_{j=1}^{m} p_j^{e_{ij}} = p_1^{e_{i0}} \times \cdots p_i^{e_{ii}} \cdots \times p_m^{e_{im}} = p_1^0 \times \cdots p_i^1 \cdots \times p_m^0 = p_i$$

Bins	Domain	Counts
$[4.3, 5.2]$	Very Short (a_1)	$n_1 = 45$
$(5.2, 6.1]$	Short (a_2)	$n_2 = 50$
$(6.1, 7.0]$	Long (a_3)	$n_3 = 43$
$(7.0, 7.9]$	Very Long (a_4)	$n_4 = 12$

Contoh : Marilah kita lihat atribut sepal length (X_1) untuk data Iris seperti yang ditunjukkan dalam tabel 1.2. Kita membagi sepal length kedalam empat interval yang sama, dan memberikan nama untuk setiap interval seperti yang diunjukkan dalam tabel 3.1. Kita lihat X_1 sebagai atribut kategorical dengan domain

$$\{a_2 = \text{ VeryShort}, a_2 = \text{ Short}, a_3 = \text{Long}, a_4 = \text{Very Long}\}$$

Kita memodelkan atribut kategorical X_1 sebagai variabel X Bernoulli multivariate, didefinisikan dengan

$$X(v) = \begin{cases} e_1 = (1,0,0,0) & \text{jika } v = a_1 \\ e_2 = (0,1,0,0) & \text{jika } v = a_2 \\ e_3 = (0,0,1,0) & \text{jika } v = a_3 \\ e_4 = (0,0,0,1) & \text{jika } v = a_4 \end{cases}$$

Misalkan, simbol $x_1 = Short = a_2$ dinyatakan dengan $(0,1,0,0)^T = e_2$

Mean

Mean atau nilai harapan dari X dapat diperoleh dengan

$$\mu = E[X] = \sum_{i=1}^{m} e_i f(e_i) = \sum_{i=1}^{m} e_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = p$$