Keoptimalan Naïve Bayes Dalam Klasifikasi

M. Ammar Shadiq

Program Ilmu Komputer FPMIPA Universitas Pendidikan Indonesia

ammar.shadiq@gmail.com

Abstrak

Naïve Bayes adalah salah satu algoritma pembelajaran induktif yang paling efektif dan efisien untuk machine learning dan data mining. Performa naïve bayes yang kompetitif dalam proses klasifikasi walaupun menggunakan asumsi keidependenan atribut (tidak ada kaitan antar atribut). Asumsi keidependenan atribut ini pada data sebenarnya jarang terjadi, namun walaupun asumsi keidependenan atirbut tersebut dilanggar performa pengklasifikasian naïve bayes cukup tinggi, hal ini dibuktikan pada berbagai penelitian empiris .

Pada paper ini, penulis akan memaparkan penggunaan naïve bayes dalam tugas klasifikasi data, membuktikan potensi naïve bayes untuk digunakan dalam data yang memiliki korelasi antara atribut dan mengajukan penjelasan mengenai keoptimalan naïve bayes dalam kondisi tertentu.

Kata Kunci : Bayesian Theorem, Naïve Bayes, Data Mining, Classification, Optimal Classification.

Pendahuluan

Klasifikasi adalah salah satu tugas yang penting dalam data mining, dalam klasifikasi sebuah pengklasifikasi dibuat dari sekumpulan data latih dengan kelas yang telah di tentukan sebelumnya. Performa pengklasifikasi biasanya diukur dengan ketepatan (atau tingkat galat)[6].

Teorema Bayes adalah teorema yang digunakan dalam statistika untuk menghitung peluang untuk suatu hipotesis, Bayes Optimal Classifier menghitung peluang dari suatu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal.

Umumnya kelompok atribut *E* direpresentasikan dengan sekumpulan nilai atribut ($x_1,x_2,x_3,\ldots,x_n$) dimana x_i adalah nilai atribut X_i. *C* adalah variable klasifikasi dan *c* adalah nilai dari *C*.

Pengklasifikasian adalah sebuah fungsi yang menugaskan data tertentu kedalam sebuah kelas. Dari sudut pandang peluang [7], berdasarkan aturan bayes kedalam kelas *c* adalah :

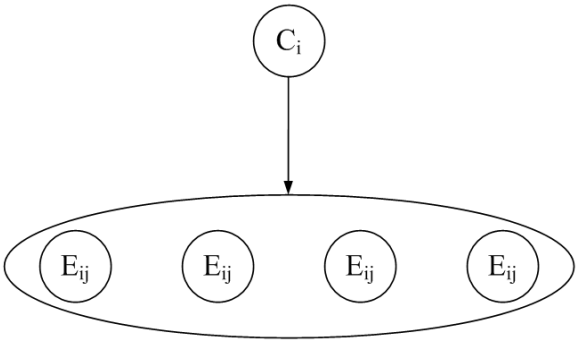
$$P(c|E) = \frac{P(E|c)P(c)}{P(E)}$$

Untuk menentukan pilihan kelas, digunakan peluang maksimal dari seluruh *c* dalam *C*, dengan fungsi :

$$\underset{c \in C}{argmax} \frac{P(E|c)P(c)}{P(E)}$$

Karena nilai $P(E)$ konstan untuk semua kelas, maka $P(E)$ dapat diabaikan. sehingga menghasilkan fungsi :

$$f_c(E) = \underset{c \in C}{argmax} P(E|c)P(c) \quad (1)$$



Gambar 1 : Ilustrasi Teorema Bayes

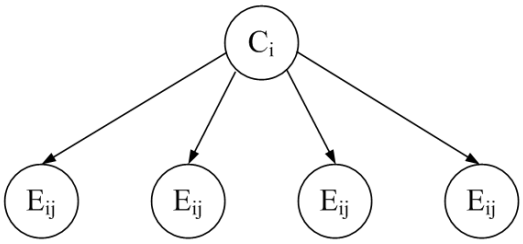
Pengklasifikasian menggunakan Teorema Bayes ini membutuhkan biaya komputasi yang mahal (waktu prosessor dan ukuran memory yang besar) karena kebutuhan untuk menghitung nilai probabilitas untuk tiap nilai dari perkalian kartesius untuk tiap nilai atribut dan tiap nilai kelas.

Data latih untuk Teorema Bayes membutuhkan paling tidak perkalian kartesius dari seluruh kelompok atribut yang mungkin, jika misalkan ada 16 atribut yang masing-masingnya berjenis Boolean tanpa *missing value*, maka data latih minimal yang dibutuhkan oleh Teorema bayes untuk digunakan dalam klasifikasi adalah 2^{16} = 65.536 data, sehingga ada 3 masalah yang dihadapi untuk menggunakan teorema bayes dalam pengklasifikasian, yaitu :

(1) kebanyakan data latih tidak memiliki varian klasifikasi sebanyak itu (oleh karenanya sering diambil sample)
(2) jumlah atribut dalam data sample dapat berjumlah lebih banyak (lebih dari 16)
(3) jenis nilai atribut dapat berjumlah lebih banyak [lebih dari 2 – Boolean] terlebih lagi untuk jenis nilai atribut yang bersifat tidak terbatas 1 - ∞ seperti numeric dan kontiniu.
(4) jika suatu data *X* tidak ada dalam data latih, maka data *X* tidak dapat di klasifikasikan, karena peluang untuk data *X* di klasifikasikan kedalam suatu kelas adalah sama untuk tiap kelas yang ada.

Untuk mengatasi berbagai permasalahan diatas, berbagai varian dari pengklasifikasian yang menggunakan teorema bayes diajukan, salah satunya adalah Naïve Bayes, yaitu penggunaan Teorema Bayes dengan asumsi keidependenan atribut. Asumsi keidependenan atribut akan menghilangkan kebutuhan banyaknya jumlah data latih dari perkalian kartesius seluruh atribut yang dibutuhkan untuk mengklasifikasikan suatu data [4].

$$f_c(E) = \underset{c \in C}{argmax}\, P(c) \prod_{j=1}^{n} P(E_j|c) \quad (2)$$



Gambar 2 : Ilustrasi Naive Bayes

Dampak negative dari asumsi naïve tersebut adalah keterkaitan yang ada antara nilai-nilai atribut diabaikan sepenuhnya. Dampak ini secara intuitif akan berpengaruh dalam pengklasifikasian, namun percobaan empiris mengatakan sebaliknya. Hal ini tentu saja cukup mengejutkan, karena dalam pengaplikasian dunia nyata, asumsi diabaikannya keterkaitan antara atribut selalu dilanggar[1].

Pertanyaan yang muncul adalah apakah yang menyebabkan baiknya performa yang didapatkan dari pengaplikasian asumsi naïve ini? Karena secara intuitif, asumsi keidependenan atribut dalam dunia nyata hampir tidak pernah terjadi. Seharusnya dengan

asumsi tersebut performa yang dihasilkan akan buruk.

Domingos dan Pazzani (1997) pada papernya untuk menjelaskan performa naïve bayes dalam fungsi zero-one loss. Fungsi zero-one loss ini mendefinisikan error hanya sebagai pengklasifikasian yang salah. Tidak seperti fungsi error yang lain seperti squared error, fungsi zero-one loss tidak member nilai suatu kesalahan perhitungan peluang selama peluang maksimum di tugaskan kedalam kelas yang benar. Ini berarti bahwa naïve bayes dapat mengubah peluang posterior dari tiap kelas, tetapi kelas dengan nilai peluang posterior maksimum jarang diubah. Sebagai contoh, diasumsikan peluang sebenarnya dari $P(\oplus|\mathrm{E}) = 0.9$ dan $P(\ominus|\mathrm{E}) = 0.1$, sedangkan peluang yang dihasilkan oleh naïve bayes adalah $P'(\oplus|\mathrm{E}) = 0.6$ dan $P'(\ominus|\mathrm{E}) = 0.4$. nilai peluang tersebut tentu saja berbeda jauh, namun pilihan kelas $\oplus$ tetap tidak terpengaruh.

Bukti Naïve Bayes tidak saja optimal pada asumsi idependen

Seperti yang telah di ketahui bahwa naïve Bayes bernilai optimal ketika seluruh atribut bernilai independen terhadap atribut lainnya. Pada bagian ini akan dibandingkan antara nilai naïve bayes yang seluruh atribut independen terhadap atribut lainnya dan nilai naïve bayes yang tidak seluruh atributnya independen.

Misalkan sebuah data latih, dengan atribut *A*, *B* dan *C* yang bersifat Boolean, dan kelas $\oplus$ dan $\ominus$, dengan peluang yang sebanding untuk tiap kelas $P(\ominus) = P(\oplus) = \frac{1}{2}$. *A* dan *B* berkorelasi penuh (*A* = *B*), sehingga *B* dapat diabaikan.

Prosedur klasifikasi optimal untuk sebuah data tuple adalah untuk menugaskan data tuple tersebut kedalam kelas $\oplus$ jika :

Kelas positif $\oplus$:

$$\oplus \Longrightarrow P(A|\oplus)P(C|\oplus) - P(A|\ominus)P(C|\ominus) > 0$$

Dan sebaliknya, menugaskan kelompok atribut kepada kelas $\ominus$ jika :

Kelas negatif $\ominus$:

$$\ominus \Longrightarrow P(A|\oplus)P(C|\oplus) - P(A|\ominus)P(C|\ominus) < 0$$

Kelas acak $\odot$:

$$\odot \Longrightarrow P(A|\oplus)P(C|\oplus) - P(A|\ominus)P(C|\ominus) = 0$$

Sedangkan prosedur klasifikasi Naïve Bayes yang tidak optimal memperhitungkan juga nilai *B* seperti halnya nilai *B* sama sekali tidak berkorelasi dengan nilai *A*. hal ini sama dengan menghitung nilai *A* dua kali. Untuk naïve bayes rumusnya adalah :

Kelas positif $\oplus$:

$$\oplus \Longrightarrow P(A|\oplus)^2P(C|\oplus) - P(A|\ominus)^2P(C|\ominus) > 0$$

Kelas negatif $\ominus$:

$$\ominus \Longrightarrow P(A|\oplus)^2P(C|\oplus) - P(A|\ominus)^2P(C|\ominus) < 0$$

Kelas Acak $\odot$:

$$\odot \Longrightarrow P(A|\oplus)^2P(C|\oplus) - P(A|\ominus)^2P(C|\ominus) = 0$$

Dengan mengaplikasikan naïve bayes untuk pengklasifikasian yang optimal, maka $P(A|+)$ dapat di representasikan sebagai

$$\oplus \Longrightarrow \frac{P(A)P(\oplus|A)}{P(\oplus)}\frac{P(C)P(\oplus|C)}{P(\oplus)} - \frac{P(A)P(\ominus|A)}{P(\ominus)}\frac{P(C)P(\ominus|C)}{P(\ominus)} > 0$$

Karena $P(\ominus) = P(\oplus)$, maka nilai $P(\oplus)$ dan $P(\ominus)$ tidak perlu dihitung dan dapat diabaikan dalam perhitungan, nilai *P(A)* dan *P(C)* juga mengeliminasi satu sama lainnya dalam operasi pengurangan, sehingga nilai *P(A)* dan nilai *P(C)* tidak perlu di hitung, sehingga setelah pengeliminasian perhitungan yang tidak di perlukan dan didapatkan :

$$\oplus \Longrightarrow P(\oplus|A)P(\oplus|C) - P(\ominus|A)P(\ominus|C) > 0$$

Untuk perhitungan korelasi optimal.

Sedangkan untuk perhitungan korelasi dengan Naïve Bayes :

$$\oplus \Longrightarrow P(\oplus|A)^2P(\oplus|C) - P(\ominus|A)^2P(\ominus|C) > 0$$

Karena dalam peluang nilai peluang maksimal adalah 1, maka dapat dituliskan

P(⊖|A) + P(⊕|C) = 1

P(⊕|A) =1 - P(⊕|C)

Misalkan P(⊕|A) = p dan P(⊕|C) = q

Sehingga rumusnya menjadi

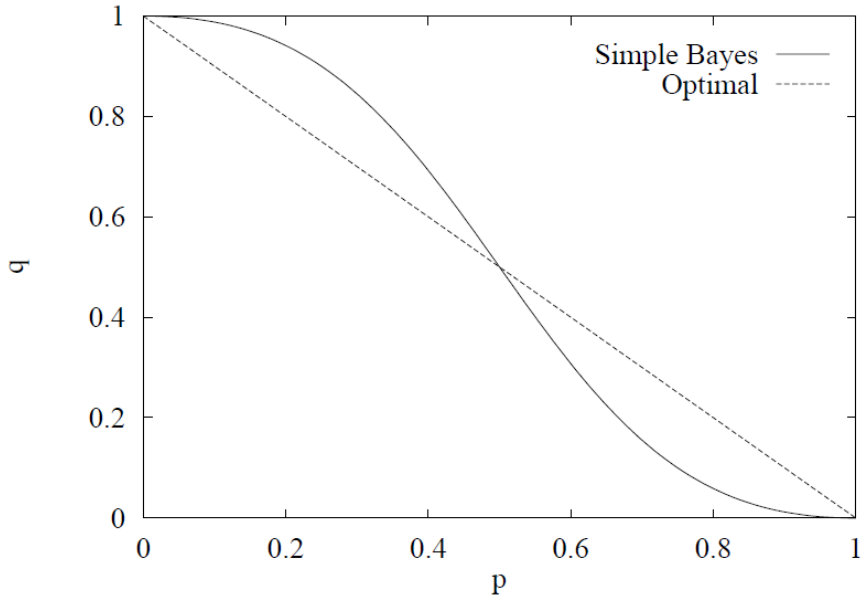
$$\oplus \Longrightarrow \mathrm{pq} - (1-\mathrm{p})(1-\mathrm{q}) > 0 = q > 1 - p$$

untuk nilai peluang optimal dengan asumsi keidependenan atribut.

$$\oplus \Longrightarrow \mathrm{p}^2\mathrm{q} - (1-\mathrm{p})^2(1-\mathrm{q}) > 0 = \mathrm{q} > \frac{(1-\mathrm{p})^2}{\mathrm{p}^2+(1-\mathrm{p})^2}$$

untuk nilai peluang naïve bayes tanpa keidependenan atribut.

Kedua kurva fungsi diatas digambarkan sebagai berikut :



Gambar 3 : Kurva Perbandingan Naive Bayes

Kurva diatas memperlihatkan bahwa walaupun asumsi keidependenan atribut dilanggar, karena *B=A*, pengklasifikasian naïve bayes dengan asumsi atribut yang tidak independen tidak sama dengan pengklasifikasian naive bayes optimal dengan keidependenan atribut hanya di dua bagian sempit, satu diatas kurva dan satu lagi dibawah, di tempat lain, naïve bayes menghasilkan klasifikasi yang benar, yaitu pada (0,1) ($\frac{1}{2}$, $\frac{1}{2}$,) (1,0) ini menunjukkan bahwa

penggunaan klasifikasi naïve bayes bisa lebih luas daripada yang dikira sebelumnya.

Keoptimalan Lokal

Keoptimalan lokal adalah nilai keoptimalan yang didapatkan untuk sebuah kumpulan atribut saja, sedangkan keoptimalan global adalah untuk seluruh kumpulan atribut.

Sebelumnya didefinisikan beberapa hal :

Definisi 1

Misalkan C(E) adalah kelas sebenarnya dari contoh E, dan C_x(E) adalah kelas yang di tugaskan oleh pengklasifikasi X, maka *zero-one loss* dari *X* pada *E* , didefinisakan sebagai :

$$L_X(E) = \begin{cases} 0, & jika\ C_X(E) = C(E) \\ 1, & jika\ C_X(E) \neq C(E) \end{cases} \qquad (3)$$

Zero-one loss adalah ukuran yang tepat jika tugas yang harus dilakukan adalah klasifikasi. Dimana *zero-one loss* memberikan ukuran nilai 1 kepada kesalahan pengklasifikasian. Pada situasi tertentu, kesalahan pengklasifikasian memiliki ukuran prioritas yang berbeda, sebagai contohnya, pada diagnosa medis, ukuran kesalahan mengklasifikasikan seorang pasien yang sakit sebagai sehat berbeda dengan mengklasifikasikan pasian sehat sebagai sakit.

Umumnya, seringkali muncul data latih dengan nilai kelompok atribut yang sama tetapi memiliki kelas yang berbeda. Ini merefleksikan fakta bahwa atribut-atribut tersebut tidak mengandung seluruh informasi untuk menentukan kelas. Maka, secara umum, sebuah data latih *E* tidak akan dihubungkan dengan suatu kelas saja, tetapi dengan peluang kelas $P(C_i|E)$ yang berbentuk vektor, dimana komponen ke *I* merepresentasikan perbandingan nilai munculnya *E* pada kelas C_i.

Ukuran Kesalahan zero-one loss dari *X* pada *E* adalah :

$$L_X(E) = 1 - P(C_X|E)$$

Dimana $C_X(E)$ adalah kelas yang ditugaskan *X* kepada *E* dan $P(C_X|E)$ adalah keakuratan dari *X* pada *E*. definisi ini disederhanakan menjadi persamaan 3 saat sebuah kelas memiliki probabilitas 1 diberikan *E*.

Definisi 2 :

Ukuran bayes untuk sebuah data latih adalah nilai galat *zero-one loss* yang terendah yang didapatkan dari pengklasifikasian manapun pada data latih tersebut [1].

Definisi 3:

sebuah pengklasifikai adalah optimal secara lokal untuk sample jika dan hanya jika nilai zero-one loss pada sample tersebut adalah sama dengan ukuran bayes.

Definisi 4:

Sebuah pengklasifikasi adalah optimal secara global untuk sample jika dan hanya jika pengklasifikasian tersebut bernilai optimal untuk tiap sample pada kumpulan sample tersebut. Sebuah pengklasifikasi adalah optimal secara global untuk sebuah masalah jika dan hanya jika pengklasifikasi tersebut optimal secara lokal untuk tiap sample yang mungkin dari masalah tersebut.

Zero –one loss harus dibedakan dengan *squared error loss* untuk perhitungan galat peluang, perbedaan ini didifenisikan sebagai :

$$SE_X = [P(C|E) - P_X(C|E)]^2$$

Dimana *X* adalah prosedur hampiran dan *C* adalah variable kelas dimana peluangnya ingin dicari. Jika ada ketidakpastian yang

berhubungan dengan *P(C|E), square error loss* didefinisikan sebagai nilai yang diharapkan dari expresi diatas. Fikiran utama dari paper ini, di deskripsikan pada bagian ini, yang dapat dijelaskan sebagai berikut. Saat asumsi independen dilanggar, persamaan 2 akan menjadi suboptimal sebagai probabilitas.

Sebagai contoh, misalkan ada dua kelas, yaitu kelas $\oplus$ dan $\ominus$, dan $P(\oplus|\mathrm{E}) = 0.51$ dan $P(\ominus|\mathrm{E}) = 0.49$ sebagai nilai peluang kedua kelas yang sebenarnya. Klasifikasi optimal adalah menugaskan *E* kepada kelas $\oplus$. Misalkan naïve bayes mendapatkan $P(\oplus|\mathrm{E}) = 0.99$ dan $P(\ominus|\mathrm{E}) = 0.01$. asumsi independen dilanggar dengan sangat jauh, dan square error loss sangat besar, tetapi naïve bayes masih mendapatkan keputusan klasifikasi yang benar, dan meminimalisir zero-one loss.

Misalkan ada dua kelas secara umum, yaitu kelas $\oplus$ dan $\ominus$ seperti sebelumnya,

$$p = P(\oplus|E)$$

$$r = P(\oplus)\prod_{j=1}^{a} P(A_j = v_{jk}|\oplus)$$

$$q = P(\ominus)\prod_{j=1}^{a} P(A_j = v_{jk}|\ominus)$$

Sekarang kita akan menciptakan kondisi yang dibutuhkan untuk keoptimalan local dari naïve bayes dan memperlihatkan bahwa volume dari daerah keoptimalan naïve bayes adalah setengah dari volume $p, r\ dan\ s$.

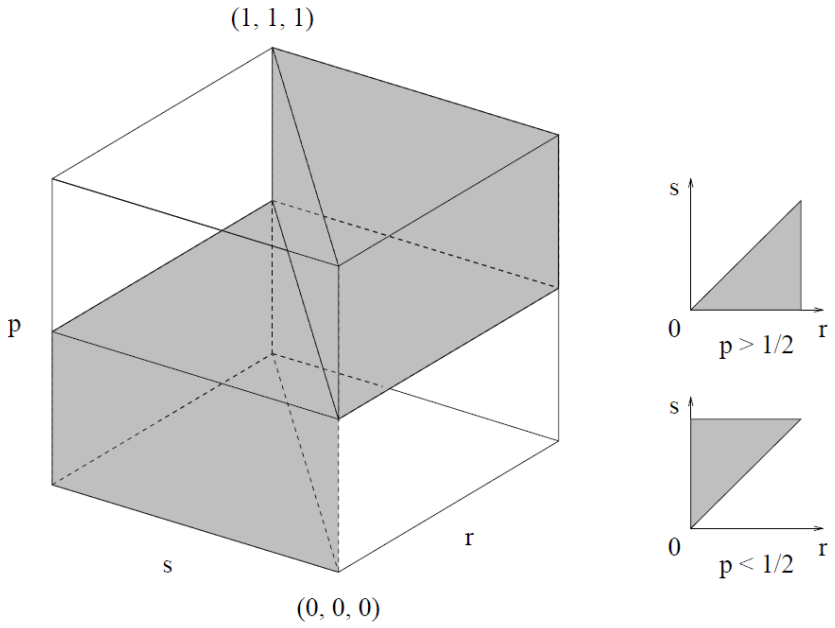
Teorema 1

Naïve bayes optimal secara local dibawah zero-one loss untuk data E jika dan hanya jika $(\boldsymbol{p} \geq \frac{1}{2} \wedge \boldsymbol{r} \geq \boldsymbol{s}) \vee (\boldsymbol{p} \leq \frac{1}{2} \wedge \boldsymbol{r} \leq \boldsymbol{s})$ untuk E.

Bukti : *Pengklasifikasian naïve bayes optimal saat zero-one loss memiliki nilai yang paling minimum. Saat $p = P(\oplus|E) > \frac{1}{2}$ minimum loss adalah $1 - p$ didapatkan dari menugaskan E ke kelas $\oplus$. Pengklasifikasi naïve bayes menugaskan E ke kelas $\oplus$ saat $f_{\oplus}(E) > f_{\ominus}(E)$ berdasarkan persamaan 2, yaitu saat $r > s$. Oleh karenanya jika $p > \frac{1}{2} \wedge r > s$, maka naïve bayes adalah optimal. Sebaliknya, saat $p = P(\oplus|E) < \frac{1}{2}$, zero-one loss minimum didapatkan dengan menugaskan E ke kelas $\ominus$, dimana pengklasifikasian naïve bayes lakukan saat $r < s$. Olehkarenanya pengklasifikasian naïve bayes optimal saat $p < \frac{1}{2} \wedge r < s$. Saat $p = \frac{1}{2}$ keputusan manapun akan optimal, sehingga pertidaksamaan dapat di representasikan sebagai berikut:*

Pengklasifikasian naïve bayes optimal di bawah zero-one loss pada setengah dari volume dari seluruh ruang nilai yang mungkin dari $(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{s})$

Bukti : *Karena p adalah sebuah peluang, dan r dan s adalah produk dari peluang, (p, r, s) hanya menempati nilai dalam kubus [0,1]^3. Daerah dari kubus tersebut yang memuaskan kondisi pada teorema 1 ditunjukkan oleh daerah abu-abu pada gambar 4. Dapat di perhatikan bahwa daerah abu-abu menempati setengah dari volume total kubus. Tetapi tidak seluruh pasangan r dan s mewakili kobinasi peluang yang benar. Karena p tidak dibatasi, maka projeksi dari ruang semesta U dari kombinasi peluang yang valid pada seluruh bidang $p = p_0$ adalah sama. Dengan teorema 1, daerah keoptimalan dari bidang $p_0 = \frac{1}{2}$ dan sebaliknya. Oleh karenanya, jika S adalah area dari projeksi U dan S_0 adalah daerah optimal dari $p_0 < \frac{1}{2}$, daerah optimal untuk $p_0 > \frac{1}{2}$ adalah $S - S_0$, dan volume total dari keoptimalan adalah $\frac{1}{2}S_0 + \frac{1}{2}(S - S_0) = \frac{1}{2}S$.*



Secara kontras dibawah *squared error loss*, persamaan 2 optimal sebagai kumpulan estimasi peluang P(Ci|E) hanya pada saat asumsi independen bertahan, yaitu pada bidang $r = p$ dan $s = 1 - p$ bertemu. Oleh karenanya daerah dari keoptimalan persamaan 2 dibawah *squared error loss* adalah sangat kecil dibandingkan dengan *zero-one loss*. Pengklasifikasian naïve bayes efektif sebagai pemprediksi optimal untuk kelas yang paling sering muncul pada sebuah kondisi yang lebih besar dimana asumsi independen dilanggar. Notasi sebelumnya dari keterbatasan pengklasifikasi naïve bayes sekarang dapat dilihat sebagai kesalahan pengaplikasian intuisi berdasarkan keterbatasan *squared error loss* pada performa pengklasifikasi naïve bayes pada *zero-one loss*.

Keoptimalan global

Ekstensi dari teorema 1 pada keoptimalan global adalah langsung. Misalkan p,r dan s pada data E di indexkan sebagai $p_E, r_E \; dan \; s_E$.

Teorema 2

Pengklasifikasian naïve bayes optimal secara global pada zero-one loss untuk sebuah sample (data set) ∑ jika dan hanya jika
$\forall \boldsymbol{E} \in \sum \left(\boldsymbol{p_E} \geq \frac{1}{2} \wedge \boldsymbol{r_E} \geq \boldsymbol{s_E}\right) \vee (\boldsymbol{p_E} \leq \frac{1}{2} \wedge \boldsymbol{r_E} \leq \boldsymbol{s_E})$

Bukti : *dengan definisi 4 dan teorema 1*

Membuktikan kondisi ini secara langsung pada test sample secara umum tidak dapat dilakukan, karena pembuktian membutuhkan penemuan peluang kelas yang sebenarnya dari setiap kelompok atribut tersebut pada sample. Lebih jauh, membuktikannya pada sebuah permasalahan membutuhkan komputasi yang seukuran dengan banyaknya kumpulan atribut yang dimungkinkan.

Kesimpulan

Pada paper ini telah ditunjukkan bahwa pengklasifikasian Naïve Bayes dibawahpengukuran galat *zero-one loss* memiliki potensi pengaplikasian yang lebih luas

dari yang dikira sebelumnya dan menunjukkan perbedaan pengaplikasian *zero-one loss* dan *squared error loss* dalam pengklasifikasian data, walaupun pembuktian secara mendalam belum dapat dilakukan karena sifat abstraksi data yang sangat tinggi, asumsi-asumsi keoptimalan yang telah dijabarkan diatas paling tidak dapat memberikan acuan untuk pengaplikasian pada data untuk klasifikasi pada sebuah permasalahan tertentu.

Daftar Pustaka

[1] Domingos, P., and Pazzani, M. (1997). *On the optimality of the Simple Bayesian Classifier under Zero-One Loss*.

[2] Tom M. Mitchell (1997). *Machine Learning*. New York, NY: McGraw-Hill.

[3] Duda, R.O., and Hart, P.E. (1973). *Pattern classification and scene analysis.* New York, NY: Wiley.

[4] Berson, A., and Smith S. J. (2001). *Data Warehousing, Data Mining, & OLAP*. New York, NY : McGraw-Hill.

[5] Han, J., and Kamber M. (2000). Data Mining, Concept and Techniques. New York, NY : Morgan Kaufmann.

[6] Walpole, E. R., Myers, R. H. (1995). *Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuan, Edisi ke-4*. Bandung, ITB.

[7] Prof. DR. Sudjana., M.A., M.Sc (1996). *Metoda Statistika, Edisi ke-6*. Bandung, Tarsito.