Klasifikasi: Naïve Bayes

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

March 13, 2017

Selayang Pandang

1. Pendahuluan
2. Naïve Bayes
   - Klasifikasi Bayesian
   - Conditional Independence
   - Kasus Kontinu
   - Kasus Diskrit
3. Pros & Cons
   - Masalah pada Naïve Bayes
   - Keuntungan Naïve Bayes

Bahan Bacaan

1. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. (Chapter 4. Algorithms: the basic method)
2. Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press. (Chapter 3. Generative models for discrete data)
3. Tan, P. N. (2006). Introduction to data mining. Pearson Education India. (Chapter 5. Classification)
4. Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all?. International statistical review, 69(3), 385-398.

Pendahuluan

Peubah Acak

- Dalam pendekatan probabilistik, data dapat dilihat sebagai observasi yang muncul dari model probabilitas untuk sebuah peubah acak (*random variables* atau r.v.)

Peubah Acak

- Dalam pendekatan probabilistik, data dapat dilihat sebagai observasi yang muncul dari model probabilitas untuk sebuah peubah acak (*random variables* atau r.v.)
- Jika diberikan peubah acak diskrit A, maka $P(A)$ adalah fungsi yang memetakan kemungkinan munculnya kelas, kategori, atau kondisi dari A, atau dikenal sebagai **probability mass function** (PMF)

Peubah Acak

- Dalam pendekatan probabilistik, data dapat dilihat sebagai observasi yang muncul dari model probabilitas untuk sebuah peubah acak (*random variables* atau r.v.)
- Jika diberikan peubah acak diskrit A, maka $P(A)$ adalah fungsi yang memetakan kemungkinan munculnya kelas, kategori, atau kondisi dari A, atau dikenal sebagai **probability mass function** (PMF)
- Jika diberikan peubah acak kontinu x, maka $P(x)$ adalah fungsi yang memetakan probabilitas munculnya suatu nilai berdasarkan semua nilai yang ada, atau dikenal sebagai **probability density function** (PDF)

Notasi

- Terkadang, $P(A = a)$ dengan a adalah salah satu kondisi dari r.v. A disingkat sebagai $P(a)$ saja
- Demikian halnya dengan x_1 untuk menggambarkan nilai r.v. x sehingga $P(x = x_1)$ dapat dituliskan sebagai $P(x_1)$ saja

Product Rule & Sum Rule

Product rule
$P(A, B) = P(A|B)P(B)$

Product Rule & Sum Rule

Product rule

$P(A, B) = P(A|B)P(B)$

Sum Rule

Jika diberikan *joint probability* variabel $X_1, X_2, ..., X_N$, *marginal probability* dari sebuah variabel bisa didapatkan dengan penjumlahan dari semua variabel yang lainnya.

Product Rule & Sum Rule

Product rule

$P(A, B) = P(A|B)P(B)$

Sum Rule

Jika diberikan *joint probability* variabel $X_1, X_2, ..., X_N$, *marginal probability* dari sebuah variabel bisa didapatkan dengan penjumlahan dari semua variabel yang lainnya.

Sum Rule

$$P(X_1) = \sum_{x_2} ... \sum_{x_N} P(X_1, X_2 = x_2, ..., X_N = x_N)$$

Marginalisasi

- Notasi pada *sum rule* dapat disederhanakan menjadi

$$P(x_1) = \sum_{x_2} ... \sum_{x_N} P(x_1, x_2, ..., x_N)$$

- Untuk r.v. kontinu, penjumlahannya diganti dengan integral

$$P(x_1) = \int_{x_2} ... \int_{x_N} P(x_1, x_2, ..., x_N) dx_2 ... dx_N$$

- Prosedur ini dikenal dengan nama **marginalisasi**

Berdasarkan *product rule*, kita tahu bahwa

$$P(A, C) = P(A|C)P(C)$$

Namun, kita juga bisa melihat bahwa

$$P(A, C) = P(C|A)P(A)$$

sehingga dapat dirumuskan dengan Bayes' rule

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

Contoh 1

Seorang dokter tahu bahwa meningitis memiliki probabilitas menyebabkan kekakuan leher sekitar 50%. Kasus meningitis sendiri ditemukan dalam 1 dari 50,000 orang. Di sisi lain, probabilitas ditemukannya kasus kekakuan leher adalah $1/20$.

Pertanyaan: Jika seseorang menderita kekakuan leher, berapa peluangnya orang tersebut terkena meningitis?

Contoh 1

Diketahui

$P(s|m) = 0.5$
$P(m) = 1/50,000 = 2 \times 10^{-5}$
$P(s) = 1/20 = 0.05$

Solusi

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.5 \times 2 \times 10^{-5}}{0.05} = 0.0002$$

Contoh 2

Terdapat dua tim sepakbola, tim 0 dan tim 1. Tim 0 memenangkan 65% pertandingan dalam pertemuan kedua tim tersebut, sedangkan tim 1 memenangkan sisanya. Dari semua kemenangan tim 0, hanya 30% terjadi saat keduanya bertanding di kandang tim 1. Di sisi lain, 75% kemenangan tim 1 terjadi saat mereka bermain di kandang.

Pertanyaan: Berapa peluang tim 1 akan menang jika di pertandingan berikutnya mereka akan bermain di kandang?

Naïve Bayes

Klasifikasi Bayesian

- Tujuan: fungsi pembelajaran $f(x) \rightarrow y$
- Klasifikasi probabilistik: kelas yang paling mungkin jika diberikan hasil observasinya, i.e. $\hat{y} = \underset{y}{arg\,max} P(y|x)$
- Probabilitas bayesian dari sebuah kelas:

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_{y'} P(x|y')P(y')}$$

Klasifikasi Bayesian: Komponen

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_{y'} P(x|y')P(y')}$$

- $P(y)$: probabilitas prior dari kelas
  "mana kelas yang sering muncul, mana yang jarang"
- $P(x|y)$: class-conditional model
  "seberapa sering observasi x dalam kasus y"
- $P(x)$: normalisasi

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada
- "Apakah datanya lebih besar probabilitasnya sebagai kelas 1 atau kelas 0?"

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada
- "Apakah datanya lebih besar probabilitasnya sebagai kelas 1 atau kelas 0?"
- Model generatif selalu melakukan **klasifikasi probabilistik**

Model Generatif

- Naïve Bayes menghitung probabilitas **untuk masing-masing kelas** yang ada
- "Apakah datanya lebih besar probabilitasnya sebagai kelas 1 atau kelas 0?"
- Model generatif selalu melakukan **klasifikasi probabilistik**
- Klasifikasi probabilistik tidak berarti generatif, e.g. *logistic regression*

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali
- Contoh: MNIST punya 784 variabel, dengan nilai biner saja artinya ada 2^{784} kemungkinan pola!

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali
- Contoh: MNIST punya 784 variabel, dengan nilai biner saja artinya ada 2^{784} kemungkinan pola!
- Namun, kita mengetahui observasi untuk **masing-masing nilai** x_i untuk setiap kelas

Asumsi Independensi

- Kita harus menghitung $P(\mathbf{x}|y)$, tetapi variabelnya bisa banyak sekali
- Contoh: MNIST punya 784 variabel, dengan nilai biner saja artinya ada 2^{784} kemungkinan pola!
- Namun, kita mengetahui observasi untuk **masing-masing nilai** x_i untuk setiap kelas
- Asumsi yang digunakan Naïve Bayes adalah $x_1 ... x_d$ *conditionally independent* jika diberikan y

$$P(\mathbf{x}|y) = \prod_{i=1}^{d} P(x_i|x_1, ..., x_{i-1}, y) = \prod_{i=1}^{d} P(x_i|y)$$

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen
- Bisa jadi independen jika kita tahu cuaca sedang terik

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen
- Bisa jadi independen jika kita tahu cuaca sedang terik
- Cuaca terik “menjelaskan” dependensi antara pergi ke pantai dan *heatstroke*

Conditional Independence

- Probabilitas pergi ke pantai dan *heatstroke* tidak independen
- Bisa jadi independen jika kita tahu cuaca sedang terik
- Cuaca terik "menjelaskan" dependensi antara pergi ke pantai dan *heatstroke*
- Dalam klasifikasi, nilai kelas menjelaskan hubungan antaratribut

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*: $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*: $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$
- Probabilitas kelas: $P(setosa) = P(versicolor) = P(virginica) = 1/3$

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*: $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$
- Probabilitas kelas: $P(setosa) = P(versicolor) = P(virginica) = 1/3$
- Asumsi: atribut terdistribusi Gaussian dan independen jika diketahui kelasnya

Contoh Kasus Kontinu

- Identifikasi iris berdasarkan *petal length* dan *petal width*: $y = \{setosa, versicolor, virginica\}$, atribut: $\{l, w\}$
- Probabilitas kelas: $P(setosa) = P(versicolor) = P(virginica) = 1/3$
- Asumsi: atribut terdistribusi Gaussian dan independen jika diketahui kelasnya
- Dicocokkan dengan *maximum likelihood estimation* (MLE) untuk Gaussian, e.g.

$$\hat{\mu}_{l,setosa} = \frac{1}{50} \sum_{i;y=setosa} l_i$$

$$\hat{\sigma}^2_{l,setosa} = \frac{1}{50} \sum_{i;y=setosa} (l_i - \hat{\mu}_{l,setosa})^2$$

Distribusi Gaussian

PDF

$$P(x|\mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

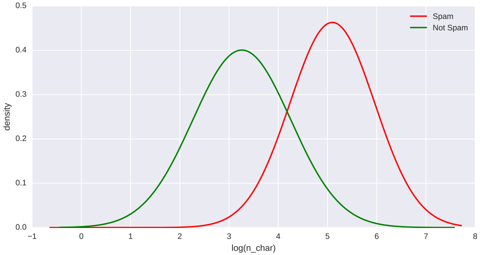
Maximum Likelihood Estimation (MLE)

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

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

Decision Boundary

- Beda rataan, variansi sama: garis lurus atau bidang lurus
- Rataan sama, beda variansi: lingkaran atau elips
- Kasus umum: kurva parabola



Gambar : Perbedaan dua Gaussian

Contoh Kasus Diskrit

Asumsi: Distribusi Bernoulli
Contoh pada kasus identifikasi spam e-mail

D1: "send us your password" (s)
D2: "send us your review" (h)
D3: "review your password" (h)
D4: "review us" (s)
D5: "send your password" (s)
D6: "send us your account" (s)
Dokumen baru: "review us now"

word	**spam**	**ham**
password	2/4	1/2
review	1/4	2/2
send	3/4	1/2
us	3/4	1/2
your	3/4	1/2
account	1/4	0/2

$P(spam) = 4/6,\ P(ham) = 2/6$

Contoh Kasus Diskrit

$$P(review\ \ us|spam) = P(0,1,0,1,0,0|spam)$$
$$P(review\ \ us|ham) = P(0,1,0,1,0,0|ham)$$
$$P(ham|review\ \ us) \approx 0.87$$

Pros & Cons

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$
- Solusi: Laplace smoothing, i.e. penambahan angka positif kecil ke semua pencacahan

$$P(w|c) = \frac{num(w, c) + \epsilon}{num(c) + 2\epsilon}$$

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$
- Solusi: Laplace smoothing, i.e. penambahan angka positif kecil ke semua pencacahan

$$P(w|c) = \frac{num(w, c) + \epsilon}{num(c) + 2\epsilon}$$

- Nilai ϵ contohnya 1 atau 0.5, tetapi bisa juga dengan $num(w)/num$

Masalah Zero-frequency

- Berdasarkan contoh sebelumnya, setiap e-mail dengan kata "account" akan dianggap spam karena $P(account|ham) = 0/2$
- Solusi: Laplace smoothing, i.e. penambahan angka positif kecil ke semua pencacahan

$$P(w|c) = \frac{num(w, c) + \epsilon}{num(c) + 2\epsilon}$$

- Nilai ϵ contohnya 1 atau 0.5, tetapi bisa juga dengan $num(w)/num$
- Kasus ini sering terjadi karena Zipf's law (50% kata hanya muncul sekali)

Masalah Conditional Independence

- Asumsi ini pada banyak kasus kurang tepat, terlalu naif
- Setiap kasus dianggap berkontribusi sama kepada kelas
- *Classifier* yang dihasilkan dapat ditipu dengan memperbanyak kata-kata yang mengindikasikan bahwa e-mail tersebut “ham”

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i =?, ..., X_d = x_d | y)$?

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i =?, ..., X_d = x_d|y)$?
- Naïve Bayes dapat mengabaikan atribut tersebut karena *conditional independence*

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i = ?, ..., X_d = x_d | y)$?
- Naïve Bayes dapat mengabaikan atribut tersebut karena *conditional independence*
- Hitung saja berdasarkan atribut yang bernilai!

Missing Values

- Misalkan kita tidak punya nilai untuk atribut X_i, bagaimana kita bisa menghitung $P(X_1 = x_1, ..., X_i =?, ..., X_d = x_d|y)$?
- Naïve Bayes dapat mengabaikan atribut tersebut karena *conditional independence*
- Hitung saja berdasarkan atribut yang bernilai!
- Nilai yang hilang tersebut tidak perlu diganti

Incremental Updates

- Dengan menyimpan data dalam bentuk jumlah, data baru dapat dimutakhirkan dengan menambahkannya ke variabel yang sudah ada
- Saat perlu diklasifikasi, baru hitung nilai yang dibutuhkan
- Berlaku untuk kasus diskrit maupun kontinu

Salindia ini dipersiapkan dengan sangat dipengaruhi oleh:
Victor Lavrenko (2014)

Terima kasih