Data Mining

Naive Bayes Classifier

Asal keluarga

- Bagian dari keluarga "pengklasifikasi probabilistik" sederhana berdasarkan penerapan teorema Bayes dengan asumsi kemandirian yang kuat (naif) di antara fitur-fitur tersebut.

Pemanfaatan

- Naif Bayes telah dipelajari secara luas sejak 1960-an dalam komunitas pencarian teks untuk pengkategorian teks, dokumen dengan frekuensi kata sebagai fitur.
- Dengan pre-pecessing yang sesuai, NBC kompetitif di domain ini dengan metode yang lebih canggih termasuk SVM.
- NBC juga digunakan dalam diagnosis medis otomatis.

Teorema Bayes

- Dalam statistik dan literatur sains komputer, model NBC dikenal dengan berbagai nama, termasuk Simple Bayes dan Independence Bayes.
- Semua nama ini merujuk pada penggunaan teorema Bayes dalam aturan keputusan classifier.

Model Probabilitas Bersyarat

- Naive Bayes adalah model probabilitas bersyarat: diberikan contoh masalah untuk diklasifikasikan, diwakili oleh vektor X = (x_1,…,x_2) yang mewakili beberapa fitur n (variabel independen), itu ditugaskan untuk probabilitas instance ini :

$$p(C_k \mid x_1, \ldots, x_n)$$

untuk setiap K kemungkinan hasil atau kelas C_k.

Probabilitas dengan teorema Bayes

- Menggunakan teorema Bayes, probabilitas bersyarat dapat didekomposisi sebagai

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Contoh Kasus

- Data terdiri dari Hari, Penampakan langit, Kondisi Kelembaban, Angin dan kolom terakhir adalah keputusan bermain, yang harus diprediksi.

Contoh Tabel

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

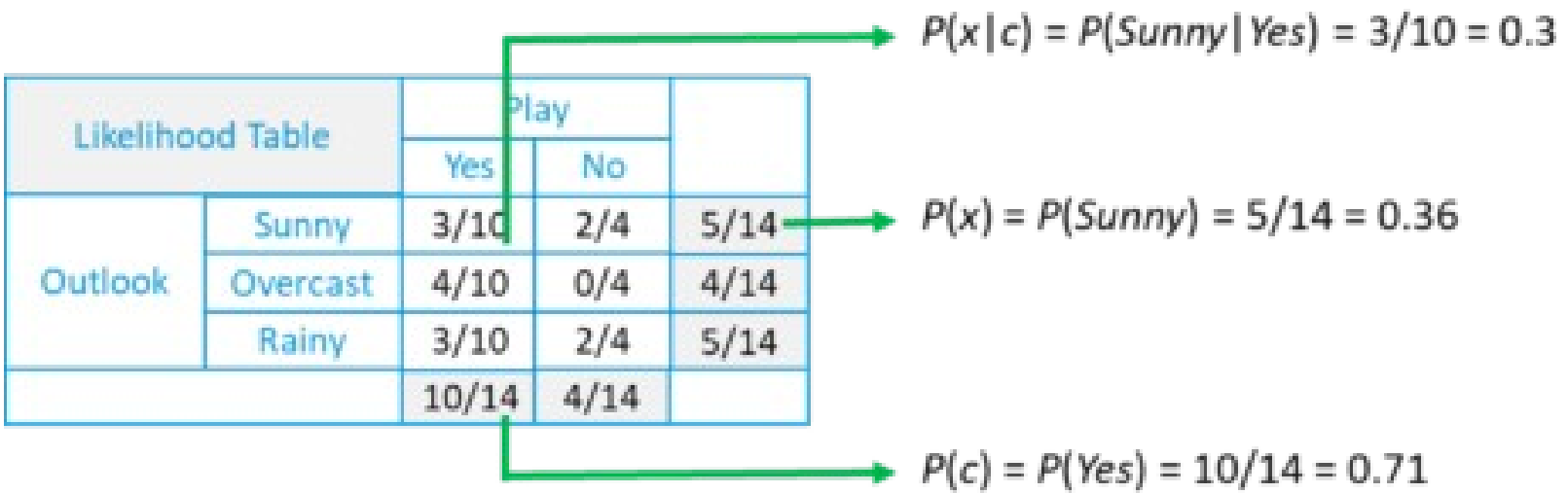
Frekuensi Tabel yang independen

Frequency Table		Play	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	3	2

Frequency Table		Play	
		Yes	No
Humidity	High	3	4
	Normal	6	1

Frequency Table		Play	
		Yes	No
Wind	Strong	6	2
	Weak	3	3

Membuat Likehood



Yes & Sunny

- Kemungkinan ‘Yes’ diberikan ‘Sunny‘ adalah:
  - P(c|x) = P(Yes|Sunny) = P(Sunny|Yes)* P(Yes) / P(Sunny) = (0.3 x 0.71) /0.36 = 0.591

No & Sunny

- Demikian pula Kemungkinan ‘No’ yang diberikan ‘Sunny‘ adalah:
  - P(c|x) = P(No|Sunny) = P(Sunny|No)* P(No) / P(Sunny) = (0.4 x 0.36) /0.36 = 0.40

Likehood yang lain

Likelihood table for Humidity

Likelihood Table		Play		
		Yes	No	
Humidity	High	3/9	4/5	7/14
	Normal	6/9	1/5	7/14
		9/14	5/14	

Likelihood table for Wind

Likelihood Table		Play		
		Yes	No	
Wind	Weak	6/9	2/5	8/14
	Strong	3/9	3/5	6/14
		9/14	5/14	

$P(Yes|High) = 0.33 \times 0.6 / 0.5 = 0.42$

$P(No|High) = 0.8 \times 0.36 / 0.5 = 0.58$

$P(Yes|Weak) = 0.67 \times 0.64 / 0.57 = 0.75$

$P(No|Weak) = 0.4 \times 0.36 / 0.57 = 0.25$

Soal

- Misalkan kita memiliki Hari dengan nilai-nilai berikut:
  - Penampakan langit = Hujan
  - Kelembaban = Tinggi
  - Angin = Lemah
  - Bermain =?

Peluang Yes

- Likelihood dari ‘Yes’ Untuk hari tersebut :
  - P(Outlook = Rain|Yes)*P(Humidity= High|Yes)* P(Wind= Weak|Yes)*P(Yes)
  - = 2/9 * 3/9 * 6/9 * 9/14 = 0.0199

Peluang No

- Likelihood dari ‘No’ pada hari tersebut :
  - P(Outlook = Rain|No)*P(Humidity= High|No)* P(Wind= Weak|No)*P(No)
  - = 2/5 * 4/5 * 2/5 * 5/14 = 0.0166

Normalisasi Peluang

- P(Yes) = 0.0199 / (0.0199+ 0.0166) = 0.55
- P(No) = 0.0166 / (0.0199+ 0.0166) = 0.45

Gaussian naive Bayes

- Ketika berhadapan dengan data kontinu, asumsi tipikal adalah bahwa nilai kontinu yang terkait dengan setiap kelas didistribusikan menurut distribusi normal (atau Gaussian).

Distriusi normal

- Data dikelompokkan berdasarkan kelas, dan dihitung mean dan varians dari x di setiap kelas.
- μ_k menjadi nilai rata-rata dalam x yang terkait dengan kelas C_k, dan biarkan σ_k^2 menjadi varian nilai dalam x yang terkait dengan kelas C_k.

Nilai Observasi

- Nilai observasi v didistribusikan dalam probabilitas v diberi kelas C_k, p (x = v | C_k), dapat dihitung dengan memasukkan v ke dalam persamaan untuk distribusi normal yang diparameterisasi oleh μ_k dan σ_k^2.

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Multinomial Naive Bayes

- Dengan model kejadian multinomial, sampel (vektor fitur) mewakili frekuensi kejadian tertentu yang dihasilkan oleh multinomial (p_1,..., p_n) di mana p_i adalah probabilitas bahwa kejadian i terjadi.
- Vektor fitur x =(x_1,…,x_n) adalah histogram, dengan x_i menghitung berapa kali peristiwa i diamati dalam contoh tertentu.

Multinomial NB dalam teks

- Ini adalah model peristiwa yang biasanya digunakan untuk klasifikasi dokumen, dengan peristiwa yang mewakili kemunculan kata dalam satu dokumen.

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{\ x_i}$$